Sense and Reference on the Web

Harry Halpin

Doctor of Philosophy

Institute for Communicating and Collaborative Systems

School of Informatics

University of Edinburgh

2009

Abstract

This thesis builds a foundation for the philosophy of the Web by examining the crucial question: What does a Uniform Resource Identifier (URI) mean? Does it have a sense, and can it refer to things? A philosophical and historical introduction to the Web explains the primary purpose of the Web as a universal information space for naming and accessing information via URIs. A terminology, based on distinctions in philosophy, is employed to define precisely what is meant by information, language, representation, and reference. These terms are then employed to create a foundational ontology and principles of Web architecture. From this perspective, the Semantic Web is then viewed as the application of the principles of Web architecture to knowledge representation. However, the classical philosophical problems of sense and reference that have been the source of debate within the philosophy of language return. Three main positions are inspected: the logicist position, as exemplified by the descriptivist theory of reference and the first-generation Semantic Web, the direct reference position, as exemplified by Putnam and Kripke's causal theory of reference and the second-generation Linked Data initiative, and a Wittgensteinian position that views the Semantic Web as yet another public language. After identifying the public language position as the most promising, a solution of using people's everyday use of search engines as relevance feedback is proposed as a Wittgensteinian way to determine sense of URIs. This solution is then evaluated on a sample of the Semantic Web discovered by via using queries from a hypertext search engine query log. The results are evaluated and the technique of using relevance feedback from hypertext Web searches to determine relevant Semantic Web URIs in response to user queries is shown to considerably improve baseline performance. Future work for the Web that follows from our argument and experiments is detailed, and outlines of a future philosophy of the Web laid out.

Acknowledgements

I dedicate this thesis to my father and mother, Harry Halpin Sr. and Rebecca Halpin, who have been a constant source of love. Likewise, the thesis would not have been possible without the support of my community of friends and colleagues across the globe and in Edinburgh. The unwavering support of my advisor, Henry S. Thompson, encouraged me to to pursue considering the Web architecture a first-rate citizen of inquiry, a brave act few advisors would have been willing to do. I can never give enough thanks to my advisor Andy Clark for philosophical inspiration and Victor Lavrenko for his invaluable help on the empirical evaluation. Conversations and support from other colleagues at Edinburgh, in particular Ewan Klein and Kavita Thomas, has been important. However, even more support has come from the global community of Semantic Web hackers and researchers. I have been particularly privileged to have had numerous discussions with Pat Hayes and Tim Berners-Lee on these subjects, and I hope I have accurately given an exegesis of their debate. My time at Duke, where I have was fortunate enough to study under Fredric Jameson and Michael Hardt, has had a decisive if subterranean influence on this thesis. Various friends and co-authors deserve my gratitude. In particular, I would like to single out Rob Didham, David Graeber, Dan Connolly, Brandon Jourdan, Jochen Leidner, Priya Reddy, Malamo Korbetis, Simon Lewis, Claire Grover, Richard Tobin, Peter Buneman, Phil Wadler, Valentin Robu, Michael Wheeler, Alan Bundy, Laura Gomez, Piotr Bultoc, Dan Brickley, Orit Halpern, Paolo Bouquet, Nicholas and Rita Tishuk, Elliott and Elena Madison, Ras Al-Majnuun, and everyone in Bilston, the Forest Cafe, and Carrboro. Lastly, I have found intellectually invaluable my time at the Santa Fe Institute, the Oxford Internet Institute, the Island seminar with Brian Cantwell Smith, and the Interface Seminar at Duke University. Special gratitude must go to the late Karen Spärck Jones, who called me out of the blue and encouraged this philosophical approach to information retrieval and the Semantic Web when I was first beginning.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Harry Halpin*)

Table of Contents

Chapter 1

Introduction

To imagine a language means to imagine a form of life. **Ludwig Wittgenstein (1953)**

The World Wide Web is without a doubt one of the most significant computational phenomena to date. Yet there are some questions that cannot be answered without a *theoretical* understanding of the Web. Although the Web is impressive as a practical success story, there has been little in the way of developing a theoretical framework to understand what – if anything – is different about the Web from the standpoint of long-standing questions of sense and reference in philosophy. While this situation may have been tolerable so far, serving as no real barrier to the further growth of the Web, with the development of the Semantic Web, a next generation of the Web "in which information is given well-defined meaning, better enabling computers and people to work in cooperation," these philosophical questions come to the forefront, and only a practical solution to them can help the Semantic Web repeat the success of the hypertext Web (Berners-Lee et al., 2001).

1.1 Motivation

There is little doubt that the Semantic Web faces gloomy prospects. On first inspection, the Semantic Web appears to be a close cousin to another intellectual project, known politely as 'classical artificial intelligence' (also known as 'Good-Old Fashioned AI'), an ambitious project whose progress has been relatively glacial and whose assumptions have been found to be cognitively questionable (Clark, 1997). The initial bet of the Semantic Web was that somehow the *Web* part of the Semantic Web would somehow overcome whatever problems the Semantic Web inherited from classical artificial

intelligence, in particular, its reliance on logic and inference as the basis of meaning (Halpin, 2004). However, progress on the Semantic Web has also been relatively slow over the last decade. Both new techniques and large amounts of data have not yet caused the Semantic Web to repeat the phenomenal success of the hypertext Web.

In order to even understand the astounding ascent of the Web we have to understand what fundamental component serves as its foundation. While we will go into this question in much greater detail in Chapter 4, tentatively we propose that the Web consists of a space of names called ***Uniform Resource Identifiers*** (URIs), *a unique identifier whose syntax is given in Berners-Lee et al. (2005).* Familiar examples of URIs include URIs for accessing web-pages, such as `http://www.example.org`, although even something as simple as a telephone number can be given a URI such as `tel:+1-816-555-1212`. It is precisely the use of URIs as their fundamental element that makes both the hypertext and Semantic Web part of the Web.

The first problem that is self-evident to anyone who actually attempts to deploy any 'real world' data on the Semantic Web is that there is little guidance on how to identify data using URIs, as well as what information to allow access to from these URIs. For a long time, this question was unanswered, and recently has only been cryptically answered (Sauermann and Cygniak, 2008). The second self-evident problem that is unavoidable to anyone using the Semantic Web for data integration is that different people create different URIs for the same thing. Recently, a set of principles known as 'Linked Data' have given some guidance, but only on a superficial level (Bizer et al., 2007).

The essential bet of the Semantic Web is that decentralized agents will come to an agreement on using the *same* URI to name a thing, including things that aren't accessible on the Web, like people, places, and abstract concepts. Yet there is virtually no ability to even find URIs for things on the Semantic Web. Currently, each application creates its own new URI for a thing, repeating the localism of classical artificial intelligence. Furthermore, it appears that most things either have no URIs or far too many.

1.2 Hypothesis

The scientific hypothesis of this thesis must be stated in a two-fold fashion, first to state the problem and then to propose a solution. The problem is the simple question: **What is the meaning of a URI?**. In order to analyze this problem further, we will

propose that **the Semantic Web is a kind of language that can be defined by its conformance to the principles of Web architecture, but nonetheless determining the meaning of a URI decomposes into a theory of sense and reference, so the Semantic Web inherits the classical problems regarding sense and reference from the philosophy of natural language**. Our proposed solution is then that **a theory of sense and reference suitable to encourage identifier re-usage on the Web can be implemented by employing relevance feedback from search engine results.**

In order to orient the reader to the Web, we give a brief introduction to its history and significance in Chapter 2. We then introduce the philosophical terminology that serves as the foundation the thesis in Chapter 3. Finally, we use this terminology to give an exegesis of Web architecture in Chapter 4. In Chapter 5 we propose that the Semantic Web, at least as embodied by the Resource Description Framework (RDF), is a kind of URI-based knowledge representation language for data integration based on the principles of Web architecture.

We address current theories of sense and reference in Chapter 6 and propose a neo-Wittgensteinian theory of sense and meaning for the Web in Chapter 8. There are three distinct positions to this question on the Semantic Web, each corresponding to a distinct philosophical theory of reference. The first response is the ***logicist position***, which states that *the referent(s) of a URI is determined by whatever model(s) satisfy the formal semantics of the Semantic Web* (Hayes, 2004). This answer is identified with both the formal semantics of the Semantic Web itself and the traditional Russellian theory of names and its descriptivist descendants (Russell, 1905). While this answer may be sufficient for automated inference engines, this answer is insufficient for humans, as it often crucially under-determines what kind of things the URI refers to. As the prevailing position in early Semantic Web research, this position has borne little fruit. Another response is the ***direct reference position*** for the Web, which states that *the meaning of a URI is whatever was intended by the owner.* This answer is identified with the intuitive understanding of many of the original Web architects like Berners-Lee and a special case of Putnam's 'natural kind' theory of meaning. This position is also nearly identical to Kripke's famous response to Russell, the causal theory of reference (Kripke, 1972; Putnam, 1975).

In Chapter 7, we describe a search engine query log from a major hypertext search engine (Microsoft *Live.com*), and how we derive query terms for people, places, and abstract concepts from this query log and then use those to derive Semantic Web URIs. From this query-driven analysis of the deployed Semantic Web, we empirically demon-

strate that following the principles of Web architecture and endorsing the direct reference position does not lead to URI re-usage, but that instead there are still likely to be multiple URIs for the same thing and that it is not easy for users to retrieve these URIs in response to a query given as keywords to a search engine. We finally turn to the third position, the ***public language position***, which states that since *the Semantic Web is a form of language* and as *a language exists as a mechanism for co-ordination among multiple agents, then the meaning of a URI is the use of the URI by a community of agents*. As vague as this position seems at first glance, we argue this analysis of sense and reference is the best fit to how natural language works, and it supersedes and even subsumes the two other positions. While there are 'semiotic' theories of reference, we will not inspect these in this thesis, although we believe that these theories can be incorporated into a public language position. As this theory of meaning works for natural language, it follows that it is a good bet for the Semantic Web, for the Semantic Web is just a form of language, albeit an unusual one.

The public language position implies a public mechanism that lets agents in turn create, find, and re-use URIs. While it may be intuitively correct to endorse a neo-Wittgensteinian theory of meaning for the Semantic Web, this does little for the Semantic Web if a practical implementation can not be demonstrated. As Wittgenstein would say, one must remember that every "language game" comes with a "form of life" (1953). Without a doubt, one activity that seems to be prevalent among users of the Web is searching for web-pages using natural language query terms via a search engine (Halpin and Thompson, 2005). Therefore, the obvious solution to the problem of finding out what a URI means is to take advantage of current search engines. Chapter 8 details on a high-level of abstraction a design for an implementation of determining URI meaning based on relevance feedback from users of keyword-based hypertext search engines. This puts the the Semantic Web in a "virtuous cycle" with the behavior of users on the hypertext Web (Baeza-Yates, 2008). Our implementation is then tested with real data and real users in Chapter 9, and we show how our results improve various baseline systems for the information retrieval of Semantic Web URIs. Finally in Chapter 10 we summarize the work so far and discuss the advantages and limitations of our particular proposed solution. We also present plans for future work as well as further philosophical questions that arise from the thesis.

Each of these chapters builds upon each other to make the thesis complete as a whole. Readers interested in particular subjects may wish to focus their attention on particular components, although they are warned that concepts and findings developed

in earlier chapters are referred to in later chapters. As the nature of the project is in an interdisciplinary and emergent area, there is no singular and comprehensive literature review in a separate chapter, but instead the literature is reviewed and mentioned as necessary throughout the thesis.

1.3 Scope

This thesis is explicitly limited in scope, concentrating only on the terminology necessary to phrase a single, if broad, question: "How can we determine the meaning of a URI on the Semantic Web?" Although the thesis is interdisciplinary, as it involves elements as diverse as the philosophy of language and machine-learning, these elements are only harnessed insofar as they are necessary to phrase our central hypothesis and present a possible solution.

Due to this constraint, this thesis is not an attempt to develop a philosophy of computation (Smith, 2002a), or a philosophy of information (Floridi, 2004), or even a comprehensive "philosophy of the Web" (Halpin, 2008b). These are much larger projects outside the scope of a single thesis, and even a single individual. However, in combination with the fully-formed work in the philosophy of mind and language, we hope that at least this thesis provides a starting point for future work in these areas. So we use notions from philosophy selectively, and then define the terms in lieu of our goal of articulating the principles of Web architecture and the Semantic Web, rather than attempting to articulate or define the terms of a systematic philosophy of the Web. Many of the philosophical terms in this thesis could be explored much further, but are necessarily not explored, as to constrain the thesis to a reasonable size. Unlike a philosophical thesis, counter-arguments and arguments are generally not given for terminological definitions, but instead references are given to the key works that explicate these notions further.

This thesis does not inspect every single possible answer to the question of *What is the meaning of a URI?*, but only three distinct positions. An inspection of every possible theory of meaning and reference is beyond the scope of the thesis, as is an inspection of the tremendous secondary literature that has accrued over the years for even those limited viewpoints that we do inspect in Chapter 6 and Chapter 8. Instead, we will focus only on theories of meaning and reference that have been brought up explicitly in the various arguments over this question in the Web by the primary architects of the Web and the Semantic Web. Our proposed solution rests on a theory of

meaning based on Wittgensteinian, one that is one of the most infamously dense and infuriatingly obscure treatments of sense and reference.

Finally, while the experimental component has done its best to be realistic, it is in no way complete. Pains have been taken to ensure that the experiment, unlike much work in the Semantic Web, at least uses real data, feedback from real users, and is properly evaluated over a wide range of algorithms and parameters. Yet a real implementation of our proposed solution would require full-scale implementation and cooperation of both a major hypertext search engine and a Semantic Web search engine. Obviously, this is beyond the means of a thesis, as is any foundational or even groundbreaking work in information retrieval. Instead, we show how information retrieval can be applied to the Semantic Web to help solve one of its most difficult problems. While various parts of the experiment could no doubt be optimized and scaled up still further, for a proof-of-concept solution to a very difficult problem, this experiment should be sufficient.

1.4 Notational Conventions

In order to aid the reader, the thesis employs a number of notational conventions. In particular, we only use “double” quotes to quote a particular author or other work. When a new word is introduced and deployed in an unusual manner to be clarified later, we use ‘single’ quotes. The use of ‘single’ quotes is also used when a word is supposed to be understood as the word *qua* word, a mention of the word, rather than a use of the word. When a term is defined, the word is first labeled using ***bold and italic*** fonts, and either immediately followed or preceded by the definition given in *italics*. Mathematical or formal terms are *italicized*, as is the use of *emphasis* in any sentence. Finally, the names of books and other large works are often italicized. In general, technical terms like HTTP are often abbreviated by their capitalized initials. One of the largest problems of this whole area historically has been a rather ad-hoc use of terms, and we hope this fairly rigorous notational convention helps separate the use, mention, definition, and direct quotations of words.

1.5 Summary

Despite its ambitious title, this thesis is a *modest* attempt to both articulate and apply the principles of Web architecture in order to answer a question at the heart of the

Semantic Web: *What does a URI mean?* We provide a solution by analyzing the primary positions in philosophy of language and Web architecture, and by constructing a proof-of-concept solution. We do not claim to provide a complete or unique solution, but do argue our solution is better than other competing positions and solutions, in particular in lieu of our implementation. We do not claim to have solved any of these problems regarding meaning and reference for language in general, especially natural language, and are fully confident that philosophers will continue arguing over these issues for at least the next century. We do present a proof-of-concept solution for these problems of meaning and reference in the special and limiting case of the Semantic Web.

Chapter 2

The Significance of the Web

If we could rid ourselves of all pride, if to determine our species we kept strictly to what historic and prehistoric periods show us to be the constant characteristic of man and of intelligence, we should not say Homo Sapiens but Homo Faber. In short, intelligence, considered in what seems to be its original feature, is the faculty of manufacturing artificial objects, especially tools for making tools. **Henri Bergson** (1911)

The subject matter of this thesis is the nature of sense and reference on the World Wide Web, and this chapter provides the necessary background information to motivate the thesis and to make the central hypothesis of the thesis comprehensible. In this thesis, we consider the World Wide Web (from hereon referred to only as 'the Web') as a first-class subject matter for study. The first chapter delves into the origins of the Web so that the question of meaning and reference on the Web can be understood in its proper context.

Why the Web? Why not look at more interesting problems in a subject like artificial intelligence? In his *One Hundred Billion Lines of C++*, computer scientist-turned-philosopher Brian Cantwell Smith notes that the models of computing used in debates over reference and representation tend to frame the debate as if it were between "classical" logic-based symbolic reasoners and some "connectionist" and "embodied" alternative ranging from neural networks to epigenetic robotics (1997). Smith then goes on to aptly state that the kinds of computational systems discussed in artificial intelligence and philosophy tend to ignore the vast majority of existing systems, for "it is impossible to make an exact estimate, but there are probably something on the order of 10, or one hundred billion lines of C++ in the world. And we are barely started. In sum: symbolic AI systems constitute approximately 0.01% of written software"

(1997). The same small fraction likely holds true of "non-symbolic AI" computational systems such as robots, artificial life, and connectionist networks. While numbers by themselves hold little intellectual weight, one could always argue that the vast majority of computational systems may have no impact on our understanding of representation and intelligence. In this thesis we argue otherwise. The wide class of computational systems present a "middle distance" where questions of reference, representation, and intelligence come to the forefront and may even be more tractable than in the case for humans (Smith, 1995). One of the the most significant members to date of this wider class of computational systems is the World Wide Web, described by Tim Berners-Lee, the person widely acclaimed to be the 'inventor' of the Web, as "a universal information space"(1992).

Michael Wheeler, a philosopher who is well-known for his Heideggerian defense of embodiment, surmises that "the power of the Web as a technological innovation is now beyond doubt" but "what is less well appreciated is the potential power of the Web to have a conceptual impact on cognitive science" and so this thesis may provide a new "fourth way" in addition to the "three kinds of cognitive science or artificial intelligence: classical, connectionist, and (something like) embodied-embedded" (2008). While countless papers have been produced on the technical aspects of the Web, very little has been done explicitly on the Web *qua* Web as a subject matter. This does not mean there has not been interest, although the interest has come in particular more from the side of those working on developing the Web rather than those already entrenched in philosophy, linguistics, and artificial intelligence. In particular, the workshop series on *Identity, Reference, and the Web* has provoked many articles on these topics from prominent Web architects, although not from philosophers per se (Halpin et al., 2006; Bouquet et al., 2007b, 2008). In this spirit, what we will undertake in this thesis as a whole is to apply many well-known philosophical theories of reference and representation to the phenomenon of the Web.

In order to establish the relative autonomy of the Web as a subject matter, we recount its origins and so its relationship to other projects, both intellectual such as Engelbart's Human Augmentation Project, as well as more purely technical projects such as the Internet (1962). It may seem odd to begin out this thesis, which involves very specific questions about meaning and reference on the Web, with a thorough history of the Web. To understand these questions we must first have an understanding of the boundaries of the Web and the normative documents that define the Web. The Web is a fuzzy and ill-defined subject matter whose precise boundaries and even def-

inition are unclear. Unlike some subject matters like chemistry, the subject matter of the Web is not necessarily very stable, for the Web is not a 'natural kind,' as it is a technical artifact. So we will take the advice of the philosopher of technology Gilbert Simondon, "Instead of starting from the individuality of the technical object, or even from its specificity, which is very unstable, try to define the laws of its genesis in the framework of this individuality or specificity, it is better to invert the problem: It is from the criterion of the genesis that we can define the individuality and the specificity of the technical object: the technical object is not this or that thing, given *hic et nunc* but that which is generated" (1958). In other words, we must first trace the creation of the Web before attempting to define it, imposing on the Web what Fredric Jameson calls "the one absolute and we may even say transhistorical imperative, that is: Always historicize!" (1981). We build on the work of this chapter in Chapter 4 to delineate the precise principles of the Web.

2.1 The Origins of the Web

What is the Web, and what is its significance? At first, it appears to be a relative upstart upon the historical scene, with little connection to anything before it, an ahistorical and unprincipled 'hack' that came unto the world unforeseen and with dubious academic credentials. The purpose of this section is to dispel this myth.

The intellectual trajectory of the Web is a fascinating, if mostly unknown, history. Although it is well-known that the Web bears some striking similarity to Vannevar Bush's 'Memex' idea from 1945 (Bush, 1945), the Web is itself usually thought more of as a technological innovation rather than an intellectually rich subject matter such as artificial intelligence or cognitive science. However, the Web's heritage is just as rich as artificial intelligence and cognitive science, and can be traced back to the same roots, namely the 'Man-Machine Symbiosis' project of Licklider (1960). The 'Man-Machine Symbiosis' project gave birth to two streams of research. The first strand is that of artificial intelligence done in the spirit of McCarthy, Minsky, and others involved in the original Dartmouth proposal (McCarthy et al., 1955). However, there exists another lesser-known strand of research, the work on 'human augmentation' exemplified by the work of Engelbart that eventually gave us both the mouse and the Internet (1962). Human augmentation, instead of hoping to replicate human intelligence as artificial intelligence did, only thought to enhance it. The Web itself is a descendant of Engelbart's vision, and this historical trajectory leading from Licklider to the creation of the

Web, is detailed in the following sections.

2.2 The Man-Machine Symbiosis Project

The first precursor to the Web was glimpsed, although never implemented, by Vannevar Bush. For Bush, the primary barrier to increased productivity was the lack of an ability to easily recall and create records, and Bush saw in microfiche the basic element needed to create what he termed the "Memex," a system that lets any information be stored, recalled, and annotated through a series of "associative trails" (1945). The Memex would lead to "wholly new forms of encyclopedias with a mesh of associative trails," a feature that became the inspiration for "linking" in hypertext (Bush, 1945). However, Bush could not implement his vision on the analogue computers of his day.

The Web had to wait for the invention of digital computers and networks, both of which bear some debt to the work of J.C.R. Licklider, a disciple of Norbert Wiener (Licklider, 1960). Wiener thought of feedback as an overarching principle of organization in any science, and one that was equally universal among humans and machines (1948). Licklider expanded this notion of feedback loops to a vision of low-latency feedback between humans and digital computers. The intellectual project of 'Man-Machine Symbiosis' is distinct and prior from cognitive science and artificial intelligence, both of which hypothesize that the human mind can be construed as either computational itself or even implemented on a computer. Licklider held that while the human mind itself might not be computational (although Licklider cleverly remained agnostic on that particular gambit), the human mind was definitely *complemented* by computers. As Licklider himself put it, "The fig tree is pollinated only by the insect Blastophaga grossorun. The larva of the insect lives in the ovary of the fig tree, and there it gets its food. The tree and the insect are thus heavily interdependent: the tree cannot reproduce without the insect; the insect cannot eat without the tree; together, they constitute not only a viable but a productive and thriving partnership. This cooperative 'living together in intimate association, or even close union, of two dissimilar organisms' is called symbiosis. The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today" (1960). The goal of 'Man-Machine Symbiosis' is then the enabling of reliable coupling between the humans and their 'external' information as given in digital computers. To obtain this

coupling, the barriers of time and space needed to be overcome so that the symbiosis could operate as a single process.

The ‘Man-Machine Symbiosis’ project was not merely an philosophical project, but an engineering project. In order to provide the funding needed to assemble what Licklider termed his “galactic network” of researchers to implement the first step of the project, Licklider became the institutional architect of the Information Processing Techniques Office at the Advanced Research Projects Agency (ARPA) (Waldrop, 2001). Licklider first tackled the barrier of time. Early computers had large time lags in between the input of a program to a computer on a medium such as punch-cards and the reception of the program’s output. This lag could then be overcome via the use of time-sharing, taking advantage of the fact that the computer, despite its centralized single processor, could run multiple programs in a non-linear fashion. Instead of idling while waiting for the next program or human interaction, in moments nearly imperceptible to the human eye, a computer would share its time among multiple humans (McCarthy, 1992).

Douglas Engelbart had independently generated a proposal for a “Human Augmentation Framework’ that shared the same goal as the ‘Man-Machine Symbiosis’ project of Licklider, although it differed by placing the human at the center, focusing on the ability of the machine to extend to the human user, while Licklider imagined a more egalitarian partnership between humans and digital computers (1962). This focus on human factors led Engelbart to the realization that the primary reason for the high latency between the human and the machine was the interface of the human user to the machine itself, as a keyboard was at best a limited channel. After extensive testing of what devices enabled the lowest latency between humans and machines, Engelbart invented the mouse and other, less successful interfaces, like the one-handed ‘chord’ keyboard (Waldrop, 2001). By employing these interfaces, the temporal latency between humans and computers was decreased even further.

2.3 The Internet

The second barrier to be overcome was space, so that any computer should be accessible regardless of its physical location. The Internet “came out of our frustration that there were only a limited number of large, powerful research computers in the country, and that many research investigators who should have access to them were geographically separated from them” (Leiner et al., 2003). Licklider’s lieutenant Bob

Taylor and his successor Larry Roberts contracted out Bolt, Beranek, and Newman (BBN) to create the Interface Message Processor, the hardware needed to connect the various time-sharing computers of Licklider's "galactic network" that evolved into the ARPANet (Waldrop, 2001). While BBN provided the hardware for the ARPANet, the software was left undetermined, so an informal group of graduate students constituted the Internet Engineering Task Force (IETF) to create software to run the Internet (Waldrop, 2001).

The IETF has historically been the main body that creates the protocols that run the Internet. It still maintains the informal nature of its foundation, with no formal structure such as a board of directors, although it is officially overseen by the Internet Society. The IETF informally credits as their main organizing principle the credo "We reject kings, presidents, and voting. We believe in rough consensus and running code" (Hafner and Lyons, 1996). Decisions do not have to be ratified by consensus or even majority voting, but require only a rough measure of agreement on an idea. The most important product of these list-serv discussions and meetings are IETF RFCs (Request for Comments) which differ in their degree of reliability, from the unstable 'Experimental' to the most stable 'Standards Track.' The RFCs define Internet standards such as URIs and HTTP (Berners-Lee et al., 1996, 2005). RFCs, while not strictly academic publications, have a *de facto* normative force on the Internet and therefore on the Web, and so they will be referenced considerably throughout this thesis.

Before the Internet, networks were assumed to be static and closed systems, so one either communicated with a network or not. However, early network researchers determined that there could be "open architecture networking" where a meta-level "internetworking architecture" would allow diverse networks to connect to each other, so that "they required that one be used as a component of the other, rather than acting as a peer of the other in offering end-to-end service" (Leiner et al., 2003). In the IETF, Robert Kahn and Vint Cerf devised a protocol that took into account, among others, four key factors, as cited below (Leiner et al., 2003):

1. Each distinct network would have to stand on its own and no internal changes could be required to any such network to connect it to the Internet.

2. Communications would be on a best effort basis. If a packet didn't make it to the final destination, it would shortly be retransmitted from the source.

3. Black boxes would be used to connect the networks; these would later be called gateways and routers. There would be no information retained by the gateways

about the individual flows of packets passing through them, thereby keeping them simple and avoiding complicated adaptation and recovery from various failure modes.

4. There would be no global control at the operations level.

In this protocol, data is subdivided into ‘packets’ that are all treated independently by the network. Data is first divided into relatively equal sized packets by TCP (Transmission Control Protocol), which then sends the packets over the network using IP (Internet Protocol). Together, these two protocols form a single protocol, TCP/IP (Cerf and Kahn, 1974). Each computer is named by an Internet Number, a four byte destination address such as 152.2.210.122, and IP routes the system through various black-boxes, like gateways and routers, that do not try to reconstruct the original data from the packet. At the recipient’s end, TCP collects the incoming packets and then reconstructs the data.

The Internet connects computers over space, and so provides the physical layer over which the “universal information space” of the Web is implemented. However, it was a number of decades before the latency of space and time became low enough for the Web to become not only universalizing in theory, but universalizing in practice. An historical example of attempting a Web-like system before the latency was acceptable would be the NLS (oNLine System) of Engelbart (1962). The NLS was literally built as the second node of the Internet, the Network Information Center, the ancestor of the domain name system. The NLS allowed any text to be hierarchically organized in a series of outlines with summaries, giving the user freedom to move through various levels of information and link information together. The most innovative feature of the NLS was a journal for users to publish information in and a journal for others to comment upon, a precursor of blogs and wikis (Waldrop, 2001).

However, Engelbart’s vision could not be realized on the slow computers of his day. Although time-sharing computers reduced temporal latency on single machines, too many users sharing a single machine made the latency unacceptably high, especially when using an application like NLS. Furthermore, his zeal for reducing latency made the NLS far too difficult to use, as it depended on obscure commands that were far too complex for the average user to master within a reasonable amount of time (Bardini, 2000). It was only after the failure of the NLS that researchers at Xerox PARC developed the personal computer, which by providing each user their own computer reduced the temporal latency to an acceptable amount (Waldrop, 2001). When these

computers were connected with the Internet and given easy-to-use interfaces as developed at Xerox PARC, both temporal and spatial latencies were made low enough for ordinary users to access the Internet. This convergence of technologies, the personal computer and the Internet, is what allowed the Web to be implemented successfully and enabled its wildfire growth, while previous attempts like NLS were doomed to failure as they were conceived before the technological infrastructure to support them had matured.

2.4 The Modern World Wide Web

Perhaps due to its own anarchic nature, the IETF had produced a multitude of incompatible protocols such as FTP (File Transfer Protocol) and Gopher (Postel and Reynolds, 1985; Anklesaria et al., 1993). While protocols could each communicate with other computers over the Internet, there was no universal format to identify information regardless of protocol. One IETF participant, Tim Berners-Lee, had the concept of a "universal information space" which he dubbed the "World Wide Web" (1992). His original proposal to his employer CERN brings his belief in universality to the forefront, "We should work towards a universal linked information system, in which generality and portability are more important than fancy graphics and complex extra facilities" (Berners-Lee, 1989). The practical reason for Berners-Lee's proposal was to connect the tremendous amounts of data generated by physicists at CERN together. Later as he developed his ideas, Berners-Lee came into direct contact with Engelbart, who encouraged him to continue with the idea of the Web despite his academic work being rejected at conferences like ACM Hypertext 1991.[1]

In the IETF, Berners-Lee, Fielding, Connolly, Masinter, and others spear-headed the development of URIs (Universal Resource Identifiers), HTML (HyperText Markup Language) and HTTP (HyperText Transfer Protocol). By being able to reference anything with equal ease due to URIs, a web of information would form based on "the few basic, common rules of 'protocol' that would allow one computer to talk to another, in such a way that when all computers everywhere did it, the system would thrive, not break down" (Berners-Lee, 2000). The Web is a *virtual space for naming information* built on top of the physical infrastructure of the Internet that could move bits around, and the Web was built through specifications that could be implemented by anyone, "What was often difficult for people to understand about the design was that

[1] Personal communication with Berners-Lee.

there was nothing else beyond URIs, HTTP, and HTML. There was no central computer 'controlling' the Web, no single network on which these protocols worked, not even an organization anywhere that 'ran' the Web. The Web was not a physical 'thing' that existed in a certain 'place.' It was a 'space' in which information could exist" (Berners-Lee, 2000).

The very idea of a *universal* information space seemed at least ambitious, if not *de facto* impossible, to many. The IETF rejected Berners-Lee's idea that any identification scheme could be universal. In order to get the initiative of the Web off the ground, Berners-Lee surrendered to the IETF and changed the name of his universal naming system from *Universal Resource Identifiers* (URIs) to *Uniform Resource Locators* (URLs) (Berners-Lee, 2000). The Web begin growing at a prodigious rate once the employer of Berners-Lee, CERN, released any intellectual property rights they had to the Web. The growth of the Web increased even more dramatically after Mosaic, the first graphical browser, was released. However, browser vendors started adding supposed 'new features' that soon led to a 'lock-in' where certain sites could only be viewed by one particular corporate browser. These 'browser wars' began to fracture the rapidly growing Web into incompatible information spaces, thus nearly defeating the proposed universality of the Web (Berners-Lee, 2000).

Berners-Lee in particular realized it was in the long-term interest of the Web to have a new form of standards body that would preserve its universality by allowing corporations and others to have a more structured contribution than possible with the IETF. With the informal position of merit Berners-Lee had as the supposed inventor of the Web (although he freely admits that the invention of the Web was a collective endeavor), he and others constituted the World Wide Web Consortium (W3C); a non-profit dedicated to "leading the Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web" (Jacobs, 1999). In the W3C, membership was open to any organization, commercial or non-profit organization. Unlike the IETF, W3C membership came at a considerable membership fee. The W3C is organized as a strict representative democracy, with each member organization sending one member to the Advisory Committee of the W3C, although decisions technically are always made by the Director, Berners-Lee himself. By opening up a "vendor neutral" space, companies who previously were interested primarily in advancing the technology for their own benefit could be brought to the table. The primary product of the World Wide Web Consortium is a W3C Recommendation, a standard for the Web that is explicitly voted on and endorsed by the W3C membership. W3C

Recommendations are thought to similar to IETF RFCs, with normative force due to the degree of formal verification given via voting by the W3C Membership. A number of W3C Recommendations have become very well known technologies, ranging from the vendor-neutral versions of HTML (Raggett et al., 1999), which stopped the fracture of the universal information space at the hands of the browser wars, to XML, which has become a prominent transfer syntax for almost any type of data (Bray et al., 1998). This thesis will cite W3C Recommendations when appropriate, as these are one of the main normative documents that define the Web. With IETF RFCs, these normative standards collectively define the foundations of the Web. It is by agreement on these standards that the Web functions as a whole. However, the rough-and-ready process of the IETF and even W3C has led to a terminological confusion that must be sorted in order to inspect the problem of how URIs can identify things outside the Web itself.

Chapter 3

Philosophical Prolegomenon

Philosophy, more rigorously understood, is the discipline that consists of creating concepts. **Gilles Deleuze and Felix Guattari** (1991)

A major focus of this thesis is to use terminology from philosophy of computation, language, and the mind to produce a small set of fairly well-defined terms that we can use to express the question: What does a URI refer to? Afterwards, we use these terms to determine what the boundaries of the Web are in Chapter 4 and to clarify the Semantic Web in Chapter 5.

For the sake of brevity we will not in this chapter explore all the nuances and consequences arising from our admittedly broad-sweeping terminology. This is unfortunate, as there is just not enough space to address, much less defuse, all possible counter-arguments. In this manner, this chapter will be decidedly non-philosophical, although we will attempt to mitigate this problem by at least providing references to well-known philosophers from whom we have adopted our terminology, although often we will use their terms in a slightly-modified form so that the terminology may fit the problem at hand. The theoretical framework and terminological definitions given in this chapter provide the foundation for the entire thesis, coming to a head in our proposed solution to the issues of reference and representation on the Semantic Web in Chapter 8. While this chapter may not appear directly relevant to the Web, the philosophical terminology established here will be used to discipline the wild and unruly terminology of Web architecture in the next chapter. Again, we claim neither that our historical and philosophical foundations of Web architecture are complete and systematic, but just systematic and complete enough to pose and solve our hypothesis, without either the question or our solution using vacuous terminology. Otherwise, the result will

be terminological confusion, causing any reader to fall into a conceptual swamp of undefined and fuzzy terms like ‘meaning’, ‘reference’, and ‘representation.’ We first explore the notion of ‘information’ at the heart of Berners-Lee’s definition of the Web as a ‘universal information space’ and then rebuild a notion of ‘digitality’ and finally ‘representation’ on top of our notion of information, since the Web is composed of not just any representations, but digital representations.

3.1 Preliminaries

On the surface a term like ‘representation’ seems to be what Brian Cantwell Smith calls “physically spooky,” since a representation can refer to something with which it is not in physical contact (Smith, 1995). This spookiness is a consequence of a violation of *common-sense* physics, since representations appear to have a non-physical relationship with things that are far away in time and space. This relationship of ‘aboutness’ or *intentionality* is often called ‘reference.’ While it would be premature to define ‘reference,’ a few examples will illustrate its usage: someone can think about the Eiffel Tower in Paris without being in Paris, or even having ever set foot in France; a human can imagine what the Eiffel Tower would look like if it were painted blue, and one can even think of a situation where the Eiffel Tower wasn’t called the Eiffel Tower. Furthermore, a human can dream about the Eiffel Tower, make a plan to visit it, and so on, all while being distant from the Eiffel Tower. Reference also works temporally as well as distally, for one can talk about someone who is no longer living such as Gustave Eiffel. Despite appearances, reference is not epiphenomenal, for reference has real effects on the behavior of agents. Specifically, one can remember what one had for dinner yesterday, and this may impact on what one wants for dinner today, and one can book a plane ticket to visit the Eiffel Tower after making a plan to visit it.

Can we get to the heart of this mystery at the heart of representation and other intentional terminology? The trick would be to define what precisely our common-sense notion of reference is, and to do this requires some terminological ground work while avoiding delving into amateur quantum physics. The terminology here is supposed to reconstruct rather carefully some common-sense demarcations in an uncontroversial yet broad manner so that these terms can deal with a suitably broad range of phenomena, including the Web. To pin the supposed ‘spookiness’ of reference down, we will introduce a few terms. A ***thing*** is a general-purpose term used to denote *events, objects, and proto-objects in a “patch of metaphysical flux,” where a thing can be defined*

by having some regularity in time and space that can distinguish it from other possible things (Smith, 1995). A ***regularity*** is *a lack of difference in time and space at a given level of abstraction*. We shall often use the term ***process*** interchangeably with things to evoke the dynamic and temporally unstable character of a 'thing.' We will also sometimes use the term ***system*** when we are emphasizing the fact *that one thing can also be, on a different level of abstraction, given as multiple things*. This can be considered a mere change of focus, for the term 'thing' emphasizes the everyday, solid, and static nature of the "metaphysical flux," while the term 'process' refers more to its dynamic aspect (Smith, 1995). *All things and processes* are the ***world***. There are generally two kinds of separation possible in processes in a relativistically invariant theory, a physical theory that obeys the rules of special relativity so that the theory looks the same for any constant velocity observer, as processes may be separated in time or space. *Things that are separated by time and space* are ***distal*** while *those things that are not separated by time and space are* ***proximal***. As synonyms for distal and proximal, we will use ***non-local*** and ***local***, or just ***disconnected*** and ***connected***. Although this may seem to be an excess of adjectives to describe a simple distinction, this aforementioned distinction will underpin our notions of representation and reference. In figures, local relationships will be marked with a dotted line, while distal (and so possibly referential) relationships are marked with the uniform bold line.

While a discussion about counterfactuals and causation is far beyond our scope, we will rely on the common-sense intuition that *if one thing is connected with another thing and a change in the former thing is followed by a change in the latter thing*, that former process may have ***caused*** the change in the latter process. In other words, the first thing is ***effective***, and *the other things that may be effected by a particular thing are within its* ***effective reach***. *Anything that appears to violate these common-sense intuitions about physics and causation* is ***spooky***, while anything that does not is ***non-spooky***. A property of the distal is that it is beyond effective reach; as Smith puts it, "distance is where no action is at" (1995). For example, a tourist hitting their toe on the Eiffel Tower has no immediate effect on someone in Edinburgh. With these preliminary terms in hand, we return to the topic of the Web.

3.2 Information, Encoding, and Content

The Web has been defined as a "universal information space" by Berners-Lee, and we will take this definition seriously and attempt to unravel it, in the hope that it will

provide clues on how we can define both ‘representation’ and ‘reference’ in a manner that can do justice to the Web (1992). The strategy to be employed is to inspect Berners-Lee’s evocative notion of the Web as a universal information space in order to provide a less complex notion of information that can serve as the foundation for building the more complex notion of representation. The first question to be answered then is the perennial question: What is information? Although we cannot comprehensively answer this question in full, we can sketch some crucial distinctions.

In order to make progress on defining the Web, we will have to reformulate the notion of information, taking inspiration from Shannon’s communication theory while allowing the central concept of information to be grounded in the wider philosophy of language. To rephrase, ***information*** is *whatever regularities held in common between two things*, a ***source*** and a ***receiver*** (Shannon and Weaver, 1963). To have something in common means to share the same regularities, e.g. parcels of time and space that cannot be distinguished at a given level of abstraction. This definition correlates with information being the inverse of the amount of ‘noise’ or randomness in a system, and the amount of information being equivalent to a reduction in uncertainty. This preservation or failure to preserve information can be thought of as the sending of a *message* between the source and the receiver over a channel. *Whether or not the information is preserved over time or space is due to the properties of a physical substrate* known as the ***channel***. The ***message*** *realizes on some level of abstraction the information*, so we will often call some particular message with some particular information an ‘information-bearing message.’ Already, information reveals itself to be not just a singular thing, but something that exists at multiple levels. In particular, we are interested in two more distinctions in information: that between abstraction and realization, and that between content and encoding.

The first distinction is between the information itself on a level of abstraction, and the particular realization of information. Information is often thought of as an abstraction, and this is true insofar as the same information can be realized by many possible messages. In order to cope with this, a distinction should be made between the information on a level of abstraction from any of the concrete realizations themselves that embody the information at a given juncture in space-time. To use an example, Daniel in Paris (the source) is trying to send a message to Amy (the receiver), a secretary in Boston, that one of her fellow workers, Ralph, has won a trip to the Eiffel Tower. Daniel can send this message in a variety of realizations: e-mail, a letter in the post, or even via a friend who happens to be passing through Boston. The information itself

is just the precise physical regularity at a level of abstraction, and these regularities can be embodied by many different possible messages, but these messages are not arbitrary, but must have a certain ability to preserve the regularity – so in the case of Daniel, it's unlikely he could convey his message from Paris to Boston using smoke signals. It would simply not reach the receiver in any recognizable form. So, a ***level of abstraction*** is *certain physical differences and regularities that can be recognized by an agent and so may have a causal effect on the agent*. For example, given a handwritten letter in English, one can focus on the low-level of abstraction, such as the details of the various pen-strokes and the texture of the paper, or progressively higher levels of abstraction, such as recognizing letters in an alphabet, words, or sentences, or even some larger units of discourse that express the thought 'Ralph won a ticket to Paris.' To say that some thing realizes the information is of course a ***realization*** of the information, which is a *the physical thing that realizes the regularities of the information due to its local characteristics*, just like a particular information-bearing message but more broadly construed. The concrete voltages down the wire realize an e-mail message, as does a physical book realize some sentences in English. It is common practice to elide various levels of abstraction and just talk about information, but often it is useful to pull apart the abstract pattern of regularities from those physical things in the world that realize them. Since the term 'information' is used indiscriminately to refer to information on a level of abstraction and the realization of some abstract information, we will use the term *information realization* or just *realization* when discussing a particular realization of information and use the term ***abstract information*** on the rare occasion when we wish to emphasize *information on a level of abstraction regardless of its particular realization*. When the term 'information' by itself is used, we are referring to both abstract information and any of its particular realizations.

The second distinction is not as obvious as the distinction between abstract information and its realization: the distinction between the content and encoding of information. Shannon's theory deals with finding the optimal encoding and size of channel so that the message can be guaranteed to get from the sender to the receiver (Shannon and Weaver, 1963). Yet, how can an encoding be distinguished from the content of information itself? Goodman defines what we would call an encoding as a series of marks, where a ***mark*** *is a physical characteristic* ranging from marks on paper one can use to discern alphabetic characters to ranges of voltage that can be thought of as bits (1968). To be reliable in conveying information, an encoding should be physically "differentiable" and thus maintain what Goodman calls "character indifference" so that

(at least within some context) each character (characteristic) can not be mistaken for another character. So, an ***encoding*** is *a set of precise regularities that can be realized by the message*. Encodings are usually given these regularities in virtue of being in a language, which is explicated in Section 3.3.

Is our distinction of 'encoding' re-stating the difference between abstract information and realization? It is not. Although it would seem that information becomes somehow concrete within a particular region of space-time when it is encoded, on closer inspection, an encoding can still exist on a level of abstraction without being concretely realized in space-time. The term 'Eiffel Tower' carries information in an encoding, but it is realized when some speaker uses it in an actual utterance. The text of *Moby Dick* can be thought of as abstract information, a story about a white whale. The text of *Moby Dick in English* is an encoding of the abstract information of *Moby Dick*, with precise regularities given by the *very letters* of the language. The content of the novel *Moby Dick* could be encoded in a different language, like French, and the precise regularities that convey the *same* information at a level of abstraction could be given by *different* physical characteristics and so *different encodings*. In the case of French versus English, different words and other linguistic nuances would exist, but the information would – at a level of abstraction, since obviously there are nuances possible in French that do not exist in English, and vice versa – be the same. So even the text of *Moby Dick* in a particular encoding like English exists at a level of abstraction, as it could be realized in multiple things in space-time, as a copy in English of *Moby Dick* could be realized by two different physical books, one in Edinburgh and the other in Jakarta. In fact, these realizations could also be quite different, such as a realization of Moby Dick in English as a web-page going down the wire as a particular set of voltages at a given time, and as a particular book on someone's bookshelf.[1]

There is more to information than encoding. Shannon's theory does not explain the notion of information fully, since giving someone the number of bits that a message contains does not tell the receiver *what* information is encoded. Shannon explicitly states that "the fundamental problem of communication is that of reproducing at one

[1]There certainly vast metaphysical difficulties that we are purposefully ignoring in our distinction between realization and abstract information. Namely, do not *realizations themselves* exist on a level of abstraction? To some extent this can be thought of as true: is a particular copy of *Moby Dick* on my shelf today the same realization tomorrow? These metaphysical conundrums can have their Gordian knots cut in a straightforward manner: A realization is composed of locally-connected causal regularities, and how this realization is thought of as varying over space-time is irrelevant for the time being, as long as the realization from one moment to another, or from one portion of space to another, is connected to its former self.

point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem" (1963). He is correct, at least for his particular engineering problem. However, Shannon's use of the term 'information' is for our purposes the same as the 'encoding' of information, but a more fully-fledged notion of information is needed. Many intuitions about the notion of information have to deal with not only how the information is encoded or how to encode it, but what a particular message is about, the *content* of an information-bearing message. 'Content' is a term we adopt from Israel and Perry, as opposed to the more confusing term 'semantic information' as employed by Floridi and Dretske (Israel and Perry, 1990; Dretske, 1981; Floridi, 2004).

While the notion of an information's content is hard to pin down, it is easy to illustrate. Just determining that a single employee out of eight won the lottery requires at least a three bit encoding and does not tell Amy which employee in particular won the lottery. Only a particular three bits will tell Amy precisely who won the lottery. Shannon's theory only measures how many bits are needed to tell Amy precisely who won. After all, the false message that another office-mate Sandro won a trip to Paris is also three bits. Yet content is not independent of the encoding, for content is conveyed by virtue of a particular encoding and a particular encoding imposes constraints on what content can be sent (Shannon and Weaver, 1963). Let's imagine that Daniel is using a code of bits specially designed for this problem, rather than natural language, to tell Amy who won the free plane ticket to Paris. The content of the encoding `001` could be Ralph while the content of the encoding `010` could be Sandro. If there are only two possible bits of information and all eight employees need one unique encoding, Daniel cannot send a message specifying which friend got the trip since there aren't enough options in the encodings to go round. An encoding of at least three bits is needed to give each employee a unique encoding. If `01` has the content that 'either Sandro or Ralph won the ticket' the message has not been successfully transferred if the purpose of the message is to tell Amy *precisely* which employee won the ticket.

One of the first attempts to formulate a theory of informational content was due to Carnap and Bar-Hillel (1952). Their theory attempted to bind a theory of content closely to first-order predicate logic, and so while their "theory lies explicitly and wholly within semantics" they explicitly do not address "the information which the sender intended to convey by transmitting a certain message nor about the information

a receiver obtained with a certain message," since they believed these notions could eventually be derived from their formal apparatus (Carnap and Bar-Hillel, 1952). Their overly restrictive notion of the content of information as logic did not gain widespread traction, and neither did other attempts to develop alternative theories of information such as that of Donald McKay (1955). In contrast, Dretske's *semantic theory of information* defines the notion of content to be compatible with Shannon's information theory, and his notions have gained some traction within the philosophical community (Dretske, 1981). To Dretske, the content of a message and the amount of information as studied by Shannon are different, for "saying 'There is a gnu in my backyard' does not have more content than the utterance 'There is a dog in my backyard' since the former is, statistically, less probable" (1981). According to Shannon, there is more information in the former case precisely because it is less likely than the latter and so would require more bits to encode (Dretske, 1981). So while information that is less frequent may require a larger number of bits in an encoding, the content of information should be viewed as separable if compatible with Shannon's information theory, since otherwise one is led to the "absurd view that among competent speakers of language, gibberish has more meaning than semantic discourse because it is much less frequent" (Dretske, 1981). Shannon and Dretkse are talking about distinct, but intertwined, notions that should be separated, namely the distinction between encoding and content.

Is there a way to precisely define the content of a message? Dretske defines the content of information as "a signal r carries the information that s is F when the conditional probability of s's being F, given r (and k) is 1 (but, given k alone, less than 1). k is the knowledge of the receiver" (1981). To simplify, the ***content*** of any information-bearing message is *whatever is held in common between the source and the receiver as a result of the conveyance of a particular message*. While this is similar to our definition of information itself, it is different. Information can measure the total in common between a source and receiver *simpliciter*. For example, two distal humans can share quite a lot in common, and so share information, despite never having conveyed a message between each other. The content is whatever is shared in common as a result of a *particular* message, such as the conveyance of sentence 'Ralph won a ticket to the Eiffel Tower.' The content of a message is called the "facts" by Dretske, (F). This content is conveyed from the source (s) successfully to the receiver (r) when the content can be used by the receiver with certainty, *and* that before the receipt of the message the receiver was not certain of that particular content. Daniel can only successfully convey the content that 'Ralph won a trip to Paris' if before receiving the message

Amy does not know that Ralph won the trip to Paris and after receiving the message Amy does know that fact. To communicate content successfully, both the source and receiver have to be using the same encoding scheme (bits, English, etc.) and the source has to encode the content relative to what the receiver already knows or capacities the receiver possesses. Thus, if Amy does not know who is specified by the term "Ralph" given by the encoding scheme, but only knows him as 'the guy with the black beard,' Daniel needs to explain in his message the additional fact that the 'fellow with the black beard at your office is Ralph.' However, we should interpret the term 'certainty' more loosely than Dretske would. Dretkse himself notes that information "does not mean that a signal must tell us everything about a source to tell us something," it just has to tell enough so that the receiver is now certain about the content within the domain (1981). Millikan rightfully notes that Dretske states his definition too strongly, for this probability of 1 is just an approximation of a statistically "good bet" indexed to some domain where the information was learned to be recognized (2004). For example, lightening carries the content that "a thunderstorm is nearby" in rainy climes but in an arid prairie lightning can convey a dust-storm. However, often the reverse is true, as the same content is carried by messages in different encodings, like the message from Daniel to Amy being encoded in either English or French.

In our example, the message that 'Ralph won a plane ticket to France' can be encoded in two different languages and still have the same relationship to content. *The relationship of an encoding to its content* is an ***interpretation***. The interpretation 'fills' in the necessary background left out of the encoding, and maps the encoding to some content. In our previous example using binary digits as an encoding scheme, a mapping could be made between the encoding `001` to the content of Ralph while the encoding `010` could be mapped to the content of Sandro. An interpretation requires an ***interpreter*** or *an agent that is capable of carrying out an interpretation from a particular encoding and a particular content*. The word 'interpretation' is probably one of the most embattled words, and an in-depth study of its usage far exceeds the scope of this thesis. Somewhat unusually, our usage of the term 'interpretation' is as a relationship between an interpreter and some encoding, not a first-order thing itself. This is done on purpose, in order to emphasize the fact that some *interpreting agent* is needed to actually make the interpretation from some encoding to content. When the word 'interpretation' is used as a noun, we mean the content given by a particular relationship between an agent and an encoding. Usual definitions of 'interpretation' tend to conflate these issues. In formal semantics, the word 'interpretation' often can be used

either in the sense of "an interpretation structure, which is a 'possible world' considered as something independent of any particular vocabulary" (and so any agent) or "an interpretation mapping from a vocabulary into the structure" or as shorthand for both (Hayes, 2004). The difference in use of the term seems somewhat divided by fields. For example, computational linguists often use "interpretation" to mean what Hayes called the "interpretation structure." In contrast, we use the term 'interpretation' to mean what Hayes called the "interpretation mapping," reserving the word 'content' for the "interpretation structure" or structures selected by a particular agent in relationship to some encoding. Also, this quick aside into matters of interpretation does not explicitly take on a formal definition of interpretation as done in model theory, although our general definition has been designed to be compatible with model-theoretic and other formal approaches to interpretation.

To uphold our requirement for physical non-spookiness, in order for an interpretation to take place, the interpreter and some realization of the encoding must be connected in some way, such as a human looking at bytes or a machine processing various voltages. In this manner, the examples of interpretation are almost always from particular information realizations in some particular encoding to some particular content. However, the relationship of interpretation is not bound to a particular realization of any information, but also functions at a level of abstraction as well, since obviously many particular realizations of the same abstract information can have the same interpretation. Imagine that Amy is bilingual, and speaks both French and English, so if Daniel had two messages, one in English and another in French, explaining that Ralph has a plane ticket, both messages would have the same interpretation to the same content. So, while information has to be realized concretely in order to be interpreted in a given message by an agent, as many messages can have the same interpretation across many agents, the interpretation is thought to be between the encoding and the content, even when the encoding is at a level of abstraction. So, the single sentence 'Ralph won a plane ticket to Paris' may have a single interpretation across many different utterances. However, if the agent and their background information changes, the interpretation may change, as obviously if the e-mail from Daniel was intercepted and read by some secret agent not at Ralph's office, obviously the secret agent may not know who Ralph is while Amy will.

The content of a particular message depends very much on the encoding scheme used by the interpreter. For example, one can interpret the encoding `11` as either the number eleven in the decimal encoding scheme, or the number three in the binary

encoding scheme. Unlike many others, including Dretske, we shall make no claims about the nature of information, interpretation, and truth, in particular if what appears to be 'false' information is really misinformation or pseudo-information. By remaining studiously neutral on this long-standing debate, our definition of information is suitably vague enough so that even encodings that are interpreted to be 'false' still count as information. For example, if Daniel was sending the message to Amy that Ralph had a free plane ticket to Paris as some sort of jest or lie, Amy could still decode and interpret the message, and by filling in normal background assumptions (as Dretske put it, the "channel assumptions") she might assume that the message was true (1981). Amy would still have an interpretation of the content of the message, it would just be different from Daniel's interpretation. In other words, information may always have an encoding and content and nothing forces some information realization to be interpreted to the same content by all interpreters.

Interestingly enough, this opens the door to the possibility of a sender sending an encoding to a receiver that lacks the necessary capacity to decode it. The encoding would not then have an interpretation to content. This would be the standard definition of ***data***, which is *information without an interpretation*. Our definition works well with other 'textbook' definitions of data and information, such as that of Davis and Olson, which states that "information is data that has been processed into a form that is meaningful to the recipient" (1985). This does not mean that the encoding does not possibly have an interpretation, but at that given moment it cannot be interpreted. One example would be if the message from Daniel that Ralph had won the plane ticket had been delivered via e-mail in French. While Amy could have been aware of some very limited aspects of the e-mail (such as the time sent and the sender), she would lack the necessary knowledge of French to decode the message's content and so to have an interpretation of the message. In this manner, the e-mail from Daniel, while having a definite interpretation for French speakers, would lack an interpretation for Amy. To Amy, the message would just be data. Of course, Amy could learn French and eventually read the message, or send it to a machine-translation program, or ask a French speaker to translate the message for her, and so could eventually transform the encoding from data to information. One can also imagine cognitive constraints leading to a lack of an interpretation. For example, the volume of data gathered by modern telescopes is absolutely enormous, so large that much of it lies as uninterpreted reams of data rather than information for scientists, as it is beyond a single human to interpret this data, and even groups of humans trying to interpret it in a distributed manner are

still struggling to catch up with the volume of data produced by the telescopes.

These terms are all illustrated in Figure 3.1. A source (Amy) is sending a receiver (Daniel) a message. The information-bearing message realizes some particular encoding such as a few sentences in English and a picture of the Eiffel Tower, and the content of the message can be interpreted to be about the actual Eiffel Tower.

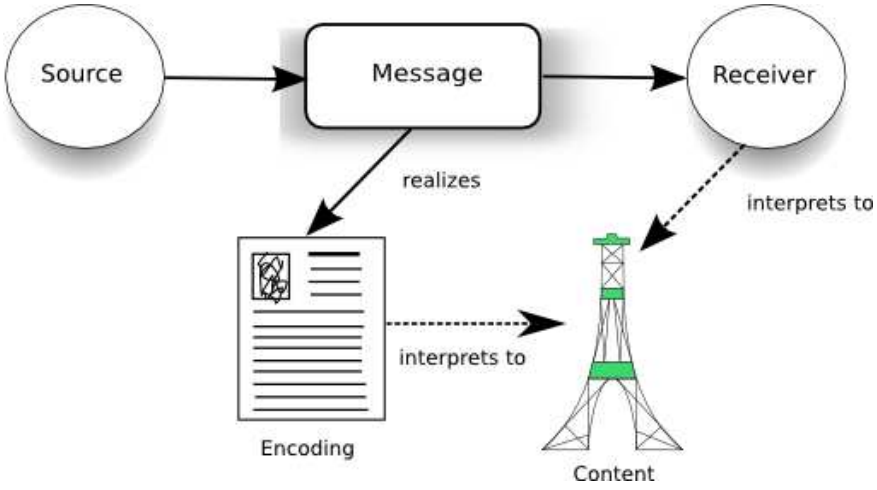

Figure 3.1: Information, Encoding, Content

Information, which appeared so simple, is now revealed to be a multi-faceted phenomenon. To summarize, information is what is held or could be held in common between a sender and a receiver. Information is always thought of at a level of abstraction, and so abstract information can be realized concretely by some realization, like a particular message. Information, on both the level of abstract information and a particular realization, has two sides: encoding and content. The encoding is the precise regularities that can convey the information in a particular message, while the content is what is in common between the receiver and the sender as a result of the conveyance of a particular message. The thought 'Ralph won a plane ticket to Paris' is the content, given an encoding in English by Daniel, and realized as some bits sent over the wire to Amy. These notions of encoding and content are not strictly separable, which is why they together compose the notion of information. An updated famous maxim of Hegel could be applied to the new-fangled concept of information: There is no encoding without content, and no content without encoding (1959). In a similar vein, while we can *imagine* there being information without any realizations, we only *know* information through its concrete realizations.

3.3 Meaning and Purpose

The notion of interpretation implies the transfer of an encoding and an act of the interpreter that relates that encoding to content, nothing more. When Daniel sends Amy the e-mail to tell her Ralph had a plane ticket to Paris, Amy interpreted the message by filling in various background information, and so determining that Ralph at her office has a plane ticket to Paris. Amy has successfully interpreted the message. The effect upon an agent of an interpretation of some encoding is difficult to visualize, and one attempt that resonates is the notion of *assertoric content* given by Dummett(1973). Ignoring his larger project, we can simply say one way to tell if an agent has interpreted an encoding to some content is that the agent would 'assent' to various questions about this content. So, if Daniel asked Amy if she got the message about Ralph, minimally she should assert that she did, and if she does not, then perhaps she did not get the message.

Yet, if Amy merely sat at her desk, content in her knowledge, but did not tell Ralph, then *something* would have gone awry from Daniel's standpoint. Obviously, the point of sending a message is for the information to have some causal effect on the agent, which would be manifested in the behavior of the agent. This *causal effect of information on agents*, often demonstrated by behavior, is the ***meaning*** of the information.[2] So, the meaning of the message for Amy that 'Ralph won the plane ticket' is precisely the behavior exhibited by Amy, such as her getting up from her desk and telling Ralph verbally that he has a plane ticket to Paris. The meaning of information is quite manifold, as it may cause the behavior of multiple agents in what are called *informational links* by Gareth Evans (1982). For example, when Ralph hears from Amy he won a plane ticket, he may go to book a hotel and tell his wife; these actions are Ralph's behavioral manifestations, the meaning of the message for Ralph, that is caused by Amy conveying the message. Since the message from Amy to Ralph realized the same abstract information that the message from Daniel to Amy realized, the behavior of both Amy and Ralph is created by the same abstract information, even if there were different distinct messages (an e-mail from Daniel, an utterance from Amy) conveying this information. The meaning of a piece of information, even a single message, may and usually does spread beyond a single agent receiving a single message. So, one can

[2]This does not necessarily mean that the receiver has changed in some observable manner, instead, the effect of the message on the receiver may cause the receiver to stay the same. This would be exemplified when measuring the degradation of information on a hard-drive, where the amount of information preserved from the selfsame hard-drive at one moment in time to another is considered the message.

legitimately use the term ‘meaning’ both in the context of a single realization or the more abstract information that can be realized by multiple realizations.

However, what if Amy doesn’t act *appropriately* when receiving the message? What if upon receiving the message, she simply deletes it? The ***purpose*** of information is *the intended meaning of information*, often given by the intended behavior of the receiver intended by the sender of a message. The sender of the message, Daniel, wants Ralph to receive it – that is the purpose of his original e-mail to Amy. Information often has a ‘purpose’ that is beyond its particular content. For example, Daniel could be trying to reward Ralph for his astounding performance in his job, and believes that a vacation to Paris may ensure his future good behavior at work. Ralph may not be able to deduce any of this from the content of the message he receives. There are numerous reasons for the purpose also being at odds with the meaning of the message; the information may not have the same meaning for the sender as it does for the receiver, and so the sender may be sending a message that causes meaningful behavior for the receiver that the sender did not predict. A single sentence like ‘Police!’ might always have the same interpretation to content (i.e. to a nearby policeman) but it would be radically different in both meaning and purpose if it was muttered by a thief who had just managed to pick-pocket a tourist than if the exact same expression was used by the tourist who had just been pick-pocketed. This shows how meaning is essentially related to the wider context of the utterance, as explored in natural language by the theory of ‘speech acts’ of Austin and Searle (Searle, 1969). Furthermore, the meaning of a message may include the attempt by the receiver to create some future behavior in the sender. Also, *when an agent is trying to determine some information in order to direct its meaningful behavior*, the agent can be said to have an ***information need***. Everything from a frog wandering around looking for flies in its environment to a student asking a teacher a question or an agent typing in search terms into a Web search engine count as information needs.

A purpose is inherently *normative*, i.e. that information *should*, but does not necessarily have to, fulfill its purpose to produce a particular meaning. This normativity could be grounded out in a number of different ways, but one prominent story is that all normativity must ultimately be grounded out in evolution, so fulfilling Dennett’s condition that “all normativity does ride on Darwin’s coat-tails” (Smith, 2002b). This is an important aspect, because it tells a story about purpose of information even when the sender is not another human agent, but the environment at large. For example, the message given by a frog’s retina that a large dark spot is nearby may cause the mean-

ingful behavior of tongue-flicking, since the tongue-flicking accomplishes the purpose of the frog feeding itself. In this way, Millikan grounds normativity out in terms on whether or not some information fulfills a "proper function" (1984). While the notion of a proper function is too large a subject to analyze thoroughly here, Millikan summarizes her more extended presentation into the evolutionary *Language, Thought, and Other Biological Categories* (1984) by saying " A thing's proper functions are effects which, in the past, have accounted for selection of its ancestors for reproduction, or accounted for selection of things from which it has been copied, or for selection of ancestors of the mechanisms" (Millikan, 2000). So, for example, the function of the eye to blink was selected because it protected the eyes from harm and so increased the survival of eye-blinking species. She later extends this definition to deal not just with natural selection of genes, but mimetic selection, where imitation counts as a form of reproduction, and in this way accounts for the extension of eye-blinking as a signal of recognition to the complex use of language (Millikan, 2004). Also, many things spread, especially by imitation, regardless of any proper function. As Millikan notices, "Many conventions seem to have no functions. They seem to proliferate only because people are creatures of habit, or unthinking conformists, or because they venerate tradition, and so forth" (Millikan, 2000). From this we can get a definition of ***convention***, such as choosing to drive on the right side of the road as opposed to the left, as the *use of a thing based purely on previous history*, without regard to imitation or natural selection. While a proper function is a natural purpose, many technological artifacts have an 'unnatural' purpose, particularly those designed in some laboratory or by some enthusiasts and not yet released 'into the wild' to suffer the travails of selection either by nature or the market. This is the purpose for which an artifact has been designed, which it may or may not succeed. In many cases, it is hard to even detect the purpose of some particular information, and the connection to evolution will be vague at best. In most of the examples we are dealing with, our notion of success is straightforward; the message to Amy that Ralph won the plane ticket is successful if Amy receives the content of the message, and this can be detected by Amy acting appropriately, such as when she tells Ralph that he has a plane ticket to Paris. Without the ability to accurately receive and transmit messages, one would assume that the species would be less likely to survive, and technology such as sending e-mails is successful insofar as it provides a benefit to its users over, say, carrier pigeons. As Andy Clark puts, it "by seeing tools as entities with their own selective histories" we can understand what Terrence Deacon calls the "flurry of adaption...going on outside the brain" (Clark, 2002; Deacon, 1997).

Despite Dretske's use of terms like "certainty" and "knowledge," we can use our story about information in ways that apply to technology such as computers whose epistemic properties are even more uncertain than those of humans. The successful conveyance of a message requires that its regularity is preserved over some channel so that the message is capable of evoking the correct and purposeful meaningful behavior from agents. What Dretske calls "knowledge" are the regularities already present in the system that may contribute to the information being successfully conveyed between agents. So one could easily replace the natural language message about a free trip to Paris between two humans to be a message to book an aeroplane ticket for Ralph from one dumb server to another over the Internet. For this to be successful, the servers must share the same encoding schemes so that the content of the message can be decoded. These computers may not interpret the content of the encoding of the message in the same manner that a human does – since the computers obviously do not know that Ralph is, say, human – but they interpret the message nonetheless, and the sign of this interpretation is that the message has some meaningful physical effect upon the machine, causing it to send other messages to other machines that eventually results in a plane ticket being printed for Ralph. However, the evoked behavior is not arbitrary, just as an interpretation is not arbitrary. If the plane ticket given by Daniel sends Ralph to Berlin, something has gone amiss in the computer's interpretation of the booking, and its meaningful behavior is no longer in line with the purpose of the message.

3.4 Language and Models

The encodings and content of information do not in general come in self-contained bundles, with each encoding being interpreted to some free-standing propositional content. Instead, encodings and content come in entire interlocking informational systems. One feature of these systems is that encodings are layered inside of each other and content is also layered upon other content. The perfect example would be an English sentence in an e-mail message, where a series of bits are used to encode the letters of the alphabet, and the alphabet is then used to encode words. Likewise, the content of a sentence may depend on the content of the words in the sentence. When this happens, one is no longer dealing with a simple message, but some form of language. A ***language*** can be defined as *a system in which information is related to other information systematically*. In a language, this is a relationship between how the encoding of some information can change the interpretation of other encodings. Messages always have

encodings, and usually these encodings are part of languages. To be more brief, information is *encoded in* languages. The relationships between encodings and content are usually taken to be based on some form of (not necessarily formalizable or even understood) rules. If one is referring to *a system in which the encoding of information is related to other encodings systematically*, then one is talking about the ***syntax*** of a language. If one is referring to *a system in which the content of information is related to other content systematically*, then one is referring to the ***semantics*** of the language. Particular encodings and content then *are accepted by* the syntax and semantics of a language respectively.

Also, we do not restrict our use of the word 'language' to primarily linguistic forms, but use the term 'language' for anything where there is a systematic relationship between syntax and (even an informal) semantics. One such investigation into non-linguistic languages is Nelson Goodman's *Languages of Art* (1968). Although our examples so far have been in natural language, our definition of language is purposefully neutral regarding languages for humans (or even possibly languages for other animals) and 'formal' languages for machines such as programming languages for computers. There are ***iconic languages*** *based on images* and ***natural languages*** based *on human linguistic expressions*, as well as ***formal languages*** *with an explicitly defined syntax and possibly model-theoretic semantics*, and so the purpose of these formal languages can be interpretation by computers. Many computer languages not considered to be programming languages are languages insofar as they have some normative or even informal interpretation, such as HTML. Furthermore, due to some bias against computer languages actually being first-class languages, sometimes the term *format* is a synonym for computer-based language, often one that cannot directly execute as a program. Lastly, just as encodings and content may be embedded in each other to form a language, languages themselves may be embedded in each other to form new languages. *A language embedded as a subset of another language* is a ***dialect*** or ***vocabulary*** of the language. Many machine languages like XML have as their primary purpose the expression of other dialects (Bray et al., 1998).

A *particular message in a language* is an ***expression*** of the language. The lower-level of a language can be ***terms***, *regularities in marks*, that may or may not have their own interpretation, such as the words or alphabet. *Any combination of terms that is valid according to the language's syntax* is a ***sentence*** in the language, and *any combination of terms that has an interpretation to content according to the language's semantics* is a ***statement*** in the language. In this way, marks form the syntax of a

language. The relationship between semantics and syntax can be straightforward or only vaguely known, depending on the language in question. For example, formal languages almost always have an explicitly humanly-defined syntax and even model-theoretic semantics, while the semantics of English seem to escape easy definition, although its syntax is reasonably well-understood. One principle used in the study of languages, attributed to Frege, is the principle of ***compositionality***, where *the content of a sentence is related systematically to terms in which it is composed*. Indeed, while the debate is still out if human languages are truly compositional (Dowty, 2007), programming languages almost always are compositional. The content of the sentence such as 'Ralph has a plane ticket to Paris so he should go to the airport!' can then be composed from the more elementary content of the sub-statements, such as 'Ralph has a plane ticket' which in turn can have its content impacted by words such as 'Paris' and 'ticket.' The argument about whether sentences, words, or clauses are the minimal building block of content (and as such can be assigned a 'truth value') is beyond our scope. Do note one result of the distinction between encoding and content is that sentences that are accepted by the syntax (encoding) of a language, such as Chomsky's famous "Colourless green ideas sleep furiously" may have no obvious interpretation (to content) outside of the pragmatics of Chomsky's particular exposition (1957). The reverse is also true. Statements that may not be grammatically correct can in the right context possess content, like most natural language utterances in speech.

An act of interpretation is usually thought of as a mapping from some sentences in a language to the content of some state-of-affairs in a world. This world is often thought to be the everyday world of concrete trees, houses, and landscapes that humans inhabit. We will not engage in any metaphysical speculation as regards the nature of the world besides our previous minimal definitions of physically connected or disconnected things and processes, so allowing for others to debate the existence of possible worlds or the metaphysical status of the past and future. Regardless, informally an interpretation can be considered to be a mapping from sentences to the physical world itself, a mapping rather appropriately labeled 'God Forthcoming' (Halpin, 2004). However, often we do not have access to the world itself and it is unclear if a simplistic definition such as "the truth of a sentence consists in its agreement with (or correspondence to) reality" makes any sense, for "all these formulations can lead to various misunderstandings, for none of them is sufficiently precise and clear" (Tarski, 1944). In an attempt to define a formal notion of truth, Tarksi defined the interpretation of a language, which he terms the "object" language, in terms of a "meta-language" (1944).

If both the language and the meta-language are suitably formalized, the interpretation of the language can then be expressed in terms of a satisfaction of a mathematical model, where ***satisfaction*** can be defined as *an interpretation to a mathematical model that defines whether or not every sentence in the language can be interpreted to content*, which in the tradition of Frege is usually thought of as a 'truth' value. The model 'stands-in' for the vague and fuzzy world or some portion thereof. While Tarksi originally applied this only to suitably formal languages, others such as Montague have tried to apply this approach, with varying degrees of success and failure, to natural language. A ***model-theoretic semantics*** is a semantics where *an interpretation of a language's sentences is to a mathematical model*. The ***model*** is *a mathematical representation of the world or the language itself*. The relationship is summarized below in Figure 3.2, where the relationship between the model and the world is thought to be distal (such that the model *represents* the world). This is not always the case, as when the model can be thought of as ranging over the world itself.

The adequacy of models is usually judged by whether or not they fulfill the purposes to which the language is designed, or whether or not their behavior adequately serves as a model of some portion of the world. Given a model-theoretic semantics, an interpretation can be given as "a minimal formal description of those aspects of a world which is just sufficient to establish the truth or falsity of any expression" in the language (Hayes, 2004). While again the history and debate over these terms is outside the scope of this thesis, in general the original notion, as pioneered by Carnap (1947), is that a certain *kind of thing may only be described*, and so given an ***intension***, while the *things that satisfy this description* (which may be more than one thing) are ***extensions***. Sentences are ***consistent*** *if they can be satisfied*, ***inconsistent*** if otherwise. Lastly, note that an ***entailment*** is *where an interpretation of one sentence to some content always satisfies the interpretation of another sentence to some content*, i.e. the first statement entails the second. In contrast, an ***inference*** is a *syntactic relationship where one sentence can be used to construct another sentence in a language*. In detail, as shown in Figure 3.2, the syntactic inference mechanisms over time produce more valid inferences, and because these inferences 'line up' with entailments, they also may accurately describe the world outside the formal system. Ideally, this model also 'lines-up' with the world, so the inferences give one more correct statements about the world. Models can be formally captured using various mathematical techniques, of which we have primarily described what is known as denotational semantics, but axiomatic and operational semantics are equally powerful formalisms. Inference can

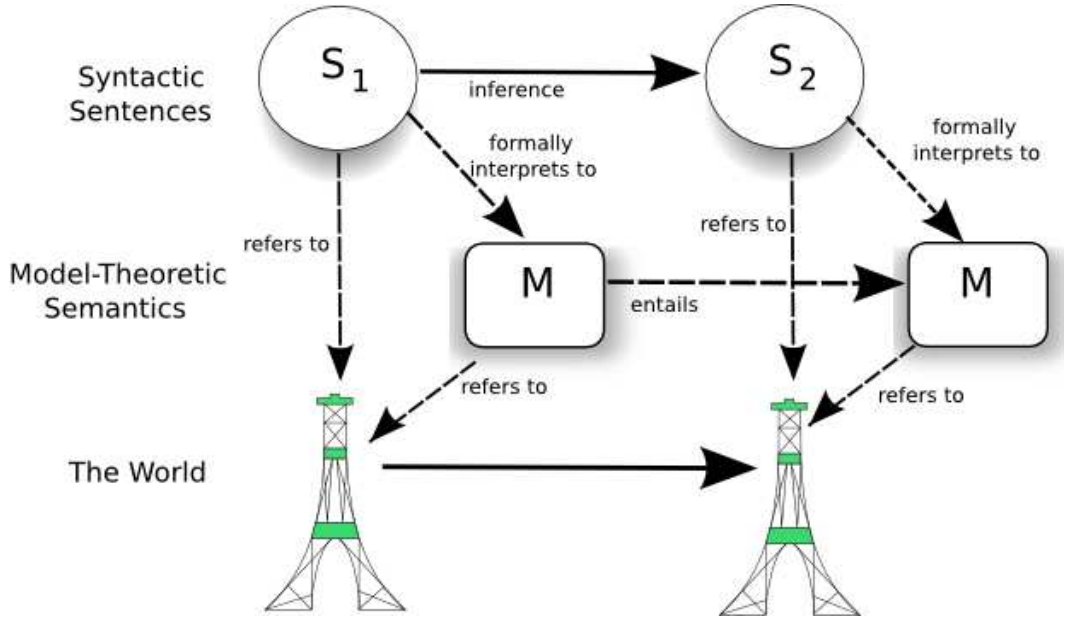

Figure 3.2: Models, Entailment, and Inference

usually be accomplished by some local inference procedure, like a computer program. The inference procedure of a language is ***sound*** *if every inferred sentence can be satisfied* (i.e. the inference mechanism preserves 'truth'), and it is ***complete*** *if every satisfied sentence can be shown to be entailed* (i.e. all 'true' statements can be proven). This is necessarily a quick overview of the large field of formal semantics, and these issues are discussed more in depth in Chapter 6. This is illustrated in Figure 3.2 as the parallel between the causal relationships of the syntactic sentences and their interpretations to a model that *semantically* refers to the world.

3.5 Digitality, Concepts, and Entities

One of the defining characteristics of information on the Web is that this information is digital, bits and bytes being shipped around by various protocols. Yet there is no clear notion of what 'being' digital consists of, and a working notion of digitality is necessary to understand what can and can not be shipped around as bytes on the Web. Much like the Web itself, we can know something digital when we spot it, and we can build digital devices, but developing an encompassing notion of digitality is a difficult task, one that we only characterize briefly here.

One philosophical essay that comes surprisingly close to defining a notion of digi-

tality is Nelson Goodman's *Languages of Art*: Given some physically distinguishable marks, which could compose an encoding, Goodman (1968) defined marks as "***finitely differentiable***" *when it is possible to determine for any given mark whether it is identical to another mark or marks*. This can be considered equivalent to how in categorical perception, despite variation in handwriting, a person perceives hand-written letters as being from a finite alphabet. So, *equivalence classes of marks can be thought of as an application of the philosophical notion of types*. This seems close to 'digital,' so that given a number of types of content in a language, a system is digital if any mark of the encoding can be interpreted to one and only one type of content. Therefore, in between any two types of content or encoding there can not be an infinite number of other types. Digital systems are the opposite of Bateson's famous definition of information: Being digital is simply having a difference that does not make difference (Bateson, 2001). This is not to say there are characteristics of a mark which do not reflect its assignment in a type, and these are precisely the characteristics which are lost in digital systems. So in an analogue system, every difference in some mark makes a difference, since between any two types there is another type that subsumes a unique characteristic of the token. In this manner, the prototypical digital system is the discrete distribution of integers, while the continuous numbers are the analogue system par excellence, since between any real number there is another real number.

Lewis took aim at Goodman's interpretation of digitality in terms of determinism by arguing that digitality was actually a way to represent possibly continuous systems using the combinatorics of discrete digital states (1971). To take a less literal example, discrete mathematics can represent continuous subject matters. This insight caused Haugeland to point out that digital systems are always abstractions built on top of analog systems (1981). The reason we build these abstractions is because digital systems allow perfect reliability, so that once a system is in a digital type (also called a 'digital state'), it does not change unless it is explicitly made to change, allowing both flawless copying and perfect reliability. Haugeland reveals the purpose of digitality to be "a mundane engineering notion, root and branch. It only makes sense as a practical means to cope with the vagarities and vicissitudes, the noise and drift, of earthy existence" (Haugeland, 1981). Yet Haugeland does not tell us what digitality actually is, although he tells us what it does, and so it is unclear why certain systems like computers have been wildly successful due to their digitally (as in the success of analogue computers was not so widespread), while others like 'integer personality ratings' have not been as successful. Without a coherent definition of digitality, it is impossible to

even in principle answer questions like whether or not digitality is *purely* subjective (Mueller, 2007).

In contrast, it seems sensible to state that certain physical processes have the potential to be digital objectively. Different interpreters can interpret the same physical encoding as 'digital' in different ways. The marks '11' can be interpreted as eleven in decimal and three in binary notation. So there are multiple ways one can state a system is digital since digitality is a convergence between an abstract mode of interpretation and an objective system that physically implements a correspondence between the possible states of the system and discrete types of content in the interpretation. An interpretation is ***discrete interpretation*** when it is *a relationship from an encoding to content where the encoding is finitely differentiable and the type of the encoding determines the content*. In order to distinguish these types in the encoding, there must be some physical regularity in the information realization that serves as a *boundary*. Due to this, digitality then allows some finitely differentiable encoding to map via an interpretation to content. When reading letters in a book, we concentrate on the letters, not any minor variations in the quality of the printing – these analogue details are left out of our discrete interpretation of the marks that represent letters to the letters themselves. Reading is a convergence between an encoding that can be discretely interpreted to the alphabet (and onwards and upwards to words, followed by language in general), and a realization in a particular book that can support and maintain the encoding. If we attempt to use an analogue substrate as a realization, such as writing letters in water, and this physical substrate does not have the proper physical characteristics then digitality seems to elude us. Any information is ***digital*** when *the boundaries in a particular encoding capable of a discrete interpretation can converge with a regularity in a physical realization*. This would include sentences in a language that can be realized by sound-waves or the text in an e-mail message that can be re-encoded as bits, and then this encoding realized by a series of voltages. In all these cases, the relevant discrete boundaries can be captured by a realization. The *particular realization of digital information* is given by a ***digital system***. Since the encoding of the information can be captured perfectly by a realization, they can be captured by many possible realizations, and thus can be copied safely and effectively, just as an e-mail message can be sent many times or a digital image reproduced countlessly.

To implement a digital system, there must be a small chance that the information realization can be considered to be in a state that is not part of the discrete types given by the encoding. The regularities that compose the physical boundary allows within a

margin of error a boundary decision to be made in the discrete interpretation of the encoding. So, an information realization is capable of upholding digitality if that buffer created by the margin of error has an infinitesimal chance at any given time of being in a state that is not part of the encoding's discrete state. For example, the hands on a clock can be on the precise boundary between the markings on the clock, just not for very long. In a digital system, on a given level of abstraction, the margin of error does not propagate upwards to higher levels of abstraction that supervene on the lower level of abstraction. This first level of abstraction is 'first-order' digital, and other latter levels can be 'higher-order' digital. First-order digital systems are created from analogue physics, as we have outlined earlier, and of course higher-order digital systems can be built on top of lower-order digital systems. Although in a discrete interpretation, the encoding must be finitely differentiable, the content – as interpreted by an agent – does not have to be capable of being divided into a finite number of discrete types. For example, the encoding `00` could map to the content 'Any human except Ralph or Sandro.' Or, in order to capture apparently analogue music stored in a digital format, one should sample the wavelength twice as often as the highest frequency of the waveform, and this leads the human to have an analogue experience of the music when the music is interpreted by their stereo. So, higher-order analog can be built on top of lower-order digital systems. Furthermore, digital realizations interact with and are based on analogue systems. Digital information, no matter how many layers of encoding are built into each other, are realized in very concrete and therefore analogue realizations. So we will make one metaphysical claim in the spirit of Brian Cantwell Smith, by pre-supposing an analog world, not a fundamentally digital world like that proposed by Fredkin (Smith, 1995; Fredkin, 2003). Some realizations of information are better than others. Since we can create physical systems through engineering, we can create physical substrata that have low probabilities of being in states that do not map to digital at a given level of abstraction. As put by Turing, "The digital computers ... may be classified amongst the 'discrete state machines,' these are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously" (Turing, 1950).

There are many things that are not digital. Some philosophers like Brian Cantwell Smith hold this "slop" or "fuzziness" of regularities to be a fundamental property of many things in the world, like abstract concepts and rather physical people and places

(1995). The ***analogue*** is the rather large and heterogeneous set of *everything that is not digital*. This would include people, such as Ralph himself, who can be represented but not realized as a message, as well as places, like Mount Everest, whose precise boundaries are rather indeterminate. Indeed, things that are *fundamentally* analogue we will call ***entities***, where *the regularities of the thing can only be realized by the thing itself, not in another realization*. This is not to say that the content of entities is itself analogue, so that Ralph can not be distinguished from another person like Sandro, or a place like France cannot be distinguished from Britain. All we mean is that the regularities that define people and places are not finitely differentiable and so cannot be realized in a single digital message. Ralph and Sandro are deeply analogue physical bodies of skin and hair who can be represented, but not realized, by a single digital representation, while places like the Eiffel Tower are literally physical areas of space upon the earth that can likewise be represented, but not realized by some digital substratum. Even when analogue entities may be differentiable, as we can differentiate Ralph from Sandro and France from Britain, these analogue entities can themselves not be realized digitally and copied. Whole places and people cannot just be copied and shipped in a message in bits over a wire! In order to distinguish this use of the term 'entity' from the use of the term 'entity' in HTTP, we will use somewhat facetiously the adjective 'physical' to describe these kinds of entities.

Another thing that has difficulty being realized by a single message are ***concepts***, where the *the regularities of the thing only exist at a level of abstraction that cannot be encoded by a single realization*. Unlike analogue entities, one does not have a definitive local physical thing one can bump up against and touch, because a concept only exists on a level of abstraction that seems physically realized by many disparate things, and may not be completely realized by any of them. Under the rubric of concept comes many things, including imaginary things like unicorns and the concept of a 'horse.' There simply are no unicorns to bump against, and while all horses may to some extent realize the concept of a horse, the concept of a horse is not given by any single horse, but instead a way an agent has of recognizing some thing is actually a realization of a 'horse.' Furthermore, there are abstract concepts that are to some extent imagined to be infinite, such as concepts like the integers that are generated by some combinatorial rules. Obviously, no bounded and connected region of space-time can realize the concept, so concepts are different from analogue entities. It is debatable whether concepts are at some level of abstraction 'really' digital or analogue. Concepts may be differentiable, a unicorn can be distinguished from a horse, or even finitely dif-

ferentiable, since an integer like seven can be distinguished from any other integer. Yet while particular messages can represent concepts, just as some mathematical expressions can represent the concept of the integers or a picture of a unicorn can represent the concept of a unicorn, one would not say that a single realization can adequately capture all the regularities inherent in a concept. So, just as one cannot just ship a physical person as a message, one cannot completely encode a concept and then realize it as a single message. Realizations always fall short of concepts. In order to emphasize that these concepts are a broader class than a single realization or a single physical entity, we shall sometimes use the adjective 'abstract' in front of the term 'concept' in order to be clear.

To return to the Web, the success of the Web lies in no small part on the vast proliferation of digital computers that allow users to create, store, and retrieve information, and use the Web as a naming space to share this information with others. While, according to Hayles, "the world as we sense it on the human scale is basically analogue," the Web is yet another development in a long-line of biological modifications and technological prostheses to impose digitalization on an analogue world (2005). The vast proliferation of digital technologies is possible because there are physical substrata, some more so than others, which support the realization of digital information and give us the advantages that Haugeland rightfully points out is the purpose of the digital: flawless copying and perfect reliability in a flawed and imperfect world (1981).

3.6 Representations

By claiming to be a "universal space of information," the Web is asserting itself to be a space where any encoding can be transferred about any content (Berners-Lee et al., 1992). However, there are some distinct differences between kinds of content, for some content can be distal and other content can be local. In a message between two computers, if the content is a set of commands to 'display these bytes on the screen' then the client can translate these bytes to the screen directly without any worry about what those bytes represent to a human user. However, the content of the message may involve some distal components, such as the string 'Ralph won a ticket to the Eiffel Tower in Paris,' which refers to many things outside of the computer. Differences between receivers allow the self-same content of a message to be both distal and local, depending on the interpreting agent. The message to 'display these bytes on the screen' could cause a rendering of a depiction of the Eiffel Tower to be displayed on the screen,

so the self-same message causes not only a computer to display some bytes but also causes a human agent to receive information about what the Eiffel Tower in Paris looks like.

Any *encoding of information that has distal content in some respect* is called a ***representation***, regardless of the particular language the information is encoded in. Representations are then a subset of information, and inherit the characteristics outlined of all information, such as having one or more possible encodings, one or more realizations, and often a purpose and the ability to evoke meaningful behavior from agents. To have some relationship to a thing that one is disconnected from is to be *about* something else. Generally, *the relationship of a thing to another thing to which one is immediately causally disconnected* is a relationship of ***reference*** to a ***referent*** or ***referents***, *the distal thing or things referred to by a representation*. The thing which refers to the referent(s) we call the 'representation,' and take this to be equivalent to being a *symbol*. To refer to something is to *denote* something, so the content of a representation is its *denotation*. In the tradition of Bretano, the reference relation is considered *intentional* due to its apparent physical spookiness. It appears there is some great looming contradiction: If the content is whatever is held in common between the source and the receiver as a result of the conveyance of a particular message, then how can the source and receiver share some information they are disconnected from?

We will have to make a somewhat convoluted trek to resolve this paradox. The very idea of representation is usually left under-defined as a "standing-in" intuition, that a representation is a representation by virtue of "standing-in" for its referent (Haugeland, 1991). The classic definition of a symbol from the Physical Symbol Systems Hypothesis is the genesis of this intuition regarding representations: "An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y" (Newell, 1980).

There are two subtleties to Newell's definition. Firstly, the notion of a representation is grounded in the behavior of an agent. So, what precisely counts as a representation is never context-free, but dependent upon the agent completing some purpose with the representation. Secondly, the representation *simulates* its referent, and so the representation must be local to an agent while the referent may be non-local: "This is the symbolic aspect, that having X (the symbol) is tantamount to having Y (the thing designated) for the purposes of process P" (Newell, 1980). We will call X a representation, Y the *referent* of the representation, a process P the representation-using *agent*. This definition does not seem to help us in our goal of avoiding physical spookiness,

since it pre-supposes a strangely Cartesian dichotomy between the referent and its representation. To the extent that this distinction is held a priori, then it is physically spooky, as it seems to require the referent and representation to somehow magically line up in order for the representation to serve as a substitute for its missing referent.

The only way to escape this trap is to give a non-spooky theory of how representations arise from referents. Brian Cantwell Smith tackles this challenge by developing a theory of representations that explains how they arise temporally (1995). Imagine Ralph finally gets to Paris and is trying to get to the Eiffel Tower. In the distance, Ralph sees the Eiffel Tower. At that very moment, Ralph and the Eiffel Tower are both physically connected via light-rays. At the moment of tracking, connected as they are by light, Ralph, its light cone, and the Eiffel Tower are a system, not distinct individuals. An alien visitor might even think they were a single individual, a 'Ralph-Eiffel Tower' system. While walking towards the Eiffel Tower, when the Eiffel Tower disappears from view (such as from being too close to it and having the view blocked by other buildings), Ralph keeps staring into the horizon, focused not on the point the Eiffel Tower was at before it went out of view, but the point where he thinks the Eiffel Tower would be, given his own walking towards it. Only when parts of the physical world, Ralph and the Eiffel Tower, are physically separated can the agent then use a representation, such as the case of Ralph using an internal 'mental image' of the Eiffel Tower to direct his walking towards it, even though he cannot see it. The agent is distinguished from the referent of its representation by virtue of not only disconnection but by the agent's attempt to track the referent, "a long-distance coupling against all the laws of physics" (Smith, 1995). The local physical processes used to track the object by the subject are the representation.

This notion of representation is independent of the representation being either internal or external to the particular agent, regardless of how one defines these boundaries.[3] Imagine that Ralph had been to the Eiffel Tower once before. He could have marked its location on a piece of paper by scribbling a small map. Then, the marking on the map could help guide him back as the Eiffel Tower disappears behind other buildings in the distance. This characteristic of the definition of representation being capable of including 'external' representations is especially important for any definition of a representation to be suitable for the Web, since the Web is composed of information that is considered to be external to its human users.

[3]The defining of "external" and "internal" boundaries is actually non-trivial, as shown in Halpin (2008a).

However fuzzy the details of Smith's story about representations may be, what is clear is that instead of positing a connection between a referent and a representation a priori, they are introduced as products of a temporal process. This process is at least theoretically non-spooky since the entire process is capable of being grounded out in physics without any spooky action at a distance. To be grounded out in physics, all changes must be given in terms of connection in space and time, or in other words, via effective reach. Representations are "a way of exploiting local freedom or slop in order to establish coordination with what is beyond effective reach" (Smith, 1995). In order to clarify Smith's story and improve the definition of the Physical Symbol Systems Hypothesis, we consider Smith's theory of the "origin of objects" to be a ***referential chain*** with distinct stages (Halpin, 2006):

- **Presentation**: Process *S* is connected with process *O*.
- **Input**: The process *S* is connected with *R*. Some local connection of *S* puts *R* in some causal relationship with process *O*. This is entirely non-spooky since *S* and *O* are both connected with *R*. *R* eventually becomes the representation.
- **Separation**: Processes *O* and *S* change in such a way that the processes are disconnected.
- **Output**: Due to some local change in process *S*, *S* uses its connection with *R* to initiate local meaningful behavior that is in part caused by *R*.[4]

In the 'input' stage, the *referent* is the cause of some characteristic(s) of the information. The relationship of *reference* is the relationship between the encoding of the information (the representation) and the referent. The relationship of interpretation becomes one of reference when the distal aspects of the content are crucial for the meaningful behavior of the agent, as given by the 'output' stage. So we have constructed an ability to talk about representations and reference while not presupposing that behavior depends on internal representations or that representations exist a priori at all. Representations are only needed when the relevant intelligent behavior requires some sort of distal co-ordination with a disconnected thing.

As a representation is just a particular kind of encoding of information, the interpretation of a representation results in content that is dependent on a distal referent via the referential chain. In this manner, the act of reference can then be defined as

[4]In terms of Newell's earlier definition, 0 is X while S is P and R is Y.

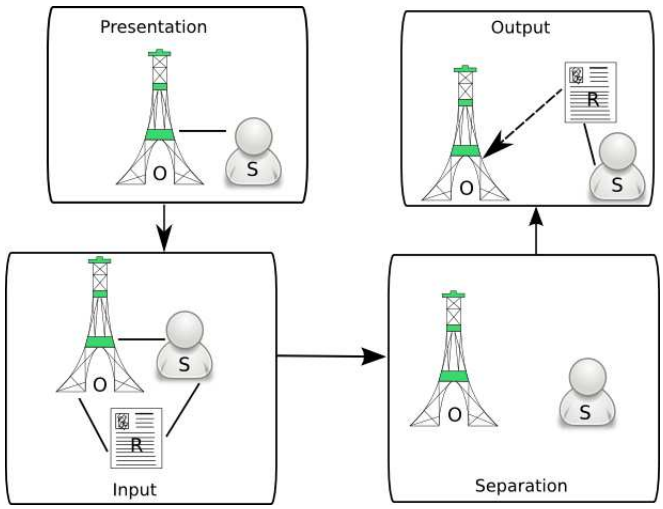

Figure 3.3: The Referential Chain

the interpretation of a representation. This would make our notion of representation susceptible to being labeled a *correspondence theory of truth* (Smith, 1987), where a representation refers by some sort of structural correspondence to some referent. However, our notion of representation is much weaker, requiring only a causation between the referent and the representation – and not just any causal relationship, but one that is meaningful for the interpreting agent – as opposed to some tighter notion of correspondence such as some structural 'isomorphism' between a representation and its "target," the term used by Cummins to describe what we have called the "referent" of a representation (1996). So an interpretation or an act of reference should therefore not be viewed as mapping to referents, but a mapping to some content where that content leads to meaningful behavior precisely because of some referential chain. This leads to the notion of a Fregean 'objective' sense, which we turn to shortly in Section 3.7. To give an example, a picture of the Eiffel Tower has an interpretation to some content that, while locally embodied as something like a mental image of the Eiffel Tower, is effective due to its historical connection to the distal and actual Eiffel Tower itself.

Up until now, it has been implicitly assumed that the referent is some physical entity that is non-local to the representation, but the physical entity was still existent, such as the Eiffel Tower. However, remember that the definition of non-local includes *anything* the representation is disconnected from, and so includes physical entities that may exist in the past or the future. The existence of a representation does not imply the existence of the referent or the direct acquaintance of the referent by the agent using a representation. A representation only implies that some aspect of the content is

non-local. However, this seems to contradict our 'input' stage in the representational cycle, which implies that part of our definition of representation is historical: for every *re*-presentation there must be a presentation, an encounter with the thing presented. By these conditions, the famous example of Putnam's example of an ant tracing a picture of Winston Churchill by sheer accident in the sand would not count as a representation (1975). If Ralph didn't know where the Eiffel Tower was, but navigated the streets of Paris and found the Eiffel Tower by reference to a tracing of a Kandinsky painting in his notebook, then Ralph would not then be engaged in any representation-dependent meaningful behavior, since the Kandinsky painting lacks the initial presentation with the Eiffel Tower. The presentation does not have to be done by the subject that encountered the thing directly. However, the definition of a representation does not mean that the *same* agent using the representation had to be the agent with the original presentation. A representation that is created by one agent in the presence of a referent can be used by another agent as a 'stand-in' for that referent if the second agent shares the same interpretation from encoding to distal content. So, instead of relying on his own vision, Ralph buys a map and so relies on the 'second-order' representation of the map-maker, who has some historical connection to someone who actually traveled the streets of Paris and figured out where the Eiffel Tower was. In this regard, our definition of representation is very much historical, and the original presentation of the referent can be far back in time, even evolutionary time, as given by accounts like those of Millikan (1984). One can obviously refer to Gustave Eiffel even though he is long dead and buried, and so no longer exists.

Also, the referent of a representation may be a *concept*, like the concept of a horse, unicorns and other imaginary things, referents to future states such as 'see you next year,' and descriptive phrases whose supposed *exact* referent is unknown, such as 'the longest hair on your head on your next birthday.' While all these types of concepts are quite diverse, they are united by the fact that they cannot be completely realized by local information, as they depend on partial aspects of an agent's local information, the future, or things that do not exist. Concepts that are constructed by definition, including imaginary referents, also have a type of 'presence,' it is just that the 'presentation' of the referent is created via the initial description of the referent. Just because a referent is a concept – as opposed to a physical entity – does not mean the content of the representation cannot have an meaningful effect on the interpreter. For example, exchanging representations of 'ghosts' - even if they do not quite identify a coherent class of referents - can govern the behavior of ghost-hunters. Indeed, it is the power

and flexibility of representations of these sorts that provide humans the capability to escape the causal prison of their local environment, to plan and imagine the future.

Our use of representation and reference is very broad, so that the phenomenon of representation can be thought of as nearly everywhere. Our message in the example, the denoting phrase that 'Ralph has won a ticket to Paris,' includes acts of reference to Ralph, Paris, the past, and implications for Ralph's future activity. Indeed, with our definition of reference, it appears that almost all linguistic sentences other than those describing the immediate local environment involve some representational aspect. Indeed, representations exist at multiple levels of abstraction and composition. For example, the 'text of Moby Dick' in English locally carries the information about the 'story about a white whale' on one level of abstraction. However, the story itself is mired in representation, involving distal referents such as whales, harpoons, and 19th-century New England. In this case, it is useful to separate from the broader class of representations those things whose primary purpose is to represent distal content from those things that only have some representational content. For example, an encyclopedia article about the Eiffel Tower or a picture of the Eiffel Tower by itself have as their primary purpose the representation of the Eiffel Tower, as opposed to a map of Paris or a movie like the 'Lavender Hill Mob' that simply features the Eiffel Tower as part of a more general or different purpose. In the cases where it is *the primary purpose of something to be a representation*, we will call that representation a ***description*** if it is in a natural or formal language or a ***depiction*** if it is in an iconic language.

3.7 Sense and Reference

The tradition most of these definitions have come from has been one strictly in line with the philosophy of cognitive science and the mind, as exemplified by Brian Cantwell Smith and Dretske, who tends to spend much energy discussing the nature of terms like 'information' and 'representation.' However, there is an important connection that seems to have been missed by Dretske and others, the connection between information, sense, and reference. This is likely because Frege himself was quite cryptic with regards to any definition of 'sense.' Therefore, we have no choice but to return to Frege's original controversial theory of sense and reference as given in *Sinn und Bedeutung* (Frege, 1892).[5]

[5]The ambiguous translation of this work from original German has been a source of great philosophical confusion. While the word 'Sinn' has almost always been translated into 'sense,' the word

The key idea lies in Frege's contention that the meaning of any term in a language is determined by what Frege calls the "sense" of the sentences that use the term, rather than any direct reference of the term (1892). For Frege, the referents of a term should be assigned to truth-values, but two statements may share the same truth-value but have different senses. According to Frege, two sentences could be the same only if they shared the same sense. Take for example the two sentences "Hesperus is the Evening Star" and "Phosphorus is the Morning Star." (Frege, 1892). Since the ancient Greeks did not know that 'The Morning Star is the same as the Evening Star,' they did not know that the names 'Hesperus' and 'Phosphorus' share the same referent when they baptized the same star, the planet Venus, with two different names (Frege, 1892). Therefore, Frege says that these two sentences have distinct 'senses' even if they share the same referent, so sense is not just a function to referents. Frege pointed out that, far from being meaningless, statements of identity that would be mere tautologies from the point of view of a theory of reference are actually meaningful if one realizes different terms can have distinct senses. One can understand a statement like 'The Morning Star is the Evening Star' without knowing that both refer to Venus. In fact, one may only know that the 'Morning Star' refers to Venus. By learning the 'Morning Star' and the 'Evening Star' are not distinct senses but a single sense, one is doing actual *meaningful cognitive work* by putting these two senses together. While the idea of a notion of 'sense' seems intuitive from the example, it is famously hard to define, even informally. Frege defines 'sense' in terms of the mysterious *mode of presentation*, for "to think of then being connected with a sign (name, combination of words, letters), besides that to which the sign refers, which may be called the reference of the sign, also what I should like to call the sense of the sign, wherein the mode of presentation is contained" (1892). This rather cryptic statement has caused multiple decades of debate by philosophers of language like Russell and Kripke who have attempted to banish the notion of sense and simply build a theory of meaning from the concept of reference. These attempts are detailed in Chapter 6.

Regardless of what precisely 'sense' is, Frege believed that the notion of sense is what allows an agent to understand sentences that may not have a referent, for "the words 'the celestial body most distant from Earth' has a sense, but it is very doubtful

'Bedeutung' has been translated into *either* 'reference' or 'meaning,' depending on the translator. While 'Bedeutung' is most usually translated into the fuzzy English word 'meaning' by most German speakers, the *use* to which Frege puts it is much more in line with how the word 'reference' is used in philosophy. So in the tradition of Michael Dummett, we will translate Frege's 'Bedeutung" into 'reference' (Dummett, 1973).

there is also a thing they refer to...in grasping a sense, one certainly is not assured of referring to anything" (Frege, 1892). So it is the concept of sense that should be given a priority over reference. This is not to deny the role of reference whatsoever, since "to say that reference is not an ingredient in meaning is not to deny that reference is a consequence of meaning..it is only to say that understanding which speaker of a language has a word in that language...can never consist merely in his associating a certain thing with it as its referent; there must be some particular *means* by which this association is effected, the knowledge of which constitutes his grasp of its sense" (Dummett, 1973).

Sense is in no way an 'encoded' referent, since the referent is distal from the sense usually. Instead, the sense of a sentence would naturally lead an agent to correctly guess the referents of the sentence. Yet how could this be detected? Again, sense is also not merely some encoded meaning, nor is sense strictly 'in the head' with no effect on meaningful behavior. As put by Wittgenstein, "When I think in language, there aren't 'meanings' going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought" (1953). Sense is the bedrock upon which meaning is constructed, and must be encoded in a language. In fact, according to Frege, sense can only be determined from a sentence in a language, and the sense of a sentence almost always requires an understanding of the other sentences in a given discourse. Without determining from a number of possible senses a sentence *may* have, which sense the sentence *does* have, one cannot meaningfully act. However, the sense used by the agent may be incorrect according to the creator of the sentence's purpose, but that does not prevent the agent from acting.

So, how can sense be determined, or at least detected? After all, almost *anything* counts as meaningful behavior. While sense determination is a difficult and context-ridden question that seems to require some full or at least 'molecular' language understanding, the best account of sense detection so far is given by the earlier notion of assertoric content of Dummett, which is simply that an agent can be thought of as interpreting to a sense if they can answer a number of 'yes-no' binary questions about the sense in a way that makes 'sense' to other agents speaking the language (Dummett, 1973). There is a tantalizing connection of Dummett's assertoric content as answers to binary questions to the information-theoretic reduction of uncertainty through binary choices (bits), as the content of information cannot be derived without enough bits in the encoding. Overall, Dummett's notion of sense as grounded in actual language use naturally leads to another question: Is sense objective?

The reason the notion of sense was thought of as so objectionable by many philosophers like Russell and Kripke was that it was viewed as a private, individual notion, much like the Lockean notion of an *idea*. Frege himself clearly rejects this, strictly separating the notion of a sense from an individual subjective idea of a referent. Far from subjective, Frege believed that sense was inherently *objective*, "the reference of a proper name is the object itself which we designate by using it; the idea which we have in that case is wholly subjective, in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself" (1892). A sense is objective insofar as it is a shared part of an inherently public language, since a sense is the "common property of many people, and so is not a part of a mode of the individual mind. For one can hardly deny that mankind has a common store of thoughts which is transmitted from one generation to another" (1892). While the exact nature of a sense is still unclear, its main characteristic is that it should be whatever is *objectively shared* between agents as regards their use of terms in a language. It is precisely this notion that sense is 'objective' that allows us to connect our work in the philosophy of information and representation to the philosophy of language.

This is namely because the Fregean notion of sense is *identical* with our reconstructed notion of informational *content*. These terms should be viewed as identical. The content of information is precisely what is shared between the source and the receiver as a result of the conveyance of a particular message. By definition, this holding of content in common which is the result of the transmission of an information-bearing message *must* by definition involve at least *two* things: a source and a receiver. Furthermore, if the source and receiver are considered to be human agents capable of speaking natural language, then by the act of sharing sentences, which are just encodings shared over written letters or acoustic waves in natural language, the two speakers of language are sharing the content of those sentences. Since the content is possessed by two people, and is by definition of information the *same* content, insofar as *subjective* is defined to be that which is only possessed by a single agent and *objective* is defined to be that which is possessed by more than one agent (although not necessarily all agents), then *content is objective*.

Most of the productive concepts reconstructed earlier then map straightforwardly to terms in philosophy of language. Sentences and terms natural in a language have both a syntactic encoding and a semantic content or sense, that can multiply realized over differing mediums. A sentence is a fully-fledged information-carrying message, that can have multiple realizations in the form of different utterances at different

points in space and time. The Gricean notion of a speaker's intentions then maps to our purposes, and his more fully-fledged notion of linguistic meaning maps closely to our notion of meaning. The problem of word senses is now revealed to be much larger than previously supposed, as it now stretches across to all sorts of non-natural languages. Everything from messages in computer protocols (formal languages) to paintings (iconic languages) are now just encodings of information, and these too have senses and possible sense ambiguities.

Representations are not just then 'in the head' but also present in sentences in the form of *names*. In particular, a name in natural language is no more than some encoding that has as its interpretation the sense of a distal referent. The class of *proper names*, long a source of interest, is just a representation in natural language whose referent is an entity, such that the name 'Ralph' refers to the person Ralph, while the larger class of names such as 'towers' or 'integers' can refer to groups of entities and concepts. There may be some objection to the idea that a mere *name* in a sentence is a full-blooded representation. However, unlike some theories of representation such as those put forward by Cummins, we do not require that there be some "isomorphism" or other structured relationship between the representation and the referent (1996). We only require the much less-demanding causal relationship with some impact upon the sense (content) and thus the meaningful behavior of the agent. While it is obvious there is nothing inherent in the term 'Eiffel Tower' that leads the letters or phonemes in the name to correspond in any significant structural way with the Eiffel Tower itself, as long as the sense of the name is dependent on *there being a referent* that the name 'stands-in' for, so a name like the 'Eiffel Tower' is still a representation of the Eiffel Tower itself.

3.8 Conclusion

In conclusion, we concur with Dummett that any account of meaning will have in essence three layers, where the outer layer has priority over the inner layers (1993). First, the "core" would be the "theory of reference" while "surrounding the theory of reference will be a shell, forming the theory of sense" so that "the theory of reference and the theory of sense form together one part of the theory of meaning: the other, supplementary, part is the theory of force", or as we would put it, a theory of purpose (Dummett, 1993). So nothing in the philosophical account presented so far is new, although the manner of reconstruction and recombination may be new. We have built

from a fairly simple account of connection and disconnection from Brian Cantwell Smith, moving to the account of information encoding and content from Dretske, to then a notion of purpose and meaning derived from Millikan, and then finally returning to an account of digitality from Haugeland and an account of reference and representation from Brian Cantwell Smith again. Then, at the last possible juncture, we show that Frege's account of sense can be seen as the same as our account of content for information in general given in Section 3.2.

The convergence of informational content with linguistic sense is liberating for the philosophy of language, because while previously, issues of sense and reference seem to have primarily been bound to natural languages, the move of identifying content with sense and sentences with encodings then opens a whole new enterprise: the impact of sense and reference on non-natural languages, in particular the study of formal languages created by digital technology. Our interest in this is how these issues of meaning, sense, and reference can be analyzed in context with the World Wide Web. Surprisingly, classical problems of sense and reference re-emerge with a vengeance on the Web. However, first we must define the foundational terminology of the Web itself.

Chapter 4

The Principles of Web Architecture

You have abandoned the old domain, the old concepts. Here you are in a new domain, for which new concepts will give you the knowledge. The sign that a real change in locus and problematic has occurred, and that a new adventure is beginning, the adventure of science in development. **Louis Althusser** (1963)

While the significance and history of the Web have been explained, the task remains to show that the Web is a well-defined system with a unique combination of properties. In Chapter 5 we will demonstrate how these principles can in turn be applied to the Semantic Web.

Can the various technologies that go under the rubric of the World Wide Web be found to have common principles and terminology? This question would at first seem to be shallow, for one could say that any technology that is described by its creators, or even the public at large, can be considered trivially 'part of the Web.' To further complicate the matter, the terms like the 'Web' and the 'Internet' are taken to be synonyms in common parlance, and so are often deployed as synonyms. In a single broad stroke, we can distinguish the Web and the Internet. The Internet is a type of packet-switching network as defined by its use of the TCP/IP protocol. The purpose of the Internet is to get data from one computer to another. In contrast, the Web is a space of names for information defined by its usage of URIs. So, the purpose of the Web is the use of URIs for accessing and referring to information. The Web and the Internet are then strictly separable, for the Web, as a space of URIs, could be realized on top of other types of networks that move bits around, much as the same virtual machine can be realized on top of differing physical computers. For example, one could imagine the Web being built on top of a network built on principles different from TCP/IP, such as

OSI, an early competitor to the TCP/IP stack of networking protocols (Zimmerman, 1980). Likewise, before the Web, there were a number of different protocols with their own naming schemes built upon the Internet like Gopher (Anklesaria et al., 1993).

Is it not presumptuous of us to even hope that such an unruly phenomenon such as the Web even has guiding principles? Again we must appeal to the fact that unlike natural language or chemistry, the Web is like other engineered artifacts, created by particular individuals with a purpose, and designed with this purpose in mind. Unlike the case of the proper function of natural language, where natural selection itself will forever remain silent to our questions, the principal designers of the Web are still alive to be questioned in person, and their design rationale is overtly written down on various notes, often scribbled on some of the earliest web-pages of the Web itself. It is generally thought of that the core of the Web consists of the following standards, given in their earliest incarnation: HTTP (Berners-Lee et al., 1996), URI (Berners-Lee, 1994a), and HTML (Berners-Lee and Connolly, 1993). So the basic protocols and data formats that proved to be successful were the creation of a fairly small number of people, such as Tim Berners-Lee, Roy Fielding, and Dan Connolly.

The primary source for our terminology and principles of Web architecture is a document entitled *The Architecture of the World Wide Web* (AWWW), a W3C Recommendation edited by Ian Jacobs and Norm Walsh to "describe the properties we desire of the Web and the design choices that have been made to achieve them" (Jacobs and Walsh, 2004). The AWWW is an attempt to systematize the thinking that went into the design of the Web by some of its primary architects, and as such is both close to our project and an inspiration. In particular, this document is an exegesis of Tim Berners-Lee's notes on "Design Issues: Architectural and philosophical points"[1] and Roy Fielding's dissertation *Architectural Styles and the Design of Network-based Software Architectures* (Fielding, 2000). The rationale for the creation of such a document of principles of the Web developed organically over the existence of the W3C, as new proposed technologies were sometimes considered to be either informally compliant or non-compliant with Web architecture. When the proponents of some technology were told that their particular technology was not compliant with Web architecture, they would often demand that somewhere there be a description of this elusive Web architecture. The W3C in response set up the Technical Architecture Group (TAG) to "document and build consensus" upon "the underlying principles that should be ad-

[1]There exist a collection of unordered personal notes available at: `http://www.w3.org/DesignIssues/`, which we also refer directly to in the course of this chapter.

hered to by all Web components, whether developed inside or outside W3C," as stated in its charter.[2] The TAG also maintains a numbered list of problems (although the numbers are in no way sequential) that attempts to resolve issues in Web architecture by consensus, with the results released as notes called 'W3C TAG findings,' which are also referred to in this discussion. The TAG's only Recommendation at the time of writing is the aforementioned *Architecture of the Web: Volume 1* but it is reasonable to assume that more volumes of *Architecture of the Web* may be produced after enough findings have been accumulated. The W3C TAG's AWWW is a blend of commonsense and sometimes surprising conclusions about Web architecture that attempts to unify diverse web technologies with a finite set of core design principles, constraints, and good practices (Jacobs and Walsh, 2004). However, the terminology in AWWW is often thought to be too informal and ungrounded to use by many, and we attempt to remedy this in the next few chapters by fusing the terminology of Web architecture with our philosophical terminology developed in Chapter 3.

4.1 The Terminology of the Web

To begin our reconstruction of Web architecture, the first task is the definition of terms, as otherwise the technical terminology of the Web can lead to as much misunderstanding as understanding. To cite an extreme example, people coming from communities like the artificial intelligence community use terms like 'representation' in a way that is different from those involved in Web architecture. We begin with the terms commonly associated with a typical exemplary Web interaction. For an agent to learn about the *resource* known as the Eiffel Tower in Paris, a person can access its *representation* using its *Uniform Resource Identifier (URI)* `http://www.tour-eiffel.fr/` and retrieve a webpage in the HTML *language* using the HTTP *protocol*.

4.1.1 Protocols

A ***protocol*** is *a convention for transmitting information between two or more agents*, an equally broad definition that encompasses everything from computer protocols like TCP/IP to conventions in natural language like those employed in diplomacy. A protocol often specifies more than just the particular encoding, but also may attempt to

[2] Quoted from their charter, available on the Web at: `http://www.w3.org/2001/07/19-tag` (last accessed April 20th, 2007).

specify the interpretation of this encoding and the meaningful behavior that the sense of the information should engender in an agent. A ***payload*** is *the information transmitted by a protocol*. Galloway notes that protocols are "the principle of organization native to computers in distributed networks" and that agreement on protocols are necessary for any sort of network to succeed in the acts of communication (2004). The paradigmatic case of a protocol is TCP/IP, where the payload transmitted is just bits in the body of the message, with the header being used by TCP to ensure the lossless delivery of the bytes. TCP/IP transmits strictly an encoding of data as bits and does not force any particular interpretation on the bits; the payload could be a picture of the Eiffel Tower, web-pages about the Eiffel Tower, or just meaningless random bits. All TCP/IP does is move some particular bits from one individual computer to another, and any language that is built on top of the bit-level is strictly outside the bounds of TCP/IP. Since these bits are usually communication with some purpose, the payload of the protocol is almost always an encoding to some sense above and beyond that of the raw bits themselves.

The Web is based on a ***client-server architecture***, meaning that *protocols take the form of a request for information and a response with information*. The ***client*** is defined as *the agent that is requesting information* and the ***server*** is defined as *the agent that is responding to the request*. In a protocol, an ***endpoint*** is *any process that either requests or responds to a protocol*, and so includes both client and servers. The client is often called a ***user-agent*** since it is the user of the Web. A user-agent may be anything from a web-browser to some sort of automated reasoning engine that is working on behalf of another agent, often the specifically human user. The main protocol in this exposition will be the ***HyperText Transfer Protocol*** (HTTP), as most recently defined by IETF RFC 2616 (Fielding et al., 1999). HTTP is a protocol originally intended for the transfer of hypertext documents, although its now ubiquitous nature often lets it be used for the transfer of almost any encoding over the Web, such as its use to transfer XML-based SOAP (originally the *Simple Object Access Protocol*) messages in Web Services (Box et al., 2000). HTTP consists of sending a ***method***, *a request for a certain type of response from a user-agent to the server*, including information that may change the state of the server. These methods have a list of ***headers*** that *specify some information that may be used by the server to determine the response*. The ***request*** is *the method used by the agent and the headers, along with a blank line and an optional message body*.

The methods in HTTP are HEAD, GET, POST, PUT, DELETE, TRACE, OP-

```
GET /index.html HTTP/1.0
User-Agent: Mozilla/5.0
Accept: */*
Host: www.example.org
Connection: Keep-Alive
```

Figure 4.1: An HTTP Request from a client

TIONS, and CONNECT. We will only be concerned with the most frequently used HTTP method, GET. GET is informally considered 'commitment-free,' which means that the method has no side effects for either the user-agent or the server, besides the receiving of the response (Berners-Lee et al., 1996). So a GET method should not be used to change the state of a user-agent, such as charging someone for buying a plane ticket to Paris. To change the state of the information on the server or the user-agent, either PUT (for uploading data directly to the server) or POST (for transferring data to the server that will require additional processing, such as when one fills in a HTML form) should be used. A sample request to `http:///www.example.org` from a Web browser user-agent is given in Figure 4.1.

The first part of an HTTP response from the server then consists of an HTTP ***status code*** which is *one of a finite number of codes which gives the user-agent information about the server's HTTP response itself.* The two most known status codes are HTTP 200, which means that the request was successful, or 404, which means the user-agent asked for data that was not found on the server. The first digit of the status code indicates what general class of response it is. For example, the two hundred level response codes mean in general a successful request, although 206 means partial success. The four hundred level response codes indicate that the user-agent asked for a request that the server could not fulfill, while the one hundred level is informational, three hundred level is redirectional, and five hundred level means server error. After the status codes there is an ***HTTP entity*** which is "*the information transferred as the payload of a request or response*" (Fielding et al., 1999). This technical use of the word 'entity' should be distinguished from our earlier use of the term 'entity' to describe a thing like the Eiffel Tower that can only be realized by itself, not transferred as abstract information in another realization. In order to do so, we will take care to preface the protocol name 'HTTP' before any 'HTTP entity,' while the term 'entity' by itself refers to the more philosophical notion of an entity. An HTTP entity consists of "entity-header

```
HTTP/1.1 200 OK
Date: Wed, 16 Apr 2008 14:12:09 GMT
Server: Apache/2.2.4 (Fedora)
Accept-Ranges: bytes
Connection: close
Content-Type: text/html; charset=ISO-8859-1
Content-Language: fr
```

Figure 4.2: An HTTP Response from a server

fields and... an entity-body" (Fielding et al., 1999) An ***HTTP response*** consists of *the combination of the status code and the HTTP entity*. These responses from the server can include an additional header, which specifies the date and last modified date as well as optional information that can determine if the desired representation is in the cache and the content-type of the representation. A sample HTTP response to the previous example request, excluding the HTTP entity-body, is given in Figure 4.2.

In the HTTP response, an HTTP entity body is returned. The encoding of the HTTP entity body is given by the HTTP entity header fields that specify its `Content-type` and `Content-language`. These are both considered different languages, as a single webpage can be composed in multiple languages, such as the text being given in English with various formatting given in HTML. Every HTTP entity body should have its particular encoding specified by the `Content-type`. *The formal languages that can be explicitly given in a response or request in HTTP* are called ***content types***. In the example response, based on the header that the content type is `text/html` a user-agent can interpret ('display as a web-page') the encoding of the HTTP entity body as HTML. Since the same encoding can theoretically represent many different languages besides HTML, a user-agent can only know definitely how to process a message through the content type. If no content type is provided, the agent can guess the content type through various heuristics including looking at the bytes themselves, a process informally called *sniffing*. A user-agent can specify what media types they (can) prefer, so that a web-server that can only present JPEG images can specify this by also asking for the content type `image/jpeg` in the request.

Content-types in HTTP were later generalized as 'Internet Media Types' so they could be applied with any Internet protocol, not just HTTP and MIME (*Multimedia Internet Message Extensions*, an e-mail protocol) (Postel, 1994). A ***media type*** con-

sists of *a two-part scheme that separates the type and a subtype of an encoding*, with a slash indicating the distinction, as in `text/html`. Internet media types are centrally registered with IANA at `http://www.iana.org/assignments/media-types/`, although certain 'experimental' media types (those beginning with 'x-') can be created in a decentralized manner (Postel, 1994). A central registry of media types guarantees the interoperability of the Web, although increasingly new media-types are dependent on extensions to specific applications (plug-ins) in order to run. Support for everything from new markup languages to programming languages such as Javascript can be declared via support of its media type.

To move from concrete bits to abstract definitions, a protocol can be defined and implemented in many different types of way. In the early ARPANet, the first wide-area network and foundation of the Internet, the protocol was 'hard-wired' in the hardware of the Interface Message Processor (IMP), a separate machine attached to computers in order to interface them with ARPANet (Hafner and Lyons, 1996). As more and more networks multiplied, these heterogeneous networks began using different protocols. While the invention of TCP/IP let these heterogeneous networks communicate, TCP/IP does not interpret messages beyond bits. Further protocols built on top of TCP/IP, such as FTP (File Transfer Protocol) for the retrieval of files (Postel and Reynolds, 1985), Gopher for the retrieval of documents (Anklesaria et al., 1993), and SMTP (Simple Mail Transfer Protocol) for the transfer of mail (Postel, 1982). Since one computer might hold many different kinds of information, IP addresses were not enough as they only identified where a particular device was on the network. Thus each protocol created its own naming scheme to allow it to refer to and access things on a more fine-grained level than IP addresses. Furthermore, each of these protocols was often associated (via registration with a governing body like IANA, the *Internet Assigned Numbers Authority*) with particular ports, such that port 25 was used by SMTP and port 70 by Gopher. With this explosion of protocols and naming schemes, each Internet application was its own 'walled garden.' Names created using a particular protocol were incapable of being used outside the original protocol, until the advent of the naming scheme of the Web (Berners-Lee, 2000).

4.1.2 Uniform Resource Identifiers

The World Wide Web is defined by the AWWW as "an information space in which the items of interest, referred to as resources, are identified by global identifiers called

Uniform Resource Identifiers (URI)" (Jacobs and Walsh, 2004). This naming scheme, not any particular language like HTML, is the primary identifying characteristic of the Web. URIs arose from a need to organize the "many protocols and systems for document search and retrieval" that were in use on the Internet, especially considering that "many more protocols or refinements of existing protocols are to be expected in a field whose expansion is explosive" (Berners-Lee, 1994a). Despite the "plethora of protocols and data formats," if any system was "to achieve global search and readership of documents across differing computing platforms," gateways that can "allow global access" should "remain possible" (Berners-Lee, 1994a). The obvious answer was to consider all data on the Internet to be a single space of names with global scope.

URIs accomplish their universality over protocols by moving *all the information used by the protocol within the name itself*. The information needed to identify any protocol-specific information is all specified in the name itself: the name of the protocol, the port used by the protocol, any queries the protocol is responding to, and the hierarchical structure used by the protocol. The Web is then first and foremost a naming initiative "to encode the names and addresses of objects on the Internet" rather than anything to do with hypertext (Berners-Lee, 1994a). The notion of a URI can be viewed as a 'meta-name,' a name which takes the existing protocol-specific Internet addresses and wraps them *in the name itself*, a process analogous to reflection in programming languages (Smith, 1984). Instead of limiting itself to only existing protocols, the URI scheme also abstracts away from any particular set of protocols, so that even protocols in the future or non-Internet protocols can be given a URI; "the Web is considered to include objects accessed using an extendable number of protocols, existing, invented for the Web itself, or to be invented in the future" (Berners-Lee, 1994a).

One could question why one would want to name information outside the context of a particular protocol. The benefit is that the use of URIs "allows different types of resource identifiers to be used in the same context, even when the mechanisms used to access those resources may differ" (Berners-Lee et al., 2005). This is an advantage precisely because it "allows the identifiers to be reused in many different contexts, thus permitting new applications or protocols to leverage a pre-existing, large, and widely used set of resource identifiers" (Berners-Lee et al., 2005). This ability to access with a single naming convention the immense amount of data on the entire Internet gives an application such as the ubiquitous Web browser a vast advantage over an application that can only consume application-specific information.

Although the full syntax in Backus-Naur form is given in IETF RFC 3986 (Berners-Lee et al., 2005), a URI can be given as the regular expression `URI= [scheme ":"] [hierarchical component]* [ "?" query ]? [ "#" fragment]?`. A ***scheme*** is *a name of the protocol or other naming convention used to begin the URI.* The scheme of a URI does not determine the protocol that a user-agent has to employ to use the URI. For example, a HTTP request may be used on `ftp://www.example.org`. The scheme of a URI merely indicates a preferred protocol for use with the URI. A ***hierarchical component*** is *the left to right dominant component of the URI that syntactically identifies the resource.* URIs are federated, insofar as each scheme identifies the syntax of its hierarchical component. For example, with HTTP the hierarchical component is given by `[authority] [//] [":" port]? ["/" path component]*`. The ***authority*** is *a name that is usually a domain name, naming authority, or a raw IP address, and so is often the name of the server*. However, in URI schemes like `tel` for telephone numbers, there is no notion of an authority in the scheme. The hierarchical component contains special reserved characters that are in HTTP characters such as the backslash for locations as in a file system. For ***absolute URIs***, *there must be a single scheme and the scheme and the hierarchical component must together identify a resource* such as `http://www.example.com:80/monument/EiffelTower` in HTTP, which identifies the resource accessible from port 80 of the authority `www.example.com` with the path component `/monument/EiffelTower`. The port authority is usually left out, and assumed to be 80 by HTTP-enabled clients. Interestingly enough there are also ***relative URIs*** *in some schemes like HTTP, where the path component itself is enough to identify a resource within certain contexts*, like that of a web-page. This is because the scheme and authority itself may have substituted some special characters that serve as indexical expressions, such as '.' for the current location in the path component and '..' as the previous level in the path component. So, `../EiffelTower` is a perfectly acceptable relative URI. Relative URIs have a straightforward translation into absolute URIs, and it is trivial to compare absolute URIs for equality (Berners-Lee et al., 2005).

The 'hash' (#) and 'question mark' (?) are special characters at the end of URI. The question mark denotes 'query string.' The 'query string' allows for the parametrization of the HTTP request, typically in the cases where the HTTP response is created dynamically in response to specifics in the HTTP request. The 'hash' traditionally declares a ***fragment identifier***, which *identifies fragments of a hypertext document* but according to the TAG, it can also identify a "secondary resource," which is defined as "some portion or subset of the primary resource, some view on representations of the pri-

mary resource, or some other resource defined or described by those representations" where the "primary resource" is the resource identified by the URI without reference to either a hash or question mark (Jacobs and Walsh, 2004). The fragment identifier (specified by a 'hash' followed by some string of characters) is stripped off for the request to the server, and handled on the client side. Often the fragment identifier causes the local client to go to a particular part of the accessed HTTP entity. If there was a web-page about Gustave Eiffel, its introductory paragraph could be identified with the URI `http://www.example.com/EiffelTower#intro`. Figure 4.3 examines a sample URI, `http://www.example.org/EiffelTower#intro`:

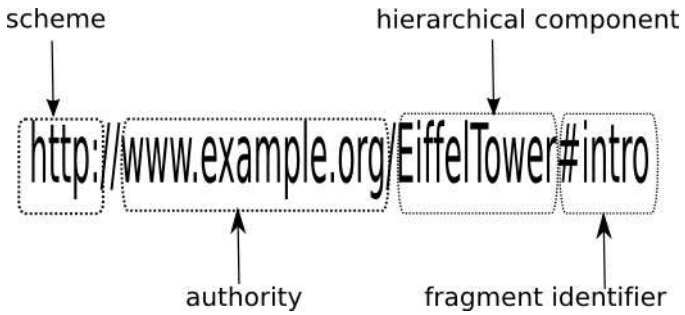

Figure 4.3: An example URI, with components labeled.

The first feature of URIs, the most noticeable in comparison to IP addresses, is that they can be human-readable, although they do not have to be. As an idiom goes, URIs can be 'written on the side of a bus.' URIs can then have an interpretation due to their use of terms from natural language, such as `www.whitehouse.gov` referring to the White House or the entire executive branch of the United States government. Yet it is considered by the W3C TAG to be ill-advised for any agent to depend on whatever information they can glean from the natural language terms used in URI itself, since to a machine the natural language terms used by the URI have no interpretation. For an agent, all URIs are opaque, with each URI being just a string of characters that can be used to either refer to or access information, and so syntactically it can only be checked for equality with other URIs and nothing more. This is captured well by the good practice of ***URI opacity***, which states that "agents making use of URIs should not attempt to infer properties of the referenced resource" (Jacobs and Walsh, 2004). To rephrase, we could state that *a URI should never itself have an interpretation, only the information referred to or accessed by that URI should have an interpretation.* This point becomes crucial in trying to determine 'what a URI identifies' as inspected in

detail in Chapter 6.

Second, a URI has an owner. The ***owner*** is *the agent that is accountable for originally determining what the URI identifies*. Usually for URIs schemes such as HTTP, where the hierarchical component begins with an authority, the owner of the URI is simply whoever controls that authority. In HTTP, since URIs can delegate their relative components to other users, the owner can also be considered the agent that has the ability to create and alter the information accessible from the URI, not just the owner of the authority. Each scheme should in theory specify what ownership of a URI means in context of the particular scheme.

4.1.3 Resources and Web Representations

While we have explained what a URI *does* in terms of the Internet, we have yet to define what a URI *is*. To inspect the acronym itself, a Uniform Resource Identifier (URI) is an identifier for a 'resource.' Yet this does not solve any terminological woes, for the term 'resource' is undefined in the earliest specification for "Universal Resource Identifiers" (Berners-Lee, 1994a). Berners-Lee has remarked that one of the best things about resources is that for so long he never had to define them (Berners-Lee, 2000). Eventually Berners-Lee attempted to define a resource as "anything that has an identity" (Berners-Lee et al., 1998). Other specifications were slightly more detailed, with Roy Fielding, one of the editors of HTTP, defining (apparently without the notice of Berners-Lee) a resource as "a network data object or service" (Fielding et al., 1999). However, at some later point Berners-Lee decided to generalize this notion, and in some of his later works on defining this slippery notion of 'resource,' Berners-Lee was careful not to define a resource only as information that is accessible via the Web, since not only may resources be "electronic documents" and "images" but also "not all resources are network retrievable; e.g., human beings, corporations, and bound books in a library" (Berners-Lee et al., 1998). Also, resources do not have to be singular but can be a "collection of other resources" (Berners-Lee et al., 1998).

Resources are not only a concrete realization or sets of possible realizations at a given temporal juncture, but are a looser category that includes things that change over time, as "resources are further carefully defined to be information that may change over time, such as a service for today's weather report for Los Angeles"(Berners-Lee et al., 1998). Obviously, a web-page with 'today's weather report' is going to change over time, so what is it that unites the notion of a resource over time? One early IETF RFC

for URIs, RFC 2396, defines this tentatively as a 'conceptual mapping' (presumably located in the head of an individual creating the representations for the resource) such that "the resource is the conceptual mapping to an entity or set of entities, not necessarily the entity which corresponds to that mapping at any particular instance in time. Thus, a resource can remain constant even when its content – the entities to which it currently corresponds – changes over time, provided that the conceptual mapping is not changed in the process" (Berners-Lee et al., 1998). This obviously begs an important question: If resources are identified as conceptual mappings in the head of an individual(s), then how does an agent know, given a URI, what the resource is? Is it our conceptual mapping, or the conceptual mapping of the owner, or some consensus conceptual mapping? This question and further questions of identity come to center stage in Chapter 6. The latest version of the URI specification deletes the confusing jargon of "conceptual mappings" and instead re-iterates that URIs can also be things above and beyond concrete individuals, for "abstract concepts can be resources, such as the operators and operands of a mathematical equation" (Berners-Lee et al., 2005). After providing a few telling examples of precisely how wide the notion of a resource is, the URI specification finally ties the notion of resource directly to the act of identification given by a URI, for "this specification does not limit the scope of what might be a resource; rather, the term 'resource' is used in a general sense for whatever might be identified by a URI" (Berners-Lee et al., 2005). Although this definition seems at best tautological, the intent should be clear. A ***resource*** is *any thing capable of having a sense*, or in other words, an 'identity' in a language. Since a sense is not bound to particular encoding, in practice within certain protocols that allow access to information, *a resource is typically not a particular encoding of a sense but a sense that can be given by many encodings*. To rephrase in terms of sense, *the URI identifies a sense on a level of abstraction, not the encoding of the sense or a particular realization of the sense*. So, a URI identifies the 'sense' of the Eiffel Tower, even if the web-accessible realization of it in the form of a web-page was accessible from that URI.

However, while this is best practice on the Web, there is nothing to forbid someone from identifying a particular encoding of information with its own URI and resource. For example, one could also have a distinct URI for a webpage about the Eiffel Tower in English, or a webpage about the Eiffel Tower in English in HTML. In other words, a resource can identify anything at a level of abstraction, and the same thing, such as a web-page, can be given *multiple URIs*, each corresponding to a *different level of abstraction*. Furthermore, due to the decentralized nature of URIs, often different agents

create *multiple URIs for the same sense*, which are then called in Web architecture ***co-referential URIs***.

We illustrate these distinctions in a typical HTTP interaction in Figure 4.4, where an agent via a web browser wants to access some information about the Eiffel Tower via its URI. While on a level of abstraction a protocol allows a user-agent to identify some resource, what the user-agent usually accesses concretely is some realization of that resource in a particular encoding, such as a webpage in HTML or a picture in the JPEG language (Pennebaker and Mitchell, 1992). In our example, the URI is resolved using the domain name system to an IP address of a concrete server, which then transmits to the user-agent some concrete bits that realize the resource, i.e. that can be interpreted to the sense identified by the URI. In this example, most of the interactions are local, since the webpage *encodes* the sense of the resource. This HTTP entity can then be interpreted by a browser as a rendering on the screen of Ralph's browser. Note this is a simplified example, as some status codes like `307` may cause a redirection to yet another URI and so another server, and so on possibly multiple times, until an HTTP entity may finally be retrieved.

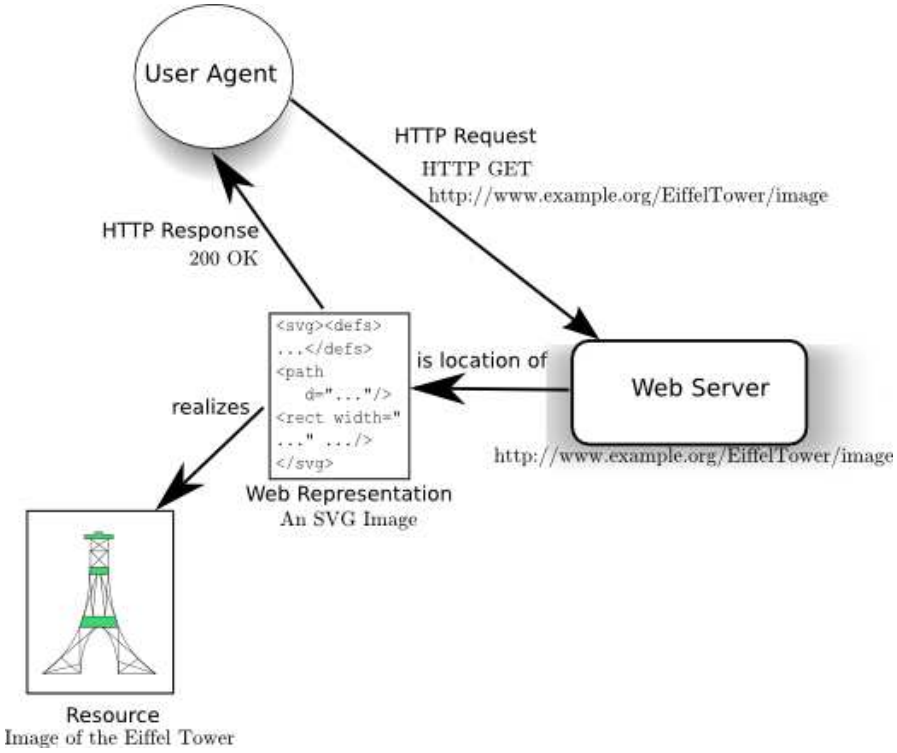

Figure 4.4: A user agent accessing a resource

One of the most confusing issues of the Web is that a URI does not necessarily

retrieve a single HTTP entity, but can retrieve multiple HTTP entities. This leads to a surprising and little-known aspect of Web architecture known as content negotiation. ***Content Negotiation*** is *a mechanism defined in a protocol that makes it possible to respond to a request with different Web representations of the same resource depending on the preference of the user-agent*. This is because information may have multiple encodings in different languages that all encode the same sense, and thus the same resource should have a singular URI. A representation on the Web is then just "an entity that is subject to content negotiation" (Fielding et al., 1999). Historically, the term 'representation' on the Web was originally defined in HTML as "the encoding of information for interchange" (Berners-Lee and Connolly, 1993). A later definition given by the W3C did not mention content negotiation explicitly, defining a representation on the Web as just "data that encodes information about resource state" (Jacobs and Walsh, 2004). To descend further into a conceptual swamp, "representation" is one of the most confusing terms in Web architecture, as the term 'representation' is used differently across philosophy. In order to distinguish the technical use of the term 'representation' within Web architecture from the philosophical use of the term "representation," we shall use the term 'Web representation' to distinguish it from the ordinary use of the term 'representation' as given earlier in Section 3.6. A ***Web representation*** is *the encoding of the sense given by a resource given in response to a request*, which must then include any headers that specify an interpretation, such as character encoding and media type. So a Web representation can be considered to have *two* distinct components, and the headers such as the media type that lets us interpret the encoding, and the payload itself, which is the encoding of the state of the resource at a given point in time. Notice that Web representations, being digital information, can be perfectly realized by messages, and the realization of a particular Web representation is the concrete bits sent across the 'wire' at a given point in space and time. Also, ***web-pages*** are *Web representations given in HTML*. Lastly, note that while HTTP entities can be a request (such as using HTTP PUT) and response from a server, Web representations can only be given as a response to a request like HTTP GET.

Our typical Web transaction, as given earlier in Figure 4.4, can become more complex due to this possible separation between sense and encoding on the Web. Different kinds of Web representations can be specified by user-agents as preferred or acceptable, based on the preferences of its users or its capabilities, as has been explained in Section 4.1.1. The owner of a web-site about the Eiffel Tower decides to host a resource for images of the Eiffel Tower. The owner creates a URI for this resource,

`http://www.eiffeltower.example.org/image`. Since a single URI is used, the sense (the depiction) that is encoded in either SVG or JPEG is the same, namely that of an image of the Eiffel Tower, that is, there are two distinct encodings of the image of the Eiffel Tower available on a server in two different iconic languages, one in a vector graphic language known as SVG and one in a bitmap language known as JPEG (Ferraiolo, 2002; Pennebaker and Mitchell, 1992). These encodings are rendered identically on the screen for the user. If a web-browser only accepted JPEG images and not SVG images, the browser could request a JPEG by sending a request for `Accept: image/jpeg` in the headers. Ideally, the server would then return the JPEG-encoded image with the HTTP entity header `Content-Type: image/jpeg`. Had the browser wished to accept the SVG picture as well, it could have put `Accept: image/jpeg, image/svg+xml` and received the SVG version. In Figure 4.5, the user agent specifies its preferred media type as `image/jpeg`. So, both the SVG and JPEG images are Web representations of the same resource, an image of the Eiffel Tower, since both the SVG and JPEG information realize the same content, albeit using different languages for encoding. Since a single resource is identified by the same URI `http://www.example.org/EiffelTower/image`, different user-agents can get a Web representation of the resource in a language they can interpret, even if they cannot all interpret the same language.

In Web architecture, content negotiation can also be deployed over not only differing formal languages, but differing natural languages, as the same content can be encoded in different natural languages such as French and English. An agent could request the description about the Eiffel Tower from its URI and set the preferred media type to '`Accept-Language: fr`' so that they receive a French version of the webpage as opposed to an English version. Or they could set their preferred language as English but by using '`Accept-Language: en`.' The preferences specified in the headers are not mandatory for the server to follow, the server may only have a French version of the resource available, and so send the agent a French version of the description, encoded in HTML or some other formal language, regardless of their preference. This extension of content negotiation to operate over different natural languages can be considered controversial. Different natural languages may not be able to encode the same content. Is it really true that two different languages can, even on a high level of abstraction, encode the same information? In some cases, this seems reasonable. Yet it is well-known there are some words in French that are difficult if not impossible to translate into English, such as 'frileusement.' Indeed, saying that

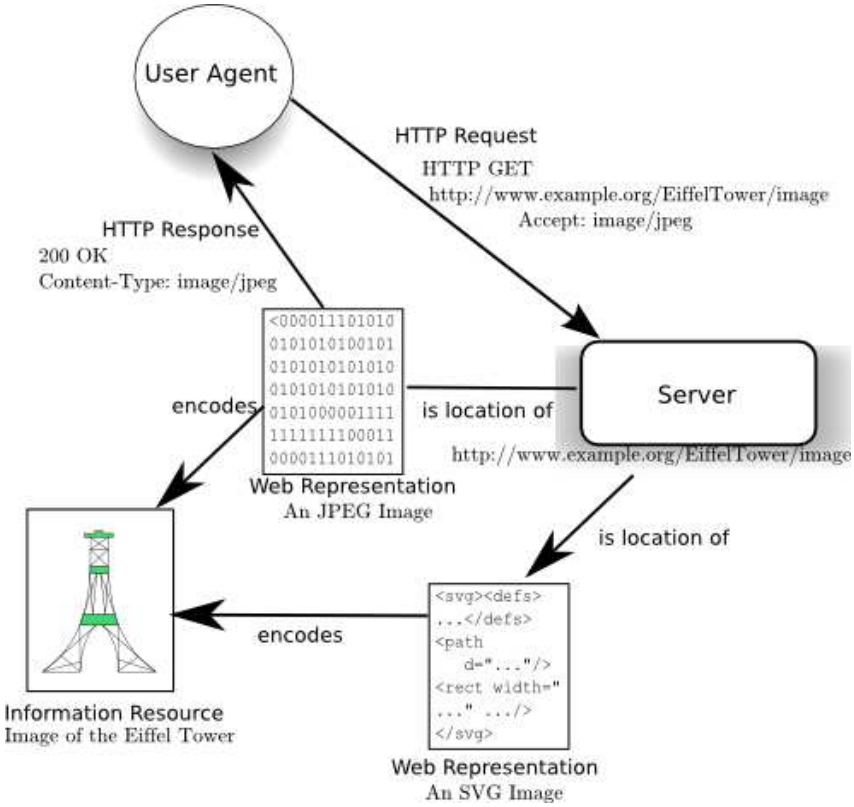

Figure 4.5: A user agent accessing a resource using content negotiation

one natural language encodes the same content as another natural language is akin to hubris in the general case. If this is the case, then it is perfectly reasonable to establish different resources and so URIs for the French and English language encodings of the resource, such as `http://www.eiffeltower.example.org/francais` and `http://www.eiffeltower.example.org/english`. In fact, if one believes the same image cannot be truly expressed by both SVG and JPEG image formats, one could give them distinct URIs as well. Regardless, what Figure 4.5 shows is that the Web representations are distinct from the resource, even if the Web representations are bound together by realizing the same information given by a resource, since accessing a resource via a single URI can return *different* Web representations depending on content negotiation.

The only architectural constraint that connects Web representations to resources is that they are retrieved by the same URI. So one could imagine a resource with a URI called `http://www.example.org/Moon`, that upon accessing using English as the preferred language would provide a web-page with a picture of the moon, and upon accessing with something other than English as the preferred language would provide a

picture of blue cheese. While this seems odd, this situation is definitely possible. What binds Web representations to a resource? Is a resource *really* just a random bag of Web representations? Remember that the answer is that the Web representations should have the same *sense* regardless of their particular encoding if it is accessible from the same URI. This notion depends on our notion of informational content (sense) as given in Section 3.2, which we define by an appeal to Dretske's semantic theory of information (Dretske, 1981). To recall, Dretske's definition of semantic information, "a signal r carries the information that s is F when the conditional probability of s's being F, given r (and k) is 1 (but, given k alone, less than 1). k is the knowledge of the receiver" (1981). We can then consider the signal r to be a Web representation, with s being a resource and the receiver being the user-agent. However, instead of some fact F about the resource, we want an interpretation of the Web representation by *different* user-agents to be to the same sense. Of course, one cannot control the interpretations of yet unknown agents, so all sorts of absurdities are possible in theory. As the interpretation of the same encoding can differ among agents, there is a possibility that the owner of the URI `http://www.example.org/Moon` really thinks that for French speakers a picture of blue cheese has the same sense as a picture of the Moon for English speakers, even if users of the resource disagree. However, it should be remembered that the Web is a space of communication, and that for communication to be successful over the Web using URIs, it is in the interest of the owner of the resource to deploy Web representations that they believe the users will share their interpretation. So content negotiation between a picture of blue cheese and a picture of the moon for a resource that depicts the Moon is, under normal circumstances, the Web equivalent of insanity at worse or bad manners at best. From a purely normative viewpoint in terms of relevant IETF and W3C standards, it is left to the owner to determine whether or not two Web representations are equivalent and so can be hosted using content negotiation at the same URI.

The key to content negotiation is that the owner of a URI never knows what the capabilities of the user-agent are, what natural and formal languages are supported by it. This is analogous to what Dretske calls the "knowledge" or k of the receiver (1981). The responsibility of the owner of a URI should be, in order to share their resource by as many user-agents as possible, to provide as many Web representations in a variety of formats as they believe are reasonably necessary. So, the owner of the URI for a website about the Eiffel Tower may wish to have a number of Web representations in a wide variety of languages and formats. By failing to provide a Web representation

in Spanish, they prevent speakers of only Spanish from accessing their resource. Since the owner of a URI cannot reasonably be expected to predict the capabilities of all possible user-agents, the owner of the URI should try their best to communicate their interpretation within their finite means.

The reason URIs identify resources, and not individual Web representations, is that Web representations are too ephemeral to want to identify in of themselves, being by definition the response of a server to a *particular* response and request for information. While one could imagine wanting to access a particular Web representation, in reality what is usually wanted by the user-agent is the sense of the resource, which may be present in a wide variety of languages. What is important is that the content (sense) gets transferred and interpreted by the user agent, not the individual bytes of a particular encoding in a particular language at a particular time.

With this insight in hand, some clarification on the relationship between representations, resources, and URIs should be given. First, a URI may identify only a single resource, as otherwise multiple resources would have both the same URI and an identical set of Web representations with the same sense, and so the resources would be indistinguishable. The opposite of this is *when the same resource has multiple URIs*, which is called ***URI collision***, and *URIs that identify the same resource* are considered ***co-referential URIs*** (Jacobs and Walsh, 2004). However, a single URI may not be identified only with its currently accessible Web representations, since those representations may change in the future as the resource changes. A resource for the weather in Paris will have to change in order to remain accurate. Likewise, two sets of otherwise identical Web representations may be for different resources. These Web representations may be identical at one point in time but diverge in the future. A resource for pictures of the tallest monument in Paris would (at the time of writing) be encoded by the same Web representations as a picture of the Eiffel Tower but if an even larger monument was built in Paris, then the Web representations for the two resources would diverge.

4.2 The Principles of Web Architecture

In light of having both the philosophical terminology defined in Chapter 3 and the terminology of the Web defined Section 4.1, it is now possible to show how the various Web terms are related to each other in a more systematic way. These relationships are phrased as five finite principles that serve as the normative Principles of Web architec-

ture: The Principles of Universality, Linking, Self-Description, the Open World, and Least Power. In practice many applications violate these principles, and by virtue of their use of URIs and the HTTP protocol, many of these applications would be in some sense 'on the Web.' However, these principles are normative insofar as they define what could be considered as compliance with Web architecture, and so an application that embodies them is compliant with Web architecture.

4.2.1 Principle of Universality

The ***Principle of Universality*** states that *any resource can be identified by a URI*. The notion of both a resource and a URI were from their onset universal in ambition, as Berners-Lee said, "a common feature of almost all the data models of past and proposed systems is something which can be mapped onto a concept of 'object' and some kind of name, address, or identifier for that object. One can therefore define a set of name spaces in which these objects can be said to exist. In order to abstract the idea of a generic object, the web needs the concepts of the universal set of objects, and of the universal set of names or addresses of objects" (1994a). The more informal notes of Berners-Lee are even more startling in their claims for universality, stating that the first 'axiom' of Web architecture is "Universality" where "by 'universal' I mean that the Web is declared to be able to contain in principle every bit of information accessible by networks" (1996c). Although it appears the germ of the idea of universality was clearly present in the earliest IETF Internet Drafts for 'Universal Resource Identifiers' in IETF 1630 (Berners-Lee, 1994a), in works like HTTP IETF RFC 1945 with co-authors like Fielding, Berners-Lee constrained himself to only talk about digital 'network data objects' that are accessible over the Internet (1996). However, in later IETF RFCs like RFC 2396, the principle quickly ran amok, as URIs were allowed to refer to "human beings, corporations, and bound books in a library" (Berners-Lee et al., 1998).

There seems to be a certain way that web-pages are 'on the Web' in a way that human beings, corporations, unicorns, and the Eiffel Tower are not. Accessing a web-page in a browser means to receive some bits, while one cannot easily imagine what accessing the Eiffel Tower itself or the concept of a unicorn in a browser even means. This property of being 'on the Web' is a common-sense distinction that separates things like a web-page about the Eiffel Tower from things like the Eiffel Tower itself. The core of the problem is that the use of term 'identify' in URIs is overloaded with two

distinctions. This distinction is a matter of between the use of URIs to *access* and *reference*, between using the URI to access local and refer to the distal. The early notes of Berners-Lee address this distinction between access and reference, phrasing it as a distinction between locations and names. As Berners-Lee states, "conventionally, a 'name' has tended to mean a logical way of referring to an object in some abstract name space, while the term 'address' has been used for something which specifies the physical location" (1991). So, a ***location*** is *a term that can be used to access the thing*, while a ***name*** is *a term that can be used to refer to a thing*. Unlike access, reference is the use of an identifier for a thing *to which one is immediately causally disconnected*. ***Access*** is *the use of an identifier to create immediately a causal connection to the thing identified* (Hayes and Halpin, 2008). The difference between the use of a URI to access a hypertext web-page or other sort of information-based resource and the use of a URI to refer to some non-Web accessible entity or concept ends up being quite important, as this ability to representationally use URIs as 'stands-in' for referents forms the basis of the distinction between the hypertext Web and the Semantic Web.

As noticed in Chapter 3, names can serve as identifiers for distal things. However, Berners-Lee immediately puts forward the hypothesis that "with wide-area distributed systems, this distinction blurs" so that "things which at first look like physical addresses...cease to give the actual location of the object. At the same time, a logical name...must contain some information which allows the name server to know where to start looking" (1991). He posits a third neutral term, "identifier" that was "generally referred to a name which was guaranteed to be unique but had little significance as regards the logical name or physical address" (Berners-Lee, 1991). In other words, an ***identifier*** is a *term that can be used to either access or refer, or both access and refer to, a thing*. The problem at hand for Berners-Lee was how to provide a name for his distributed hypertext system that could get "over the problem of documents being physically moved" (1991). Using simple IP addresses or any scheme that was tied to a single server would be a mistake, as the resource that was identified on the Web should be able to move from server to server without having to change identifier.

For at least the first generation of the Web, the way to overcome this problem was to provide a translation mechanism for the Web that could provide a methodology for transforming "unique identifiers into addresses" (Berners-Lee, 1991). Mechanisms for translating unique identifiers into addresses already existed in the form of the domain name system that was instituted by the IETF in the early days of the expansion of ARPANet (Mockapetris, 1983). Before the advent of the domain name system, the

ARPANet contained one large mapping of identifiers to IP addresses that was accessed through the Network Information Center, created and maintained by Engelbart (Hafner and Lyons, 1996). However, this centralized table of identifier-to-address mappings became too unwieldy for a single machine as ARPANet grew, so a decentralized version was conceived based on ***domain names***, where each domain name is *a specification for a tree structured name space, where each component of the domain name (part of the name separated by a period) could direct the user-agent to a more specific 'domain name server' until the translation from an identifier to the name to IP address was complete.*

Many participants in the IETF felt like the blurring of this distinction that Berners-Lee made was incorrect, so URIs were bifurcated into two distinct specifications. ***Uniform Resource Locations*** (URLs) are *a scheme for locations that allowed user-agents via an Internet protocol to access a realization of information* (Berners-Lee et al., 1994). In contrast, ***Uniform Resource Names*** (URNs) are *a scheme whose names that could refer to things outside of the causal reach of the Internet* (Sollins and Masinter, 1994). Analogue things like concepts and entities naturally had to be given URNs, and digital information that can be transmitted over the Internet, like web-pages, were given URLs. Interestingly enough, URNs count *only* as a naming scheme, as opposed to a protocol like HTTP, because they cannot access any information. While one could imagine a particular Web-accessible realization, like a web-page, disappearing from the Web, it was felt that identifiers for things that were not accessible over the Web should "be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name" (Mealling and Daniel, 1999). Precisely because of their lack of ability to access information, URNs never gained much traction, while URLs to access web-pages became the norm. Building on this observation about the "blurring of identifiers," the notion of URIs implodes the distinction between identifiers used only for access (URLs) and the identifiers used for reference (URNs).

A ***Uniform Resource Identifier*** is *a unique identifier that may be used to either or both refer to or access a resource*, whose syntax is given in the latest URI IETF RFC, currently (Berners-Lee et al., 2005). URIs subsume both URLs and URNs, as shown in Figure 4.6. Berners-Lee and others were only able to push this standard through the IETF process years after the take-off of the Web. Indeed, early proposals for universal names, ranging from attempts to find the 'true' names of things in various mystical traditions to Engelbart's 'Every Object Addressable' principle (1990), all missed the

crucial advantage of the Web: Classically names in natural language are usually used for reference, yet on the Web names are can also used to access information. In a decentralized environment this is crucial for discovering the sense of a URI, as illustrated by the notions of 'linking' and 'self-description' detailed next in Section 4.2.2 and Section 4.2.3.

The fact that URIs can be used as names to access as opposed to just refer to information is *not* a direct contrast between the use of names in natural language and the use of URIs on the Web. Trivially, names in natural language can access things as well, such as when one is knocking on a door and says "Ralph, come open the door!" or when one picks a friend out of a large crowd by simply yelling their name. Furthermore, these examples of natural language use of names to access holds in an even more interesting fashion for information that can be realized by the message itself. One example of this would be the sentence "In 'Moby Dick,' one has the immortal opening line 'Call me Ishmael'..." where it is clear that the name 'Moby Dick' refers to some of the text which is directly uttered in the same sentence. So the difference between URIs being used for access as opposed to names being used in natural language for reference is not an absolute distinction, but simply two different kinds of functions that both kinds of names, both URIs on the Web and names in natural language, can perform. The matter is more one of emphasis; names in natural language have a tendency to be used often for reference in speech, as the amount of things that are distal that an agent may wish to talk about far outweighs the amount of things in their immediate vicinity they could also discuss. Likewise, in common parlance, URIs on the Web are almost synonymous for their ability to access web-pages. Many people would even not even consider the fact that a URI can be used to refer to some thing to be important, as crucial as this usage is for the Semantic Web. Thus, names of any sort can usually be used for both access and reference, but URIs are often mostly used for access while natural language names are more often used for reference. So one of the tricks the Semantic Web has to play is to convey to agents that URIs should be used for reference as well, in other words, to treat URIs more like natural language names.

4.2.2 Principle of Linking

The ***Principle of Linking*** states that *any resource can be linked to another resource identified by a URI*. No resource is an island, and the relationships between resources are captured by the linking, transforming lone resources into a Web. A ***link*** is *a connec-*

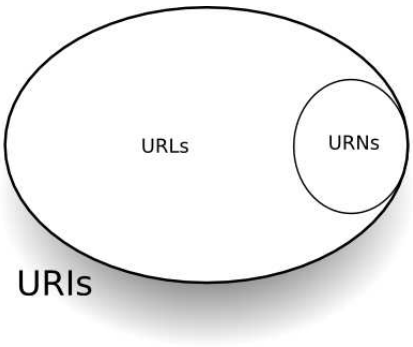

Figure 4.6: A Venn Diagram describing the relationships between URIs, URNs, and URLs

tion between resources. The *resource that the link is directed from* is called its ***starting resource*** while the *resource a link is directed to* is the ***ending resource*** (DeRose et al., 2001).

What are links for? Just as URIs links may be used for either access or reference, or even both. In particular, in HTML the purpose of links is for access to additional hypertext documents, and so they are sometimes called 'hyperlinks.' This access is often called *following* the link, a traversal from one Web representation to another, that results in access to Web representations of the ending resource. A unidirectional link that allows access of one resource from another is the predominant kind of link in hypertext. Furthermore, access by linking is transitive, for if a user-agent can access a Web representation of the ending resource from the starting resource, then it can access any links present in the Web representation, and thereby access a Web representation of an ending resource. It is precisely this ability to transitively access documents by following links that led the original Web to be a seamless Web of hypertext. While links can start in Web representations, the main motivation for using URIs as the ending resource of a link as opposed to a specific Web representation is to prevent *broken links*, where a user-agent follows a link to a resource that is no longer there, due to the Web representation itself changing. As put by the W3C TAG, "Resource state may evolve over time. Requiring a URI owner to publish a new URI for each change in resource state would lead to a significant number of broken references. For robustness, Web architecture promotes independence between an identifier and the state of the identified resource" (Jacobs and Walsh, 2004).

However, one of the distinguishing features of the Web is that links may be broken by having access to a Web representation disappear, due to simply the lack of

hosting a Web representation, loss of ownership of the domain name, or some other reason. These reasons are given in HTTP status codes, such as the infamous `404 Not Found` that signals that while there is communication with a server, the server does not host the resource. Further kinds of broken links are possible, such as `301 Moved Permanently` or a five hundred level server error, or an inability to even connect with the server leading to a time-out error. This ability of links to be 'broken' contrasts to previous hypertext systems. Links were not invented by the Web, but by the hypertext research community. Constructs similar to links were enshrined in the earliest of pre-Web systems, such as Engelbart's *oNLine System* (NLS) (1962), and were given as part of the early hypertext work by Theodor Nelson (1965). The plethora of pre-Web hypertext systems were systematized into the Dexter Reference Model (Halasz and Schwartz, 1994). According to the Dexter Reference Model, the Web would not even qualify as hypertext, but as "proto-hypertext," since the Web did not fulfill the criteria of "consistency," which requires "in creating a link, we must ensure that all of its component specifiers resolve to existing components" (Halasz and Schwartz, 1994). To ensure a link must resolve and therefore not be broken, this mechanism requires a centralized link index that could maintain the state of each resource and not allow links to be created to non-existent or non-accessible resources. Many early competitors to the Web like HyperG had a centralized link index (Andrews et al., 1995). As an interesting historical aside, it appears that the violation of this principle of maintaining a centralized link index was the main reason why the World Wide Web was rejected from its first academic conference, ACM Hypertext 1991, although Engelbart did encourage Berners-Lee and Connolly to pursue the Web further.[3] While a centralized link index would have the benefit of not allowing a link to be broken, the lack of a centralized link index removes a bottleneck to growth by allowing the owners of resources to link to other resources without updating any index besides their own Web representations. This was doubtless important in enabling the explosive growth of linking. The lack of any centralized link index, and index of Web representations, is also precisely what search engines like Google create post-hoc through spidering, in order to have an index of links and web-pages that enable their keyword search and page ranking algorithms. As put by Dan Connolly in response to Engelbart, "the design of the Web trades link consistency guarantees for global scalability" (2002). So, broken links and `404 Not Found` status codes are purposeful *features*, not defects, of the Web.

[3] Personal communication with Tim Berners-Lee.

4.2.3 Principle of Self-Description

One of the goals of the Web is for resources to be 'self-describing,' currently defined as "individual documents become self-describing, in the sense that only widely available information is necessary for understanding them" (Mendelsohn, 2006). While it is unclear what "widely-available" means, one notion of "widely-available" is that in order for some sort of new information to have an interpretation, its interpretation must build on top of various implicit and 'common-sense' information that the interpreting agent already possesses (Mendelsohn, 2006). The idea that 'common-sense' information is crucial to intelligence and sharing information has long been held central by artificial intelligence (McCarthy, 1959). The question that confronts the Web is similar in many regards, but with a change of focus due to the open ended nature of the Web: Given a URI, how can an agent discover the interpretation of the URI? In many cases, the answer may be similar to how humans learn foreign languages, in which case the URI's interpretation can be given by its implicit context. However, due to the fact that the agents are often machines lacking the ability to rely on sophisticated common-sense interpretative capacities, often the additional information needed to interpret a URI needs to be made explicit. Of course, at some point even for machine agents there must be a base-line of capacity that allows the *some* information on the Web to be interpreted, but the question is how such interpretive abilities can be boot-strapped in the face of new and possibly unknown URIs and Web representations?

The ***Principle of Self Description*** states that *if an interpretation of a URI is not possible with the implicit capabilities of the agent, information that can aid an agent in discovering an interpretation of the URI should be accessible from the Web representation accessible from the URI.* Note that the interpretation of a URI can be grounded in the interpretations of Web representations accessible from the URI, or the use of the URI in other media. How many and what sort of links are necessary to adequately describe a resource? A resource is successfully described if an interpretation of a sense is possible. Any representation can have links to other resources which in turn can determine valid interpretations for the original resource. This process of following whatever data is linked in order to determine the interpretation of a URI is informally called 'following your nose' in Web architecture.

The ***Follow-Your-Nose algorithm*** states that if a user-agent encounters a representation in a language that the user-agent cannot interpret, the user-agent should, in order:

1. **Dispose of Fragment Identifiers:** As mandated by the URI specification (Berners-Lee et al., 2005), user-agents can dispose of the fragment identifier in order to retrieve whatever Web representations are available from the racine (the URI without fragment identifier). For example, in HTML the fragment identifier of the URI is stripped off when retrieving the webpage, and then when the browser retrieves a Web representation, the fragment identifier can be used to locate a particular place within the Web representation.

2. **Inspect the Media Type:** The media type of a Web representation provides a normative declaration of how to interpret a Web representation. Since the number of IETF media-types is finite and controlled by the IETF, a user-agent should be able to interpret these media types.[4]

3. **Follow any Namespace Declarations:** Many Web representations use a generic format like XML to in turn specify a customized dialect. In this case, a language or dialect is itself given a URI, called a ***namespace URI***, *a URI that identifies that particular dialect*. A namespace URI then in turn allows access to a ***namespace document***, *a Web representation that provides more information about the dialect*. In a Web representation using this dialect, a ***namespace declaration*** then *specifies the namespace URI*. In this case, the user-agent may follow these namespace declarations in order to get the extra information needed to interpret the Web representation. As a single Web representation may be encoded in multiple languages, it may have multiple namespace URIs.

4. **Follow any links:** The user-agent can follow any links. There are some links in particular languages that may be preferred, such as the ending resource of a `link` header in HTML or in RDF Schema links such as *rdfs:isDefinedBy* links, or links like OWL by the *owl:imports* (See Chapter 5 for the definition of RDF and OWL). If links are typed in some fashion, each language may define or recommend links that have the normative status, and normative links should be preferred. However, for many kinds of links, their normative status is unclear, so the user-agent may have to follow any sort of link as a last resort.

Using this algorithm, the user-agent can begin searching for some information that allows it to interpret the Web representation. It can follow the first three guidelines

[4]The finite list is available at *http://www.iana.org/assignments/media-types/*, and a mapping from media types to URIs has been proposed at *http://www.w3.org/2001/tag/2002/01-uriMediaType-9*.

and then follow the fourth, applying the above guidelines recursively. Eventually, this recursive search should bottom out either in a program that allows an interpretation of the Web representation, such as new inferences produced by the metadata gathered by the follow-your-nose algorithm or the natural bottoming out point of specifications given by the IETF in plain, human-readable text. This final fact brings up the point that the information that gets one an interpretation is not necessarily a program, but could be a human-readable specification that requires a human to make the mapping from the names to the intended sense.

4.2.4 The Open World Principle

The ***Open World Principle*** states that *the number of resources on the Web can always increase*. There can always be new acts of identification, carving out a new resource from the world and identifying it with a URI. At any given moment, a new webpage may appear on the Web, and it may or may not be linked to. This is a consequence of the relatively decentralized creation of URIs for resources given by the Principle of Universality and the decentralized creation of links by the Principle of Linking. Without any centralized link index, there is no central repository of the state of the *entire* Web. While approximations of the state of the entire Web are created by indexing and caching web-pages by search engines like Google, due to the Open World Principle, none of these alternatives will necessarily ever be guaranteed to be complete. Imagine a web-spider updating a search engine index. At any given moment, a new URI could be added to the Web that the web-spider may not have crawled, or a previously crawled Web representation may change. So to assume that any collection of resources of the Web can be a complete picture of the whole Web is at best impudent.

The ramifications of the Open World Principle are surprising, and most clear in terms of judging whether a statement is true or false. This repercussions transform the Open World Principle into its logical counterpart, the ***Open World Assumption***, which logically states that *statements that cannot be proven to be true cannot be assumed to be false*. Intuitively, this means that the world cannot be bound. On the Web, the Open World Principle holds that since the Web can always be made larger, with any given set of statements that allows an inference, a new statement relevant to that inference may be found. So any agent's knowledge of the Web is always partial and incomplete, and thus the Open World Assumption is a safe bet for agents on the Web. The Open World Principle is one of the most influential yet challenging principles of the Web, the one

that arguably separates the Web from traditional research in artificial intelligence and databases in practice. In these fields, systems tend to make the opposite of the Open World Assumption, the Closed World Assumption. The ***Closed World Assumption*** states that logically *statements that cannot be proven to be true can be assumed to be false*. Intuitively, this means that somehow the world can be bounded. The Closed World Assumption has been formalized on a number of different occasions, with the first formalization being due to Reiter (1978). *Negation as failure* is an implementation of the Closed World assumption in both logic programming and databases, where failure for the program to prove a statement is true implies the statement is false (Clark, 1978).

4.2.5 Principle of Least Power

The Principle of Least Power states that a *Web representation given by a resource should be described in the least powerful but adequate language*. This principle is also normative, for if there are multiple possible Web representations for a resource, the owner should chose the Web representation that is given in the 'least powerful' language. The Principle of Least Power seems odd, but it is motivated by Berners-Lee's observation that "we have to appreciate the reasons for picking not the most powerful solution but the least powerful language" (1996c). The reasons for this principle are rather subtle. The receiver of the information accessible from a URI has to be able to decode the language that the information is encoded in so the receiver can determine the sense of the encoding. Furthermore, an agent may be able to decode multiple languages, but the owner of the URI does not know what languages an agent wanting to access their URI may possess. Also, the same agent may be able to interpret multiple languages that can express the same sense. The question always facing any agent trying to communicate is: what language to use? In closed and centralized systems, this is ordinarily not a problem, since each agent can be guaranteed to use the same language. In an open system like the Web, where one may wish to communicate a resource to an unknown number of agents, each of which may have different language capabilities, the question of which language to deploy becomes nearly insurmountable. Obviously, if an agent is trying to encode some sense, then it should minimally choose a language which is capable of conveying that sense. Yet the same sense can be conveyed by different languages, as languages in effect encode systems of senses.

The Principle of Least-Power is a common-sense engineering solution to this prob-

lem of language choice. The solution is simply to build first a common core language that fulfills the minimal requirements to communicate whatever sense one wishes to communicate, and then extend this core language. Using HTML as an example, one builds first a common core of useful features such as the ability to have text be bold and have images inserted in general areas of the text, and then as the technology matures, to slowly add features such as the precise positioning of images and the ability to specify font size. The Principle of Least Power allows a straightforward story about compatibility to be built to honor the maxim that an agent should "be strict when sending and tolerant when receiving," since it makes the design of a new version an exercise in strictly extending the previous version of the language (Carpenter, 1996). A gaping hole in the middle of the Principle of Least Power is no consistent definition of the concept of 'power,' and the W3C TAG seems to conflate power with the Chomsky Hierarchy. At this stage, the problem of defining 'power' formally must be left as an open research question.

4.3 Conclusions

The Web, while to a large extent being an undisciplined and poorly-defined space, does contain a set of defining terms and principles. While previously these terms and principles have been scattered throughout various informal notes, IETF RFCs, and W3C Recommendations, in this chapter we have systematized both the terminology and the principles in a way that reveals how they internally build of each other. In general, when we are referring to the ***hypertext Web***, we are referring *to the use of URIs and links to access hypertext web-pages using HTTP*. Yet there is more to the Web than hypertext. The next question is how can these principles be applied to domains outside the hypertext Web, and this will be the topic of Chapter 5, as we apply these principles to the notion of a knowledge representation language for the Web, a vast project tantalizing called the 'Semantic Web.'

Chapter 5

The Semantic Web

All the important revolutions that leap into view must be preceded in the spirit of the era by a secret revolution that is not visible to everyone, and still less observable by contemporaries, and that is as difficult to express in words as it is to understand. **G.W. F. Hegel** (1959)

The Web is a universal information space, but so far it has been one composed entirely of hypertext documents. As said by Berners-Lee at the World Wide Web conference in 1994, "to a computer, then, the web is a flat, boring world devoid of meaning...this is a pity, as in fact documents on the web describe real objects and imaginary concepts, and give particular relationships between them" (1994b). The heart of this particular insight is the realization that it is the content – the sense – of the information, not its encoding in hypertext, that is of central importance to the Web. The purpose of the architecture of the Web is to connect information of any kind in a decentralized manner, and this architecture can be applied beyond the hypertext of its initial incarnation.

The next step in Berners-Lee's programme to expand the Web beyond hypertext is called the ***Semantic Web***. The most cited definition of the Semantic Web is given by Berners-Lee et al. as "*the Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*" (2001). How can information be added to the Web without encoding it in hypertext? The answer is to find a language capable of representing the information about "real objects and imaginary concepts." This requires a ***knowledge representation language****, a language whose primary purpose is the representation of non-digital content in a digital formal language*.

As the previous exposition of Web architecture explained in detail, resources on the Web are given by a URI that identifies the same sense on the Web across different encodings. What drives the Semantic Web is the realization that at least some of the information on the Web is primarily representational, i.e. information about distal content. Then instead of the hypertext language, which is mainly concerned with the presentation and linking of natural language for humans, the Web needs a knowledge representation language which describes the represented referents as fully as possible without regard to presentation for humans. The mixture of content and encoding for presentation forces web-spiders to 'scrape' valuable content out of hypertext. In theory, encoding information directly in a knowledge representation language gives a spider more reliable and direct access to the information. As Berners-Lee puts it, "most information on the Web is designed for human consumption, and even if it was derived from a database with well defined meanings (in at least some terms) for its columns, that the structure of the data is not evident to a robot browsing the Web" (1998b). This has led him to consider the Semantic Web as a Web "for expressing information in a machine processable form" and so making the Web "machine-understandable" (Berners-Lee, 1998b). This leads to the contrast between the Semantic Web as a 'Web of data' as opposed to the hypertext 'Web of documents.' W3C standards such as XML were originally created, albeit rarely used, precisely in order to separate content and presentation (Connolly, 1998).

Furthermore, the purpose of the Semantic Web is to expand the scope of the Web itself. Most of the world's digital information is not natively stored in hypertext. Instead, it is stored in databases and other non-hypertext documents and spreadsheets. As more and more of this information is being slowly but surely migrating to the Web via scripts that automatically and dynamically convert data from databases into HTML, the advocates of the Semantic Web imagine that by having a common knowledge representation language across the entire Web, all information that is not currently on the Web can become part of the Web. This makes the Semantic Web not a different and parallel Web to the hypertext Web, but an extension of the current Web, where hypertext serves as just one possible language.

5.1 A Brief History of Knowledge Representation

The creation of the Semantic Web then depends on the creation of a knowledge representation language for the Web, and so the Semantic Web inherits both the successes

and failures of previous efforts to create knowledge representation languages in artificial intelligence. The earliest work in digital knowledge representations was spearheaded by John McCarthy's attempts to formalize elements of human knowledge in first-order predicate logic, where the primary vehicle of intelligence was to be considered some form of inference (1959). These efforts reached their apex in Hayes's *Naive Physics Manifesto*, which called for parts of human understanding to be formalized as first-order logic. Although actual physics was best understood using mathematical techniques such as differential equations, Hayes conjectured that most of the human knowledge of physics, such as "water must be in a container for it not to spill" could be conceptualized better in first-order logic (1979). Hayes took formalization as a grand long-term challenge for the entire AI community to pursue, as he said that "we are never going to get an adequate formalization of common sense by making short forays into small areas, no matter how many of them we make" (1979). While many researchers took up the grand challenge of Hayes in various domains, soon a large number of insidious problems were encountered, primarily in terms of the expressivity of first-order logic and its undecidability of inference. In particular, first-order logic formalizations were viewed as not expressive enough, being unable to cope with temporal reasoning as shown by the Frame Problem, and so had to be extended with fluents and other techniques (McCarthy and Hayes, 1969). Since the goal of artificial intelligence was to create an autonomous human-level intelligence, another central concern was that predicate calculus did not match very well with how humans actually reasoned. For example, humans often use default reasoning, and various amendments must be made for predicate calculus to support this (McCarthy, 1980). Further efforts were made to improve first-order logic with temporal reasoning to overcome the Frame Problem, as well as the use of fuzzy and probabilistic logic to overcome issues brought up by default reasoning and the uncertain nature of some knowledge (Koller and Pfeffer, 1998).

Under increasing criticism from its own former champions like McDermott, first-order predicate calculus was increasingly abandoned by those in the field of knowledge representation (1987). McDermott pointed out that formalizing knowledge in logic requires that all knowledge be formalized as a set of axioms and that "it must be the case that a significant portion of the inferences we want...are deductions, or it will simply be irrelevant how many theorems follow deductively from a given axiom set" (1987). McDermott found that in practice neither can all knowledge be formalized and that even given some fragment of formalized knowledge, the inferences drawn

are usually trivial or irrelevant (1987). The debate focused on whether or not there was a more appropriate manner for AI to model human intelligence besides first-order logic. Some researchers championed a *procedural* view of intelligence that regarded the representation as itself irrelevant if the program could successfully solve some task given some input and output. This contrasted heavily with earlier attempts to formalize human knowledge that it was called the *declarative versus procedural* debate. Procedural semanticist Terry Winograd stated that "the operations on symbol structures in a procedural semantics need not correspond to valid logical inferences about the entities they represent" since "the symbol manipulation processes themselves are primary, and the rules of logic and mathematics are seen as an abstraction from a limited set of them" (1976). While the procedural view of semantics first delivered impressive results through programs like SHRDLU (Winograd, 1972), since the 'semantics' were ad-hoc and task-dependent, so they could not be used outside the limited domain in which they were created. Furthermore, there became a series of intense debates on whether these programs often purported to do what they wanted even within their domain, as Dreyfus argued that it was ridiculous that just because a program was labeled `UNDERSTAND` that it did actually in any way actually *understand* (1979). Interestingly enough, the debate between declarative and procedural semantics is, under the right formal conditions, a red herring since the Curry-Howard Isomorphism states that given the right programming language, there is a tight coupling between logical proofs and programs so that the simplification of proofs can be equivalent to steps of computation (Wadler, 2003).

Within AI, research began into other forms of declarative knowledge representation languages besides first-order logic that were supposed to be in greater concordance with human intelligence and that could serve as more stable substrates for procedural knowledge-based systems. Most prominent among these alternatives were ***semantic networks***, "a graphic notation for representing knowledge in patterns of interconnected nodes and arcs" (1987). Semantic networks are as old as classical logic, dating back to Porphyry's explanation of Aristotelian categories (Sowa, 1987). The term 'semantic network' was coined by Richard Richens to describe a common knowledge-representation system for machine-translation systems at the Cambridge Language Research Unit (1956). While the work at the Cambridge Language Research Unit moved more towards different knowledge representation languages to represent the underlying structure of thesauri, such as Masterman's semantic lattices and fans (1961), the simplistic 'node-arc-node' structure of semantic networks soon found favor elsewhere. Soon semantic networks were being used to represent everything from human mem-

ory to first-order logic itself (Quillian, 1968; Sowa, 1976). Semantic networks also continued to be used as an intermediate knowledge representation for natural language systems by systems like Shapiro's 'Semantic Network Processing System,' as the node and arc formulation computationally could be detected in the various dependencies given by words (1979). The approach of semantic networks was given some credibility by the fact that often when attempting to make diagrams of 'knowledge,' humans often start by drawing circles connected by lines, with each component labeled with some human-readable description. A semantic network about 'The architect of the Eiffel Tower was Gustave Eiffel' is given in Figure 5.1. Note that it refers declaratively to things in the world, but uses 'natural-language-like' labels on its nodes and edges.

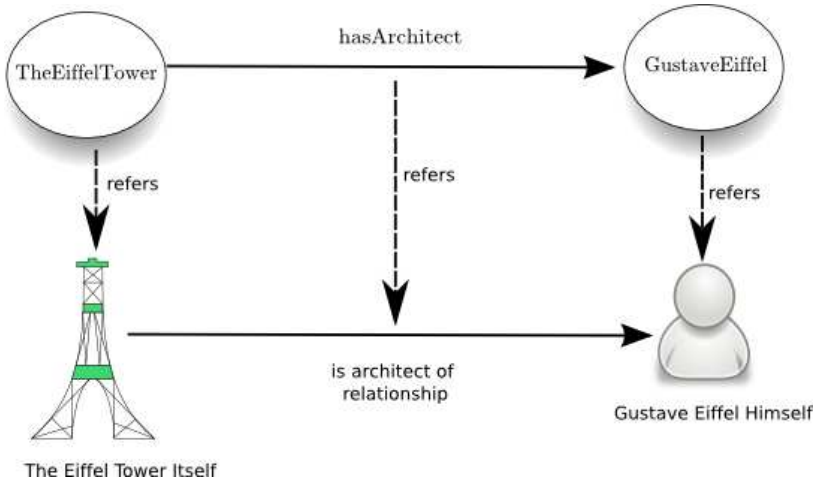

Figure 5.1: An example semantic network

When researchers attempted to communicate or combine their knowledge representation schemes, no-one really knew what the natural language description *meant* except the author, even when semantic networks were used as a formal language. The link in semantic networks was interpreted in at least three different ways (Woods, 1975) and no widespread agreement existed on the most common sort-of link, the `IS-A` link, which could represent both subclassing, instantiation, close similarity, and more. This led to an assault on semantic networks by champions of first-order logic like Hayes, who believed that by providing a formal semantics that defined 'meaning', first-order logic at least allowed knowledge representations to be transportable across domains, and that many alternative knowledge representations could be re-expressed in first order-logic (Hayes, 1977). In response, the field of knowledge representation bifurcated into separate disciplines. Many of the former champions of logic currently

do not believe that human intelligence can be construed as logical inference, but researchers still actively pursue the field as first order logic is of crucial importance to many systems such as mathematical theorem-proving and it is still used in many less ambitious knowledge-reasoning systems such as ISO Common Logic (Delugach, 2007).

The classical artificial intelligence programme, while fixated on finding a formal language capable of expressing human knowledge, had ignored the problem of inference. This problem came to attention abruptly when KRL (the self-titled Knowledge Representation Language), one of the most flexible knowledge representation languages pioneered by Winograd, was found to have intractable inference even on simple problems of cryptarithmetic, because of its representational richness (Bobrow and Winograd, 1977).[1] Furthermore, while highly optimized inference mechanisms existed for first-order logic, even first-order predicate logic was known to be undecidable. These disadvantages of alternative representational formats and first-order logic led many researchers, particularly those interested in *an alternative "slot and value" knowledge representation language* known as ***frames*** to begin researching the decidability of their inference mechanisms (Minsky, 1975). This research into frames then evolved into research on ***description logics***, where the trade-offs between the tractability and expressivity were carefully studied (Levensque and Brachman, 1987). The goal of the field was to produce a logic with decidable inference while maintaining maximum expressivity, as exemplified by languages like KL-ONE (Brachman and Schmolze, 1985). Although the first description-logic system, KL-ONE, was proven to have undecidable inference for even subsumption, later research produced a vast proliferation of description logics with carefully categorized decidability and features (Schmidt-Schauss, 1989).

Ultimately, the project of artificial intelligence to design a single knowledge representation system suitable for creating human-level intelligence has not yet succeeded and progress seems glacial at best. With no unifying framework, the field of artificial intelligence itself fragmented into many different diverse communities, each with its own family of languages and techniques. Researchers into natural language embraced statistical techniques and went back to practical language processing tasks, while logicians have produced an astounding variety of different knowledge representation languages, and cognitive scientists moved their interests towards dynamical systems and specialized biologically-inspired simulations. The lone hold-out seemed to be the Cyc

[1] Personal communication with Henry S. Thompson.

project, which continued to pursue the task of formalizing all 'common-sense' knowledge in a single knowledge representation language (Lenat, 1990). In one critique of Cyc, Smith instead asked what lessons knowledge representation languages could learn from hypertext, "Forget intelligence completely, in other words; take the project as one of constructing the world's largest hypertext system, with Cyc functioning as a radically improved (and active) counterpart for the Dewey decimal system. Such a system might facilitate what numerous projects are struggling to implement: reliable, content-based searching and indexing schemes for massive textual databases" (1991). Cantwell Smith's statement that strangely prefigures not only search engines, but the revitalization of knowledge representation languages due to the Semantic Web (1991).

5.2 The Resource Description Framework (RDF)

What makes knowledge representation language on the Web *different* from classical knowledge representation? Berners-Lee's early thoughts, as given in the first World Wide Web Conference in Geneva in 1994, were that "adding semantics to the Web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values" (Berners-Lee, 1994b). Having information in "machine-readable forms" requires a knowledge representation language that has some sort of relatively content-neutral syntax for encoding content (Berners-Lee, 1994b). The parallel to knowledge representation in artificial intelligence is striking, as it also sought to find one universal encoding, albeit encoding human-intelligence. The second point, of "allowing links," means that the basic model of the Semantic Web will be a reflection of the Web itself: the Semantic Web is constituted by connecting resources by links (Berners-Lee, 1994b). The Semantic Web is then easily construed as a descendant of semantic networks from classical artificial intelligence, where nodes are resources and arcs are links. Under the aegis of the W3C, the first knowledge representation language for the Semantic Web, the ***Resource Description Framework*** (RDF) was made a W3C Recommendation, and it is clearly influenced by work in AI on semantic networks. This should come as no surprise, for RDF was heavily inspired by the work of Ramanathan V. Guha on the Meta-Content Framework (MCF) (Guha, 1996). Before working on MCF, Guha was chief lieutenant of the aforementioned Cyc project, the last-ditch Manhattan project of classical artificial intelligence (R.V.Guha and D.Lenat, 1993). Another important influence on RDF besides semantic networks was the influence of semantic templates in information ex-

traction systems. As opposed to the ‘node-arc-node’ form, these templates normally had a ‘subject-verb-object’ form. Much of this influence from information extraction and computational linguistics in the design of RDF came from Tim Bray, who was hired by Netscape to transform Guha’s MCF system into RDF. Formerly, Bray was the manager of the project of digitizing the New Oxford English Dictionary and then later of the Open Text search engine, one of the Web’s first search engines. In fact, one of Guha’s first uses of RDF was as a light-weight knowledge representation system of subject-verb-object form for his ground-breaking ‘Semantic Search’ information extraction system (2003). There are nonetheless some key differences between semantic networks (and similar ‘subject-verb-object’ templates from information extraction) and RDF, as RDF was built in accordance with the Principles of Web Architecture as given in Chapter 4, as detailed in the next subsections.

5.2.1 RDF and the Principle of Universality

Semantic networks fell out of favor because of their use of ambiguous natural language terms to identify their nodes and arcs, which became a problem when semantic networks were transported between domains and different users, a problem that would be fatal in the decentralized and multi-lingual environment of the Web (Woods, 1975). According to the Principle of Universality, since a resource can be *anything*, then a component of the knowledge representation language should be considered a resource, and thus can be given a URI. Instead of labeling the arcs and nodes with natural language terms, in RDF all the arcs and nodes can be labeled with URIs. Although few applications had ever taken advantage of the fact before RDF, due to the Principle of Universality, URIs could be minted for things like the Eiffel Tower *qua* Eiffel-Tower, an absolute necessity for knowledge representation. Since the sense of statements in knowledge representation is usually about content in the world outside the Web, this means that the Semantic Web crucially depends on the rather strange fact that URIs can refer to things outside the Web.

This does not restrict the knowledge-representation language to merely refer to things that we would normally consider outside of the Web, since normal web-pages use URIs as well, and so the Semantic Web can easily be used to refer to normal web-pages. This has some advantages, as it allows RDF to be used to model the relationships between web-accessible resources, and even mix distal and proximal of relationships. This sort of “meta-data” is exemplified by the relationship between a

web-page and its human author, which in RDF would both be denoted by URIs. Lastly, this ability to describe everything with URIs leads to some unusual features, for RDF can then model its own language constructs using URIs, and make statements about its own core language constructs. However, just as all components of RDF may be considered resources, just as all resources may not have URIs, all components of RDF may not have URIs. For example, *a string of text or a number may be a component of RDF*, and these are called ***literals*** by RDF. In RDF *specified anonymous resources can not be given a URI*, and these are called ***blank nodes***. Yet it would be premature to declare that the deployment of URIs in RDF signal a major improvement over the natural language labels, for URIs can be just as ambiguous as natural language labels. A further analysis of the scope of this problem is in Chapter 6.

5.2.2 RDF and the Principle of Linking

The second step in Berners-Lee's vision for the Semantic Web, "allowing links to be created with relationship values," follows straightforwardly from the application of the Principle of Universality to knowledge representation. Since RDF is composed of resources, and any resource may link to another resource, then any term in RDF may be linked to another term. This linking forms the heart of RDF, as it allows disparate URIs to be linked together in order for statements in RDF to be made. The precise form of a statement in RDF is a ***triple***, which consists of two resources connected by a link, as shown in Figure 5.2. This use of RDF shows off the flexibility of using URIs and links for reference instead of access. Lastly, this use of URIs and links *outside* Web representations like those of hypertext web-pages shows the flexibility of the linking paradigm, as RDF is an example of the use of the idea of a 'linkbase' that was developed in the hypertext community, in particular in the Microcosm hypertext system (Fountain et al., 1990).

Any Web representation that contains *as its information some form of Semantic Web language* such as RDF is called a ***Semantic Web document***. There are several options for encoding Semantic Web documents. The W3C standardized an encoding of RDF is in a verbose XML format called 'RDF/XML' and a simpler encoding called *Turtle* for triples. (Beckett and Berners-Lee, 2008). In Turtle, a triple is three space-delimited terms (the subject, predicate, and object) ended in a period. Using namespaces, with `http://www.example.org/` being abbreviated as `ex`, one abbreviates the example in Figure 5.2 to `ex:EiffelTower ex:hasArchitect ex:Gustave_Eiffel`.

Comparing the example given in Figure 5.2 to Figure 5.1, the *only* noticeable difference between RDF and a classical semantic network is the use of URIs.

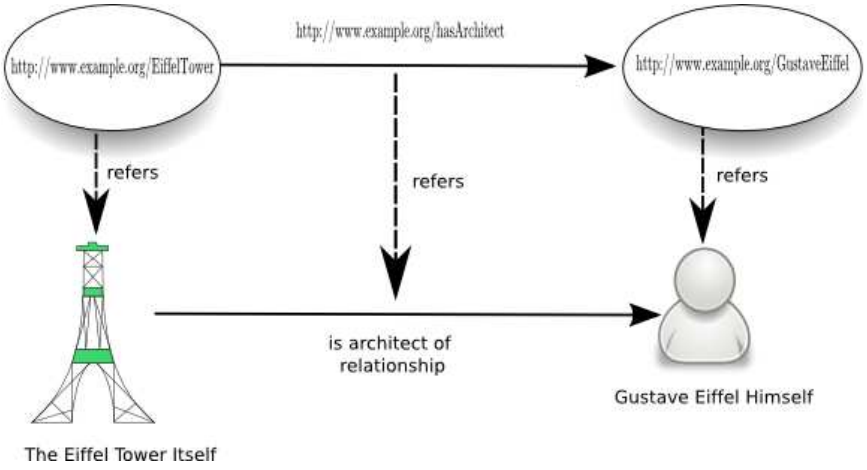

Figure 5.2: An example RDF statement

There are some restrictions to linking on the Semantic Web. As opposed to the vast numbers and kinds of links possible in XLink, linking on the Semantic Web is directed, like hyperlinks (DeRose et al., 2001) . *The starting resource in the triple* is called the ***subject***, while *the link itself* is called the ***predicate***, and *the ending resource in the triple* is the ***object***. The predicate is usually a role as opposed to an arc role. The major restriction on the Semantic Web is that the subject must be a URI or a blank node, and the predicate must also be a URI. The object, on the other hand, is given the most flexibility, as it may either be a URI, a blank node, or a literal. This predicate-argument structure is a well-known and familiar structure from logic, linguistics, and cognitive science. Triples resemble the binary predicates in propositional logic needed to express facts, relationships, and the properties of individuals. Furthermore, triples seem similar to simple natural language sentences, where the subject and objects are nouns and the predicate is a verb.

From the perspective of the traditional Web, the main feature of RDF is that links in RDF themselves have a required role URI. It is through this role that URIs are given to relationships outside the Web in RDF. For example, the relationship of ‘is architect of’ between Gustave Eiffel and the Eiffel Tower could be formalized as a link (as shown in Figure 5.2), as could the relationship between Tim Berners-Lee and the creation of his web-page. Since the relationships are abstract, these URIs then refer to these relationships, the URIs are primarily referential and may not lead to access unlike

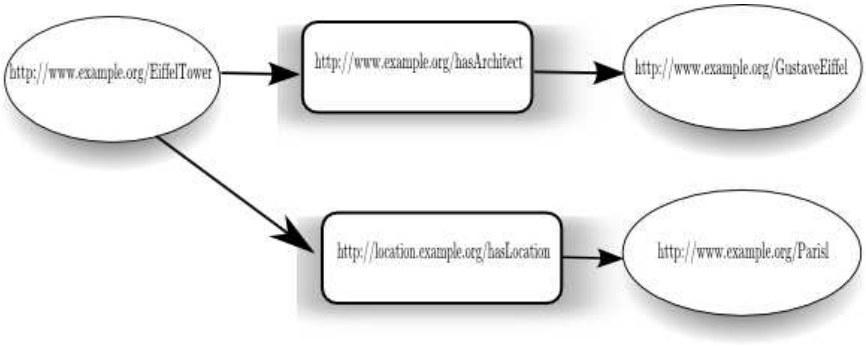

Figure 5.3: Merging RDF triples

links in traditional hypertext systems. A set of RDF triples is essentially a linkbase, such as those pioneered in earlier hypertext systems like Microcosm (Fountain et al., 1990). Similarly, a triple by itself can only state a simple assertion, but webs of links may be made between triples to explain. A set of triples that share resources is called a ***graph***, as illustrated in Figure 5.3 by two triples having the same subject, namely that 'The Eiffel Tower in Paris has an architect called Gustave Eiffel.'

With the ability to make separate statements using URIs, the main purpose of RDF is revealed to be *information integration*. Due to their reliance on URIs, RDF graphs can ***graph merge***, when *two formerly separate graphs combine with each other when they use any of the same URIs*. The central purpose of URIs is to allow independent agents to make statements about the same referent. With a common language of URIs, agents can merge information about the referents of the URIs in a decentralized manner.

5.2.3 RDF and the Principle of Self-Description

Once the Principle of Universality and the Principle of Linking are obeyed, the Principle of Self-Description naturally follows, and RDF is no exception. Self-description is a crucial advantage of RDF in decentralized environments, since an agent by following links can discover the context of a triple needed for its interpretation. As witnessed by the Brachman and Smith survey of knowledge representation systems, a bugbear of semantic networks was their inability to be transferred outside of the closed domain and centralized research group that designed them (Brachman and Smith, 1980). The crucial context for usage of a particular semantic network was always lost in transfer, so that what precisely "IS-A" means could vary immensely between contexts, such as the

difference between a sub-class relationship or individual identity (Brachman, 1983). By providing self-description, RDF triples can be transported from one context to another, at least in an ideal world where normal conditions hold, such as when the URIs in the triple can be used to access a web-page describing its content, and correct media types are used. Furthermore, as RDF is imagined to be used as a basic meta-language for other dialects, these dialects can also have their intended interpretation discovered by the follow-your-nose algorithm.

The hypertext Web, when every resource is linked together, provides a seamless space of linked documents. For example, the W3C tries to deploy its own internal infrastructure in a manner compatible with the principles of Web architecture. Its e-mail lists are archived to the Web, and each e-mail is given a URI, so an agent may follow links seamlessly from one e-mail message to another, and by following links can launch applications to send e-mail, discovers more about the group, and in new e-mails reference previous topics. Likewise, an initiative called 'Linked Data' attempts to deploy massive public data-sets as RDF, and its main tenet is to follow the Principle of Self Description (Bizer et al., 2008). The hope is that the Semantic Web can be thought of as a seamless web of linked data, so that an agent can discover the interpretation of Semantic Web data by just following links. These links will then go to more data which may host formal definitions or informal natural language descriptions and multimedia depictions. For example, if one finds an RDF triple such as `ex:EiffelTower ex:hasArchitect ex:Gustave_Eiffel` and discover more information about the Eiffel Tower, like a picture of it or the fact that construction was finished in 1889 by accessing `http://www.example.org/EiffelTower`. Still, the devil is in the details, especially when trying to decide exactly how to connect a URI for the Eiffel Tower itself and another URI for some digital information about it given in RDF and HTML, as explored in Chapter 6.

Since RDF is supposed to be an all-purpose knowledge representation system for the Web, RDF statements themselves can also be described using RDF. RDF itself has a namespace document at `http://www.w3.org/1999/02/22-rdf-syntax-ns#`, which provides a description of RDF in RDF itself. In other words, RDF can be meta-modeled using RDF itself, in a similar manner to the use of reflection in knowledge representation and programming languages (Smith, 1984). For example, the notion of a RDF predicate is `http://www.w3.org/1999/02/22-rdf-syntax-ns#predicate`, and is defined as "the predicate of the subject RDF statement." The same holds for almost all RDF constructs, and a conformant RDF processor can derive from any

RDF triple a set of axiomatic triples that define RDF itself, such as `rdf:predicate rdf:type rdf:Property` (all RDF predicates are of the type property). For any RDF statement like `ex:EiffelTower ex:hasArchitect ex:Gustave_Eiffel`, an RDF-aware agent can then infer that `ex:hasArchitect rdf:type rdf:predicate`, which states in RDF that an architect relationship is a predicate in a RDF triple. However, usually RDF is not hosted according to the Principle of Self-Description. Use of the media type `application/rdf+xml` is not consistent usually, and the namespaces URI of specifications like the RDF Syntax namespace often allows nothing more than access to some RDF triples, which is useless to a machine incapable of understanding RDF in the first place, instead of accessing a document that contains some useful human-readable information, such as a *Resource Directory Description Language* (RDDL) namespace document (Borden and Bray, 2002). A version of RDDL in RDF exists with an associated automated transform[2] makes it even easier for Semantic Web agents to follow namespace documents to associated resources (Walsh and Thompson, 2007).

5.2.4 RDF and the Open World Principle

The Principle of the Open World is the fundamental principle of inference on the Semantic Web. A relatively simple language for declaring sub-classes and sub-properties, RDF Schema, abbreviated as RDF(S), was from the beginning part of the vision of the Semantic Web and developed simultaneously with RDF. Yet determining how to specify exactly what other triples may be inferred from a given RDF triple is a non-trivial design problem, since it required adding an inference mechanism to a semantic network, which historically in AI featured little or no inference. Those that do not remember the history of artificial intelligence are bound to repeat it, and the process of specifying inference in RDF led to an almost complete repeat of the 'procedural versus declarative' semantics debate. An early W3C Recommendation for RDF defined its inference procedure by natural language and examples (Lassila and Swick, 1999). Yet differing interpretations of this early RDF W3C Recommendation led to decidedly different inference results, and so incompatible RDF processors. This being unacceptable for a Web standards organization, the original RDF W3C Recommendation was deprecated, and rewritten. The original defender of formal semantics in artificial intelligence, Pat Hayes, oversaw the creation of a declarative, formal semantics for RDF

[2] Also called a *Gleaning Resource Descriptions from Dialects of Languages* (GRDDL) transform (Connolly, 2007).

and RDF(S) in order to give them a principled inference mechanism (Hayes, 2004).

The Open World principle was considered to be a consequence of the lack of centralized knowledge implied by the decentralized creation of URIs and links as given by the Principles of Universality and Linking. The parallel to the removal of centralized link indexes is that on the Semantic Web, "we remove the centralized concepts of absolute truth, total knowledge, and total provability, and see what we can do with limited knowledge" (1998c). Hayes argued, in a similar fashion as he had argued in the original 'procedural versus declarative' semantics debate in AI, that the Semantic Web should just use standard first-order predicate logic. Yet while Berners-Lee accepted the need for a logic-based semantics, he argued against Hayes for the Principle of Open World and monotonicity, and the formal semantics of RDF was designed to obey the Open World Assumption (Hayes, 2002). The reason for maintaining the Open World Assumption was that adding triples in a graph merge should never change the meaning of a graph so one could never retract information by simply adding more triples, and so possibly invalidate previously-made conclusions. This monotonicity is considered key, since otherwise every time a RDF triple was merged into a graph the interpretation of the graph could change and so the entire graph might have to be re-interpreted, a potentially computationally expensive operation. By having a design that allows only monotonic reasoning, RDF allows interpretations to be changed incrementally in order to scale well in the potentially unbounded partial information of the Web. Hayes himself eventually came to agree with Berners-Lee on the issue, noting that reasoning on the Semantic Web "needs to always take place in a potentially open-ended situation: there is always the possibility that new information might arise from some other source, so one is never justified in assuming that one has 'all' the facts about some topic" (2002).

RDF Schema is on the surface a very simple modeling and inference language (Brickley and Guha, 2004). Due to the Open World assumption, unlike schemas in relational databases or XML Schemas, RDF Schemas are not prescriptive, but merely descriptive, and so an agent cannot validate RDF triples as being either consistent or inconsistent with an RDF Schema (Thompson et al., 2004). They cannot make the information given by a triple itself change, but only enrich the description of an existing triple. RDF Schema adds two main features to RDF. First, RDF(S) provides a notion of a class, or a set of resources. Then RDF(S) allows any resource to be given membership in classes and declare sub-classes (or subsets) of a class that inherit all the triples created to describe the class. Second, RDF(S) also allows properties to have

sub-properties and for properties to have types for domains and ranges, such that in a triple the subject is the domain and the object is the range of a property. Imagine that the property `ex:hasArchitect` has the range `ex:Person` and domain `ex:Building`. Note that RDF Schemas are not automatically applied to triples even if they are mentioned in a triple, such that for a statement like `ex:Eiffel_Tower ex:hasArchitect ex:Gustave_Eiffel`, the fact that the domain of `ex:hasArchitect` is buildings and the range is people is not known unless the RDF Schema is automatically imported and so merged with the triple itself. If the RDF Schema has been imported (either explicitly via `owl:imports` or the follow-your-nose algorithm), an RDF(S)-aware agent that has retrieved the RDF Schema can deduce from the triple that `ex:Gustave_Eiffel rdf:type ex:Person`, namely that Gustave Eiffel is indeed a person. This sort of simple reasoning is again encoded as a set of axiomatic triples and rules for inference and semantic conditions for applying these axioms to infer more triples. See the RDF Formal Semantics for full details (Hayes, 2004). From here on out, the acronym 'RDF' refers to both RDF and RDF(S), whose formal semantics are given together (Hayes, 2004).

In practice, the Principle of the Open World has surprising results. One of the ramifications in RDF is that there is no proper notion of false, but only the notion that something is either inferred or not, and if it is not inferred, it may simply be undefined. Although it seems straightforward, in practice this leads to surprising results. Take the following example: 'Gustave is the father of Valentine,' which in RDF is `ex:Gustave ex:fatherOf ex:Valentine_Eiffel`. Is George also the father of Valentine, i.e. `ex:George ex:fatherOf ex:Valentine`? Operating under the closed world assumption, the answer would be 'no.' Yet operating under the Open World Principle, that statement would be possible, for there is no restriction that the there someone can only have a single father, and in RDF(S) stating such a restriction is impossible. This restriction is possible in the ***Web Ontology Language*** (abbreviated OWL, in an obscure reference to A.A. Milne), an open-world extension of RDF that allows restrictions, such as cardinality, to be placed on predicates. However, even if one set the cardinality of the `ex:fatherOf` predicate to one (so that one could have at most one father), the results will be surprising: the reasoner will conclude that `ex:George` and `ex:Gustave` refer to the same individual. In contrast to the expected behavior of many other inference engines, including people, there is no ***Unique Name Assumption***, *the assumption that each unique name refers to a unique individual*, due to the Open World Principle. The Unique Name Assumption, while very useful for counting,

makes an implicit assumption about each name referring to only one individual, and if an individual cannot be found that satisfies the name then that individual must not exist. This further reinforces the tendency of URIs on the Semantic Web, despite their global scope, to be ambiguous, a point we shall return to in Chapter 6.

5.2.5 RDF and the Principle of Least Power

Insofar as it is applied to the Semantic Web, the Principle of Least Power is strangely counter-intuitive: traditionally knowledge representation languages were always striving for greater power, yet the Semantic Web begins with RDF, a language purposefully designed to be the least powerful language. The true bet of the Semantic Web is then on triples as the most basic language upon which other languages can be based. The challenge for the Principle of Least Power is how to build the rest of the Semantic Web by expanding on the language of triples.

Inspired by the Principle of Least Power, he envisaged that each language would extend and build upon lower-level languages. On top of RDF, Berners-Lee envisaged a whole stack of more expressive languages being constructed. Although the vagarities of the standardization process have caused various changes in the 'Semantic Web stack' and numerous conflicting versions exist, the original and most popular version of the Semantic Web stack is given in Figure 5.4 (Gerber et al., 2008). The W3C has commenced standardization efforts in a number of these areas, and research in almost all levels of the stack has begun. The majority of the research has focused on extending the Semantic Web with 'ontologies' based on description logic like OWL. As should be suspected given their heritage in artificial intelligence, most of the work in description logic applied to OWL has focused on determining the most expressive possible language that preserves decidable inference. OWL itself works well with the Open World Principle, since it only makes an inference by adding inferred statements and classifications, and so remains monotonic. While almost any possible triple is acceptable in RDF, OWL allows users to design ontologies that can even add constraints, such as cardinality and data-typing, that can make some RDF triples inconsistent with a given OWL ontology. Another part of the Semantic Web, originally unforeseen, is the query language SPARQL, a query language for RDF similar to the popular database query language SQL (Prud'hommeaux and Seaborne, 2008). Current work is focused on Rule Interchange Format (RIF), a rule-language similar to Prolog for both serializing normal rules and operating over RDF data (Boley and Kifer, 2008). Other higher-levels

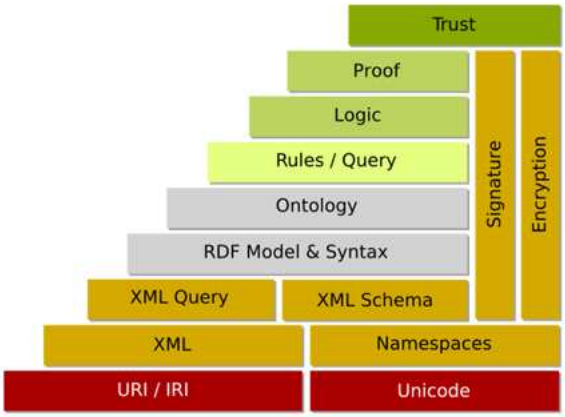

Figure 5.4: The Semantic Web stack

on the Semantic Web stack such as 'Logic' remain mysterious if evocative.

5.3 Information and Non-Information Resources

One question is whether or not there should be some way to distinguish between URIs used to access web-pages and Semantic Web data, and URIs used as names for things like physical entities and abstract concepts that are not 'on the Web.' This latter class of URIs, *URIs that are used as names for entities and abstract concepts*, are called ***Semantic Web URIs***. Should a URI be able to both name a non-Web accessible thing in addition to accessing a representation of the thing? This is a difficult question, as it seems the class of web-pages and physical people should be disjoint (Connolly, 2006). The W3C TAG took on this question, calling it the *httpRange-14* issue, which was phrased as the question: "What is the range of the HTTP dereference function?" (Connolly, 2006).

The TAG defined a class of resources on the Web called an *information resource*, which is a resource "whose essential characteristics can be conveyed in a message" (Jacobs and Walsh, 2004). In particular, this means that an ***information resource*** is a *resource that can be realized as an information-bearing message*. Note that it is not necessarily restricted to a *single* encoding, but possibly can be realized as multiple encodings, just like some fact can be realized by both natural language text in HTML and RDF. A resource is defined by its sense (content), not the encoding of its Web representations. So information resources would naturally include web-pages and so resources

on the hypertext Web, as well as most digital things. However, there are *things that cannot be realized digitally by a message*, but only described or depicted by digital information. These things are ***non-information resources***. Their only realization is themselves. Many analogue things therefore are non-information resources. It appears that this distinction between information resources and non-information resources is trying to get at the heart of the distinction between a resource being a web-page *about* the Eiffel Tower and a resource *for* the Eiffel Tower itself. A web-page is an information resource, but the Eiffel Tower itself is a non-information resource, as is the text of *Moby Dick* or the concept of red.

The distinction is more subtle than it first appears. The question is not whether something *is* accessible on the Web, but whether it *can be* accessible on the Web by being *in theory* transmitted as an encoding, and therefore as a Web representation. For example, imagine a possible world where the Eiffel Tower does not have a web-page. In this world, it would seem counter-intuitive to claim that the web-page of the Eiffel Tower is then not an information resource just because it happens not to *exist* at this moment. This is not as implausible as it sounds, for imagine if the Eiffel Tower's web server went down, so that `http://www.tour-eiffel.fr` returned a `404` status code. A more intuitive case is that of the text of *Moby Dick*. Is the text of *Moby Dick* an information resource? If the complete text of Moby Dick isn't on the Web, one day it might be. However, a particular collector's edition of *Moby Dick* could not be an information resource, since the part of that resource isn't the text, but the physical book itself. Yet do people have to have remarkably scholastic discussions about whether or not something is *essentially* information before creating a Semantic Web URI?

Our previous terminology as defined in Chapter 3 comes to the rescue. Both a web-page about the Eiffel Tower and the text of *Moby Dick* are, on some level of abstraction, carrying information about some sense in some encoding. So, if any information resource is any resource which can have its sense realized as a Web representation, then information resources *must* be on some level of abstraction digital so that they can be encoded as Web representations. Then both the text of *Moby Dick* and a web-page about the Eiffel Tower are information resources, even if they are not currently Web-accessible. Digital information can be transmitted via digital encodings, and so *can* in theory be on the Web by being realized as Web representations, even if the resource does not allow access to Web representations at a given time. Lastly, a particular edition of Moby Dick, or Moby Dick in French, or even some RDF triples about *Moby Dick*, are all information resources, with various encodings specified at certain levels

of abstraction.

It appears that the best story we have to tell about the rather clumsy term 'non-information resource' is that a non-information resource is a thing that is *analogue* and so resists direct digital encoding, but can only be indirectly encoded via representations of the thing in a suitable language. This would then at least be the rather odd combination of physical entities and abstract concepts. So the Eiffel Tower itself, Tim Berners-Lee himself, the integers, and a particular book at a given point in space-time (i.e. on a particular shelf!) are all non-information resources.

Should there be a class to which a web-page about the Eiffel Tower belongs but the text of some as-of-yet unwritten novel does not? In other words, it seems that the class of information resources is too large, and we need a term for things that are actually accessible over the Web at a given time. We call this kind of thing a ***Web resource***, *an information resource that has accessible Web representations that realize its information.* A Web resource can then be thought of as a mapping from time of request to a series of Web representation responses, where the information realized by those Web representations *are* the Web resource. This definition is close in spirit to the original pre-Semantic Web thinking behind resources in IETF 1630, as well as in IETF RFC 2616 where a 'resource' is defined as "a network data object or service " and coherent with Engelbart's original use of the term 'resource' (Engelbart and Ruilifson, 1999; Fielding et al., 1999). A ***Semantic Web resource*** is *a resource that allows access to Semantic Web documents*.

The distinction between information resources and non-information resources has real effects. When the average hacker on the streets wants to add some information to the Semantic Web, the first task is to mint a new URI for the resource at hand, and the second task is to make some of this information about the resource available as a Web representation. However, should a Web representation be accessible from a URI for a non-information resource? If not, should Web representations be accessed from such a non-information resource? This might confuse the non-information resource itself with a Web resource that merely represents that resource. Yet how else would fulfilling the Principle of Self-Description for non-information resources be possible? To refuse to allow access to any Web representations would make the Semantic Web completely separate from the Principles of Web Architecture.

Non-information resources need ***associated descriptions***, *information resources that have as their primary purpose the representation, however incomplete, of some non-information resource*. In other words, associated descriptions are classical exam-

ples of metadata. According to the TAG, since the associated description is a separate thing from the non-information resource it represents, the non-information should be given a separate URI. This would fulfill the common-sense requirement that the URI for a thing itself on the Semantic Web should be *separate* from the URI for some information about the thing. The TAG officially resolved *httpRange-14* by saying that disambiguation between these two types of resource should be done through the `303 See Other` HTTP header. The TAG's official resolution to the *httpRange-14* issue is given below:

- If an HTTP resource responds to a GET request with a two hundred level HTTP response, then the resource identified by that URI is an information resource;
- If an HTTP resource responds to a GET request with a `303 (See Other)` response, then the resource identified by that URI could be any resource;
- If an HTTP resource responds to a GET request with a four hundred level HTTP (error) response, then the nature of the resource is unknown.

To give an example, let's say an agent is trying to access a Semantic Web URI that names a non-information resource, the Eiffel Tower itself, as illustrated in Figure 5.5. Upon attempting to access that resource with a HTTP GET request using its Semantic Web URI, since the Eiffel Tower itself is not an information resource, no Web representations are directly available. The Semantic Web URI used to refer to the Eiffel Tower itself, `http://www.example.org/EiffelTower`, could be any kind of resource, and so could be a non-information resource. Instead, the agent gets a `303 See Other` that in turn redirects them to an associated description that hosts Web representations about the Eiffel Tower, such as the information resource for the homepage of the Eiffel Tower. In turn, using content negotiation, an agent could ask for either the `text/html` or `application/rdf+xml` media type and therefore get redirected to either a URI for hypertext web-page or a Semantic Web document depending on what kind of associated description is needed. When this URI returns the 200 status code in response to an HTTP GET request, the agent can infer that the homepage is actually an information resource. This `303` redirection then allows the non-information resource given by a Semantic Web URI for the Eiffel Tower itself to comply with the Principle of Self-Description.

An alternative to the obtuse `303` redirection is the *hash convention*, where one uses the fragment identifier of a URI to get redirection for free. If one wanted a Semantic

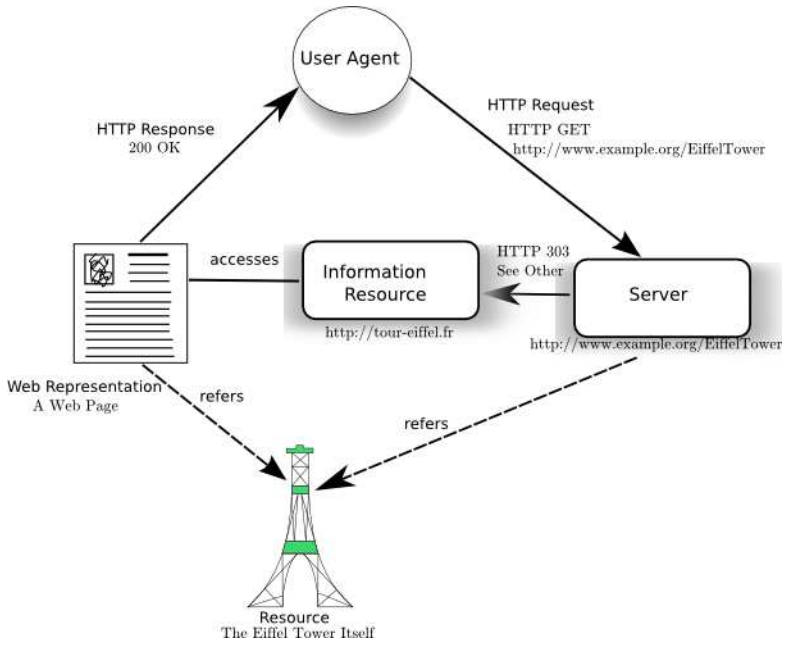

Figure 5.5: The 303 redirection for URIs

Web URI that referred to a non-information resource like the Eiffel Tower itself without the hassle of a 303 redirection, one would use `http://www.tour-eiffel.fr/#` to refer to the Eiffel Tower itself. Since browsers, following the follow-your-nose algorithm, either dispose of it or treat the fragment identifier as a fragment of a document or some other Web representation, if an agent tries to access via HTTP GET a Semantic Web URI that uses the hash convention, the server will not return a `404 Not Found` status code, but instead resolve to the URI before the hash, `http://www.tour-eiffel`, which can then be treated as an associated description. In this way, Semantic Web inference engines can keep the Semantic Web URI that refers to the Eiffel Tower itself and an associated description about the Eiffel Tower separate by taking advantage of some predefined behavior in web browsers.

While at first these distinctions between non-information resources and information resources seems ludicrously fine-grained, clarifying them and pronouncing an official W3C policy on them had an immense impact on the Semantic Web, since once there were definite guidelines on how to publish information on the Semantic Web, users could start creating Semantic Web URIs and connecting them to relevant documentation resources. The TAG's decision on redirection was made part of a tutorial for publishing Semantic Web information called *How to Publish Linked Data on the Web* (Bizer et al., 2007).

5.4 The Semantic Web: Good Old Fashioned AI Redux?

To many, it has seemed that the Semantic Web was nothing but a second coming of classical artificial intelligence. As put by Yorick Wilks, “Some have taken the initial presentation of the Semantic Web by Berners-Lee, Hendler and Lassila to be a restatement of the Good Old Fashioned AI agenda in new and fashionable World Wide Web terms” (2008a). So why would the Semantic Web succeed where classical knowledge representations failed? The first reason would be a difference in the underlying intellectual project. A second reason would be a difference in technology.

The difference of the project is one both of scope and goal. The Semantic Web is, at first glance at least, a more modest project than artificial intelligence. To review the claims of artificial intelligence in order to clarify their relation to the Semantic Web, we are best served by remembering the goal of AI as stated by John McCarthy at the 1956 Dartmouth Conference, “The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” (McCarthy et al., 1955). However, ‘intelligence’ itself is not even vaguely defined. The proposal put forward by McCarthy gave a central role to “common-sense,” so that “a program has common sense if it automatically deduces for itself a sufficient wide class of immediate consequences of anything it is told and what it already knows” (1959).

The Semantic Web does not seek to create a theory of intelligence and encode all common-sense knowledge in some universal representational scheme. The Semantic Web instead leaves “aside the artificial intelligence problem of training machines to behave like people” but instead tries to develop a representation language that can *complement* human intelligence, for “the Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help” (Berners-Lee, 1998c). Many of the most difficult problems of artificial intelligence, as laid out by McCarthy and Minsky, arise because they are interested in a theory of intelligence in general, be it human or machine, and so have to explain difficult problems ranging from natural language understanding to vision (McCarthy et al., 1955). Berners-Lee is explicit that the project of encoding intelligence in general is not part of the problem, as the Semantic Web “does not imply some magical artificial intelligence which allows machines to comprehend human mumblings. It only indicates a machine’s ability to solve a

well-defined problem by performing well-defined operations on existing well-defined data" (Berners-Lee, 1998c). The goal of the Semantic Web is not to provide a theory of intelligence, but instead to enable new – if still untheorized – forms of collective intelligence. As phrased by Licklider, this would be a "man-machine symbiosis," in which in "the anticipated symbiotic partnership, men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions" (1960). So a theory of collective intelligence may still rely on a theory of general intelligence as promised by artificial intelligence, but the Semantic Web itself will not provide such a theory.

While the goals of the Semantic Web are different, it does still employ the same fundamental technology as classical artificial intelligence: knowledge representation languages. As put by Berners-Lee, "The Semantic Web is what we will get if we perform the same globalization process to knowledge representation that the Web initially did to hypertext" (Berners-Lee, 1998c). Yet there is a question about whether or not knowledge representation *itself* might be the problem, not just scale. As put by Karen Spärck Jones, one of the founders of information retrieval, "there are serious problems about the core [Semantic Web] idea of combining substantive formal description with world-wide reach, i.e. having your cake and eating it, even if the cake is only envisaged as more like a modest sponge cake than the rich fruit cake that AI would like to have" (2004). According to Spärck Jones, the problem may lie at the heart of the Semantic Web in its very use of *knowledge representation language* itself. So far we have shown that the properties of at least RDF as a knowledge representation language puts the emphasis on 'Web' as opposed to 'Semantic' in the Semantic Web, as it has a number of properties – a graph structure, the ability to make unconstrained statements, and the like – that have their basis in the tradition of the Web, rather than knowledge representation in AI. As the Web has proved to be extraordinarily successful, the hope of the Semantic Web is that any knowledge representation language which is based on the same principles as the Web may fare better than its ancestors in artificial intelligence. However, these changes in the formalism of RDF due to the influence of the Web are all relatively minor, and while counter-intuitive to traditional knowledge representation, these changes to the formalism based on the principles of Web architecture have yet to be vindicated as the Semantic Web has not yet reached widespread use.

Overlooked by Spärck Jones in her critique of the Semantic Web, the only substantive difference between traditional knowledge representation and the Semantic Web is

the central role of URIs. Just as the later principles of Web architecture build upon the Principle of Universality, so the Semantic Web builds on top of the use of URIs as well. The true bet of the Semantic Web is *not* a bet on the return of knowledge representation languages, but a bet on the universality of URIs, namely that agents in a decentralized and global manner can use URIs to share meaning even about non-Web accessible things using URIs. As this use of URIs as the basic element of meaning is central to the Semantic Web, and as it is a genuinely *new* technical claim, it is precisely in the understanding of the status of meaning and reference of URIs that any new *theoretical* claim must be made. Furthermore, it is precisely within the realm of URIs that any *technical* claim to advance must be made.

Chapter 6

The Identity Crisis

Meaning is what essence becomes when it is divorced from the object of reference and wedded to the word. **W.V.O. Quine** (1951).

6.1 What Do URIs Refer to?

For multiple agents to exchange knowledge representations on the Semantic Web, they must share the meaning of a URI. How can agents determine what a URI refers to? The question lies at the heart of Web architecture itself, although it only becomes noticeable on the Semantic Web. On the hypertext Web URIs trivially identify the hypertext webpages that those URI allow access to, although content negotiation does complicate even that simple story. While on the hypertext Web this question could be ignored as an obscure edge-case, for the Semantic Web this question is absolutely central, since the information identified by Semantic Web URIs should be shared universally in a decentralized manner. In a nutshell, the problem is that URIs identify not only hypertext documents and other digital information, but analogue things that have no causal connection to the Web. How can a Semantic Web URI for the Eiffel Tower be used to refer to the Eiffel Tower in Paris itself? Should the Eiffel Tower itself have a URI? If so, should that URI allow access to any Web representations? This cluster of questions has been dubbed the *Identity Crisis* of the Semantic Web.

As regards any theory of meaning for URIs, in the realm of official Web standards, the jury is still out. In the specification of RDF, Hayes notes that "exactly what is considered to be the 'meaning' of an assertion in RDF or RDF(S) in some broad sense may depend on many factors, including social conventions, comments in natural language"

so unfortunately "much of this meaning will be inaccessible to machine processing" such that a "a full analysis of meaning" is "a large research topic" (Hayes, 2004). As the entire Semantic Web is built on top of the notion of URIs having some sort of sharable 'meaning' or 'referent,' there is no choice but to engage questions of meaning and reference. However, upon pursuing this question, one surprisingly finds there is no clear answer, but instead a conceptual quagmire dominated by two positions.

The first position, the ***direct reference position***, *is that the meaning of a URI is whatever was intended by the owner*. The owner of the URI should be able to unambiguously declare and communicate the meaning of any URI, including a Semantic Web URI. In this position, the referent is generally considered to be some individual unambiguous *single* thing, like the Eiffel Tower or *the* concept of unicorns. This viewpoint is the one generally held by many Web architects, like Berners-Lee, who imagine it holds not just for the Semantic Web, but the entire Web.

The second position, the ***logicist position***, *is that for the Semantic Web, the meaning of a URI is given by whatever things satisfy the model(s) given by the formal semantics of the Semantic Web*. Adherents of this position hold that the referent of a URI is ambiguous, as many different things can satisfy whatever model is given by the interpretation of some sets of sentences using the URI. There are a few minor variations on this theme, with some people believing a URI has no meaning in itself, but only in the context of its use in other triples, while others hold that one should be able to access logical descriptions from the URI itself. This position is generally held by logicians, who claim that the Semantic Web is entirely distinct from the hypertext Web.

These two antagonistic positions were subterranean in the development of the Semantic Web, until a critical point was reached in an argument between Pat Hayes, the AI researcher primarily responsible for the formal semantics of the Semantic Web, and Berners-Lee. This argument was provoked by an issue called 'Social Meaning and RDF' and was brought about by the following draft statement in the *RDF Concepts and Abstract Syntax Recommendation*, "the meaning of an RDF document includes the social meaning, the formal meaning, and the social meaning of the formal entailments" so that "when an RDF graph is asserted in the Web, its publisher is saying something about their view of the world" and "such an assertion should be understood to carry the same social import and responsibilities as an assertion in any other format" (2004). During the period of comments for the RDF Working Drafts, Bijan Parsia commented that the above-mentioned sentences do not "really specify anything and thus can be ignored" or are "dangerously underthought and underspecified" and so should

be removed (Parsia, 2003). While at first these sentences about the meaning of RDF seemed to be rather harmless and in concordance with common-sense, the repercussions on the actual implementation of the Semantic Web are surprisingly large, since "an RDF graph may contain 'defining information' that is opaque to logical reasoners. This information may be used by human interpreters of RDF information, or programmers writing software to perform specialized forms of deduction in the Semantic Web" (Klyne and Carroll, 2004). In other words, a special type of *non-logical* reasoning can therefore be used by the Semantic Web.

An example of this extra-logical reasoning engendered by the fact that URIs identify 'one thing' is as follows. Assume that a human agent has found a URI for the Eiffel Tower from DBpedia, and so by accessing the URI a Semantic Web agent can discover a number of facts about the Eiffel Tower, such that it is in Paris and that its architect is Gustave Eiffel, and these statements are accessed as an RDF graph (Auer et al., 2007). However, a human can have considerable background knowledge about the Eiffel Tower, such as a vague belief that at some point in time it was the tallest building in the world. This information is confirmed by the human agent employing the follow-your-nose algorithm, where by following the subject of any triple, the human would be redirected to the hypertext Wikipedia article about the Eiffel Tower, where the agent discovers via a human-readable description that the Eiffel Tower was in fact the tallest building until 1930, when it was superseded in height by New York City's Chrysler building. This information is *not* explicitly in the RDF graphs provided. It is difficult to even phrase this sort of temporal information in RDF. Furthermore, the human agent discovers another URI for the Eiffel Tower, a RDF version of Wordnet in the file `synset-Eiffel_Tower-noun-1.rdf` (van Assem et al., 2006). When the human agent accesses this URI, there is little information in the RDF graph except that this URI is used for a noun. However, the human-readable `gloss` property explains that the referent of this URI is 'a wrought iron tower 300 metres high that was constructed in Paris in 1889; for many years it was the tallest man-made structure.' Therefore, the human agent believes that there is indeed a singular entity called the 'Eiffel Tower' in Paris, and that this entity was in fact at some point the tallest building in the world, and so the two URIs are equivalent in some sense, although the URIs do not formally match. What the 'Social Meaning' clause was trying to state is that the human should be able to *non-logically* infer that both URIs refer to the Eiffel Tower in Paris, and they use this information to merge the RDF graphs, resulting in perhaps some improved inferences in the future.

This use-case was put forward primarily by Berners-Lee, and the W3C RDF Working Group decided that deciding on the relationship between the social and formal meaning of RDF was beyond the scope of the RDF Working Group to decide, so the RDF Working Group appealed to the W3C TAG for a decision. As TAG member Connolly noticed, they "didn't see a way to specify how this works for RDF without specifying how it works for the rest of the Web at the same time" (Berners-Lee, 2003b). In particular, Berners-Lee then put forward his own viewpoint that "a single meaning is given to each URI," which is summarized by the slogan that a URI "identifies one thing." (2003c).

In response, Hayes said that "it is simply untenable to claim that all names identify one thing" (2003a). Furthermore, he goes on to state that this is one of the basic results of the knowledge representation community and 20th century linguistic semantics, and so that the W3C cannot by fiat render the judgment that a URI identifies one thing. Berners-Lee rejects Hayes's claim that the Semantic Web must somehow build upon the results of logic and natural language, instead claiming that "this system is different from natural language: we designed it such that each URI identifies one and only one concrete thing in the real world or one and only one globally shared concept" (2003a). In exasperation, Hayes retorted that "I'm not saying that the 'unique identification' condition is an unattainable ideal: I'm saying that it doesn't make sense, that it isn't true, and that it could not possibly be true. I'm saying that it is *crazy*" (2003b). While Hayes did not explain his own position fully, as he was the editor of the formal semantics of RDF and had the support of other logicians in the RDF Working Group, the issue deadlocked and the RDF Working Group was unable to come to a consensus. In order to move RDF from a Working Draft to a Recommendation, the W3C RDF Working Group removed all references to social meaning from the RDF documents.

One should be worried when two prominent researchers such as Berners-Lee and Hayes have such a titanic disagreement, where no sort of consensus agreement seems forthcoming. Yet who is right? Berners-Lee's viewpoint seems intuitive and easy to understand, and some people would say that it qualifies as common-sense. However, the argument would seem to have been won by Hayes, as many people would also agree that his defense of ambiguity in names is also common-sense, and Hayes also has the backing of his knowledge of the formal semantics of logic. Still, there is reason to pause to consider the possibility that Berners-Lee is correct. First, while Berners-Lee's notion of unambiguous names may seem counter to many of our intuitions about the common-sense knowledge that many names are indeed of ambiguous, Berners-Lee

can claim that his viewpoint also is shared with philosophers and logicians such as Kripke, as explored in Section 6.3. Furthermore, while Hayes may appear to have a common-sense understanding of ambiguity, as explored in Section 6.2, Hayes actually is arguing for the much more radical claim that in the case of the Semantic Web only inference as defined by a formal logic can restrict interpretations, and hence ambiguity. In this vein, it should should be remembered that as far as practical results are concerned, the project of logic-based modeling of common-sense knowledge in classical artificial intelligence earlier inaugurated by Hayes is commonly viewed to be a failure by current researchers in AI and cognitive science (Wheeler, 2005). In contrast, despite the eerily similar argument that Berners-Lee had with original hypertext academic researchers about broken links and with the IETF about the impossibility of a single naming scheme for the entire Internet, the Web is without a doubt an unparalleled success. While in general the intuitions of Berners-Lee may seem to be wrong according to academia, history has proven him right in the past. Therefore, one should take his pronouncements seriously.

The Identity Crisis is not just a conflict between merely two differing individual opinions, but a conflict between two entire disciplines: the nascent discipline of 'Web Science' as given by the principles of Web architecture, and that of knowledge representation in AI and logic (Berners-Lee et al., 2006b). Berners-Lee's background is in the Internet standardization bodies like the IETF, and it is primarily his intuitions behind Web architecture as given in Chapter 5. As discussed in Chapter 2, Hayes is a formidable character in the field of artificial intelligence, since it was his background in logic that jump-started the field of knowledge representation. If two entire fields, who have joined common-cause in the Semantic Web, are at odds, then trouble at the level of *theory* is afoot.

Troubles at levels of theory invariably cause trouble in practice. So this disagreement would not be nearly as worrisome were not the Semantic Web itself not in such a state of perpetual disrepair, making it practically unusable. In a manner disturbingly similar to classical artificial intelligence, the Semantic Web is always thought of as soon-to-be arriving, the 'next' big thing, but its actual uses are few and far between. The reason given by Semantic Web advocates is that the Semantic Web is suffering from simple engineering problems, such as a lack of some new standard, some easily-accessible list of vocabularies, or a dearth of Semantic Web-enabled programs. The fact that the Semantic Web has not yet experienced the dizzying growth of the original hypertext Web, even after an even longer period of gestation, points to the fact that

something is fundamentally awry. The root of the problem is the dependence of the Semantic Web on using URIs as names for referents.

Far from being a mandarin metaphysical pursuit, this problem is the very first practical issue one encounters as soon as one wants to actually use the Semantic Web. If an agent receives a graph in RDF, then the agent should be able to determine an interpretation of these triples. The inference procedure itself may help this problem, but it may instead make it worse, simply producing more uninterpretable RDF statements. The agent could employ the follow-your-nose algorithm, but what information, if any, should be accessible at these Semantic Web-enabled URIs? If a user wants to add some information to the Semantic Web, how many URIs should they create? One for the representation, and another for the referent the representation is *about*? In other words, one for the associated description and another one for the non-information resource the associated description is about? Should the same URI for the Eiffel Tower itself be the one that is used to access a web-page about the Eiffel Tower?

What is then necessary to explain these vast differences over such a basic issue would be a more complete explanation of the differing background assumptions between Berner-Lee's direct reference position and Hayes's logicist position. URIs on the Semantic Web can be thought of as analogous to natural language *names*, as names in natural language can be used to refer as well. Therefore, what needs to be done is to distinguish within analytic philosophy the various theories on naming and reference in general, and then see how these various theories either do or do not apply to the Semantic Web. What is remarkable is that the position of Hayes, the logicist position, corresponds to a well-known theory of meaning and reference, the 'descriptivist theory of reference' attributed to early Wittgenstein, Carnap, Russell, and turned into its pure logicist form by Tarski (Luntley, 1999). However, it is common currency in philosophical circles that the descriptivist theory of reference was overthrown by the 'causal theory of reference' championed by Kripke and extended by Putnam (Luntley, 1999). It is precisely this causal theory of reference that Berners-Lee justifies in his direct reference position. Thus, the curious coincidence is that both opposing positions on the Semantic Web correspond to *equally* opposing positions in philosophy. Understanding these positions belongs primarily to the domain of philosophy, even if Hayes and especially Berners-Lee do not articulate their positions with the relevant academic citations. In this manner, the precise domain of philosophy that the Identity Crisis falls under is the philosophy of language. The purpose of the rest of this chapter is then the full explication of these two theories of reference in philosophy of language, and

then to inspect their practical success (or lack thereof) in the context of the Semantic Web, while at the end offering a critique of both, paving the way for a third theory of meaning.

6.2 The Logicist Position and the Descriptivist Theory of Reference

The origin of the logicist position is the descriptivist theory of reference. In the ***descriptivist theory of reference***, *the referent of a name is given by whatever satisfies the descriptions associated with the name*. Usually, the descriptions are thought to be logical, so a name is actually a disguised logical description. The referent of the name is then equivalent to the set of possible things, given normally by a mathematical model, such that all statements containing the name are satisfied.

6.2.1 Logical Atomism

The roots of the descriptivist theory of reference lay with the confluence of philosophers who are known as *logical atomists*, a term coined by Bertrand Russell, and influential to later epistemological projects like the *logical positivism* of Rudolf Carnap. Although eventually abandoned by Bertrand Russell, logical atomism is a vast school of thought that has proven tremendously influential, even in its current discredited state, for our purposes we will only concern ourselves with one particular doctrine: The problem of how natural language terms relate to the logical descriptions, and logical descriptions to the world. Bertrand Russell begins the investigation of the connection between logic and language is his landmark investigation *On Denoting* with a deceptively simple question: “is the King of France bald?” (1905). To what referent does the description “the King of France” refer to? (Russell, 1905) Since in Russell’s time there was no King of France, it could not refer to anything like what Carnap later called “elementary sense data” (Carnap, 1928). In this regard, Russell makes a crucial distinction. According to Russell, elementary sensory experiences are known through *acquaintance*, in which we have some sort of direct ‘presentation of’ the thing (1905). Yet knowledge of a thing can be based on ***description***, which are those “things we only reach by means of denoting phrases” (Russell, 1905). Russell believed that “all thinking has to start from acquaintance, but it succeeds in thinking *about* many things with which we have no acquaintance” via the use of description (1905). Russell was most

interested in whether those things with which we have direct acquaintance can be considered true or false, or whether a more mysterious third category such as 'nonsense' is needed. Russell opts to reject creating imaginary but true 'things' as well as any third category, but instead holds that statements such as "the King of France is bald" are false, since "it is false that there is an entity which is now the King of France and is bald" (Russell, 1905). This solution then raises the alarming possibility that "the King of France is not bald" may also come out false, which would seem to violate the Law of the Excluded Middle. So, Russell counters this move by introducing the fact that "the King of France is bald" is actually a complex logical statement involving scope and quantification, namely $(\exists x.F(x) \wedge G(x)) \wedge (\forall y.F(y) \rightarrow x = y)$, where F is "being the King of France" and G is "being bald" (Russell, 1905). According to the analysis, 'The King of France' is merely a *disguised* complex logical statement. Furthermore, this treatment can be extended to proper names such as 'Sir Walter Scott,' who can be identified with 'the author of Waverly,' so that instead of being a tautology, even a proper name of a person, even if known through acquaintance, is sort of short-hand for a large cluster of logical statements. So to use our previous example, the 'Eiffel Tower' can be thought of as a short-hand for not only that 'there exists an entity known as the Eiffel Tower' but also the logical statement was 'the aforementioned entity had Gustave Eiffel as its architect.' If someone did not know that 'the aforementioned entity was also the tallest building in the world up until 1930,' one could then make a statement such as 'The Eiffel Tower is identical to the tallest building in the world up until 1930' without merely stating a tautology, and such a statement would add true and consistent knowledge to a hearer who was not aware of the statement.

While the first proponent of logical atomism was Bertrand Russell, one of its most systematic presentations is in his student Ludwig Wittgenstein's early philosophical work the *Tractatus Logico-Philosophicus*. In it, Wittgenstein strongly argues for his own version of *logical atomism*, that *logic* is the true language of the world; "logic is not a body of doctrine, but a mirror image of the world" for "the facts in logical space" are the world (1921). So logical statements are "laid against reality like a measure" (1921). This is possible because the world is metaphysically determinate at its base, being composed of "simple" and "unalterable" objects that "make up the substance of the world" so that "the configuration of objects produces states of affairs" where "the totality of existing states of affairs is the world" (Wittgenstein, 1921). In other words, there is no – as Brian Cantwell Smith would put it – "flex" or "slop" in this picture, no underlying "metaphysical flux" that somehow resists easily being constrained into

these fully determinate "objects" (1995). Although the nature of the world consists of *true* logical facts, humans, since they "picture facts" to themselves, can nonetheless make *false* logical statements, since these pictures merely "model reality" (Wittgenstein, 1921). Contrary to his own logical atomist mentor Russell, Wittgenstein thought that the primary job of the logician is then to state true facts, and "what we cannot speak about" in the form of true logical statements "we must pass over in silence," a phrase he believed was consistently misinterpreted by even his teacher Bertrand Russell and later philosophers like Carnap (Wittgenstein, 1921). Note that unlike the more mature standpoint of Hayes, the logical atomism of Wittgenstein allowed *logical statements* to directly refer to single things in the world, Wittgenstein and other logical atomists reified *the logical model* to *be the world* itself.

This position was further developed by Rudolf Carnap. According to Carnap, in his *The Logical Structure of the World*, all statements (at least, "scientific" statements with "cognitive content" about the world) can then be reduced to logical statements, where the content of this logical language is given by sensory experiences (1928). These "elementary experiences" cannot be directly described, as they are irreducible, but only described by a network of logical predicates that treat these experiences as logical constants (Carnap, 1928). While Carnap's ultimate goal was to render any scientific hypothesis either verifiable by sense experience or not; their general position was since natural language is part of the world, the structure of language too must be logical, and range over these elementary sense experiences. In this regard, names are given to their referents by concordance with a logical structure ranging over these elementary sensory experiences. Carnap's project was similar in spirit to Chomsky's syntactic theory of language, but focused on semantics rather than syntax: Carnap hoped to develop a semantic and logical definition of meaning that would validate only sentences with 'meaning.'

As sensible as logical atomism appeared, there are difficulties in building any theory of reference on, as Quine put it, such a "slender basis" as elementary sense data and logic (1951). The crux of the problem for any descriptivist theory of names is that names for any "kind of abstract entities like properties, classes, relations, numbers, propositions" could not have an interpretation to any content using such a simple sensory epistemology (Carnap, 1950). Carnap's *Empiricism, Semantics, and Ontology* made an argument for basing such entities purely on linguistic form itself. Carnap believed that, despite the difficulty of determining the interpretation of names for abstract entities, "such a language does not imply embracing a Platonic ontology but is

perfectly compatible with empiricism" (1950). His position was that while "if someone wishes to speak in his language about a new kind of entity, he has to introduce a system of new ways of speaking, subject to new rules," which Carnap calls the "construction of a linguistic framework for the new entities in question." From *within* a linguistic framework, Carnap believed to commit to any statement about the "existence or reality of the total system of the new entities" was to make a "pseudo-statement without cognitive content" (1950). Although this particular position of Carnap's was devastated by Quine's argument against analyticity in *The Two Dogmas of Empiricism*, Carnap made an important advance in the idea of a name of even abstract things being defined by linguistic descriptions, the problems brought up by Quine forced later logicians to abandon the notion of the logic ranging over "elementary sense data" (Quine, 1951).

6.2.2 Tarski's Formal Semantics

Tarski abandoned the quaint epistemology of Russell and Carnap and defined reference purely in terms of logic in his *The Concept of Truth in Formal Languages* (Tarski, 1935). Reference was just defined as a consequence of the truth *only* in terms of satisfaction of a formal language (1935). To set up his exposition, Tarksi defines two languages, the first being the syntactic *object language L* and the second being the *meta-language M*. The *meta-language* should be *more expressive* (in the sense given in Section 5.2.5) such that it can describe every sentence in the object language, and furthermore, that it contain axioms that allow the truth of every sentence in the object language to be defined. In his first move, Tarski defines *the formal conception of truth* as 'Convention T,' namely that for a given sentence *s* in *L*, there is a statement *p* in *M* that is a theorem defining the truth of *s*, that is, the truth of *s* is determined via a translation of *s* into *M* (Tarski, 1935). Tarski then later shows that truth can be formally defined as "*s* is true if and only if *p*" (Tarski, 1944). For example, if the object language is exemplified by a sentence uttered by some speaker of English and the meta-language was an English description of the real world; 'The Eiffel Tower is in Paris' is true if and only if the Eiffel Tower is in Paris. The sentence 'The Eiffel Tower is in Paris' must be satisfied by the Eiffel Tower *actually being* in Paris. While this would at first seem circular, its non-circularity is better seen through when the object language is not English, but another language such as German. In this case, "'Der Eiffelturm ist in Paris' is true if and only if the Eiffel Tower is in Paris." However, Tarksi was not interested in informal languages such as English, but in determining the

meaning of a new formal language via translations to mathematical models or other formal languages with well-known models. If one was defining a formal semantics for some fragment of a knowledge representation language like RDF, a statement such as `http://www.eiffeltower.example.org ex:location dbpedia:Paris` is true if and only if $\exists ab.R(a,b)$ where R, a, and b are given in first-order predicate logic.

This straightforward approach to formal semantics runs into a difficulty, as shown in the above example; if one is defining a formal Tarski-style semantics for a language, what should one do when one encounters complex statements, such as 'the Eiffel Tower is in Paris and had as an architect Gustave Eiffel.' The answer is at the heart of Tarksi's project, namely that the second component of Tarski's formal semantics is to use the principle of compositionality so that any complex sentence can have its truth conditions derived from the truth conditions of its constituents. To do this, the meta-language has to have finitely many axioms, and each of the truth-defining theorems produced by the meta-language have to be generated from the axioms (Tarski, 1935). So, the aforementioned complex sentence is true if and only if $\exists ab.R(a,b) \wedge Q(a,c)$, where Q can be the *architect of* relationship, c can be Gustave Eiffel and a the Eiffel Tower. Tarksi's theory as explained so far only deals with 'closed' sentences, i.e. sentences containing no variables or quantification. The third, and final component of Tarski's formal semantics is to use the notion of satisfaction via *extension* to define truth (Tarski, 1935). For a sentence such as 'all monuments have a location,' we can translate the sentence to $\forall a,l.monument(a) \rightarrow hasLocation(a,l)$ which is true if and only if there is an extension x from the world that satisfies the logical statements made about a. In particular, Tarksi has as his preferred extensions infinite ordered pairs, where the ordered set could be anything (Tarski, 1935). For formal languages, as explained in Section 3.3, a model-theoretic semantics with a model composed by set theory was standard. For example, the ordered pairs in some model of $(EiffelTower, Paris)$ would satisfy our example statement, as would $(ScottMonument, Edinburgh)$ but not $(Paris, EiffelTower)$. However, there is no reason why these models could not be "God Forthcoming," things in the the real world itself, albeit given in set-theoretic terms that would violate the "metaphysical flux" of the world (Smith, 1995). Henceforth we will assume all extensions used by Tarski-style semantics are models. To summarize Tarksi's remarkably successful programme, model-theoretic semantics can produce a theory of truth that defines the semantics of a sentence in terms of the use of a translation of the sentence into some formal language with a finite number of axioms, then using compositionality to define the truth of complex sentences in terms of basic sentences, and finally

determining the truth of those basic sentences in terms of what things in a model satisfy the extensions of the basic sentences as given by the axioms. This work marks the high-point of the logicist programme, as all questions of meaning are reduced to questions about giving the interpretation of a sentence in terms of a formal notion of truth, and this notion of truth is not restricted by the logical atomist's quaint epistemology of elementary sense data, but instead can range over any possible formal language and any possible worlds.

6.2.3 In Defense of Ambiguity

The descriptivist theory of reference, taken to its conclusion, results in the logicist position on the Semantic Web. While this work in the descriptivist theory of reference seems distant from the Identity Crisis of the Web, it is in fact central to the position of Hayes and the Semantic Web as a whole. This is primarily because Hayes's background was in the logicist tradition, with his particular specialty being the creation of Tarski-style semantics for knowledge representation languages. What Hayes calls the "basic results in 20th century linguistic semantics" that Berners-Lee's dictum that "URIs identify one thing" violates is the interpretation of URIs in a Tarski-style formal semantics (Hayes, 2003a). For the logicist position, the *semantics* in the Semantic Web derive from the Tarski-style formal semantics Hayes created for the Semantic Web (2004).

Before delving into the RDF Formal Semantics, it should be noticed that these semantics are done by extension, including not only subjects and objects but properties, which is unusual in light of standard formal semantics given by Hayes for first-order logic in KIF (2001). The reason for this is the Principle of Linking, in particular, the unusual features of RDF that "a property may be applied to itself" and that classes "may contain themselves" (Hayes, 2004). This is done by distinguishing the class *qua* class and property *qua* property in RDF from whatever their extensions are, so while a class and property in RDF may or may not be satisfied by some model or world, the extension of the class or property are not considered to have the same *identity* as the property or class.

A simple example should suffice. What is the formal semantics of `ex:EiffelTower ex:architect ex:Gustave_Eiffel`? To simplify slightly, Hayes defines the formal semantics of set theory, where there is a set of resources that compose the model of the language, a set of properties, and a set of URIs that can refer to resources. The

interpretation of any RDF statement is then given as an extensional mapping from the set of properties to the powerset of resources, to the set of pairs of resources. So, given a set-theoretic model consisting of elements (given by italics) *Gustave Eiffel* and *the Eiffel Tower* and *being the architect of*, then `ex:EiffelTower` $\models$ *the Eiffel Tower*, `ex:Gustave_Eiffel` $\models$ *Gustave Eiffel* and `ex:architect` $\models$ *being the architect of*, so that the entire triple maps to a set of pairs: `ex:EiffelTower ex:architect ex:Gustave_Eiffel`$\models$ *(..., (the Eiffel Tower, Gustave Eiffel), ...)*. Someone using common-sense human intuitions will likely believe that this interpretation maps to our common-sense content of `ex:EiffelTower ex:architect ex:Gustave_EiffelTower`, and using the axiomatic triples defined in the RDF formal semantics, a few new triples can be inferred, such as `ex:architect rdf:type rdf:Property`.

However, the inherent pluralism of the Tarski approach to models also means that another equally valid interpretation would be the inverse, i.e. the mapping of `ex:EiffelTower` to *Gustave Eiffel* and `ex:Gustave_Eiffel` to *the Eiffel Tower*. In other words, `ex:architect` $\models$ *being the architect of*, so that the entire triple maps to a set of pairs `ex:EiffelTower ex:architect ex:Gustave_Eiffel`$\models$ *..., (Gustave Eiffel, Eiffel Tower), ...)*. Due to the unconstrained nature of RDF, `ex:architect` has no 'natural' relationship to anything in particular, but could easily be assigned either *the Eiffel Tower* or *Gustave Eiffel* just as easily as *being the architect of*.

Furthermore, the model could just as easily be given by something as abstract as the integers *1* and *2*, and an equally valid mapping would be for `ex:EiffelTower` $\models$ *1* and `ex:Gustave_Eiffel` $\models$ *2*, so that `ex:architect` $\models$ *being the architect of*, so that the entire triple maps to a set of pairs `ex:EiffelTower ex:architect ex:Gustave_Eiffel`$\models$ *(..., (1,2), ...)*. Indeed, the extreme pluralism of a Tarski-style semantics shows that, at least if all one has is a single lone triple statement, that triple can be satisfied by any model. This is no mere oddity of formal languages, this would also hold for any lone sentence in a language like English – such as "Gustave Eiffel is the architect of the Eiffel Tower" – as long as one subscribed to a Tarski-style semantics for natural language, such as Montague semantics (Montague, 1970). As the number of triples increased, the amount of possible things that satisfy the model is thought to decrease, but in such a loose language as RDF, as mandated by the Principle of Linking, Hayes notes that it is "usually impossible to assert enough in any language to completely constrain the interpretations to a single possible world, so there is no such thing as 'the' unique interpretation" (2004). This descriptivist theory of reference, where descriptions are logical statements in RDF, is illustrated in Figure 6.1.

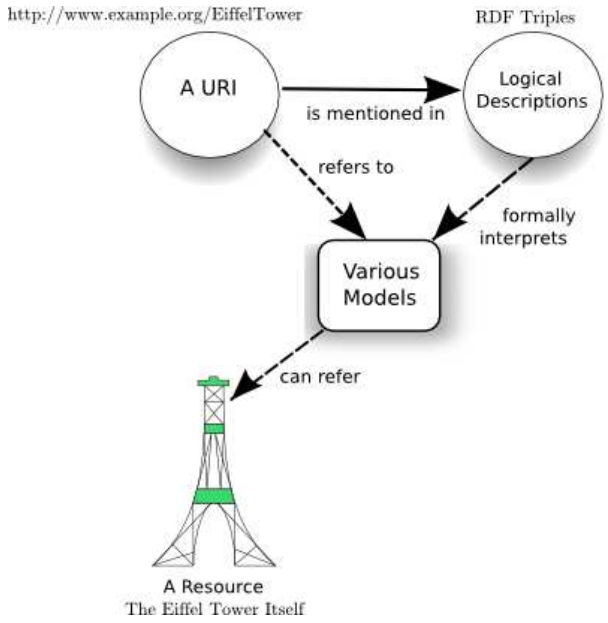

Figure 6.1: The descriptivist theory of reference for URIs

Despite appearances to the contrary, Hayes is not defending a viewpoint arguing for any common-sense understanding of ambiguity, such as how words like ‘bank’ in a natural language like English can have many possible senses. Indeed, what Hayes is arguing is the ambiguity built into formal model-theoretic semantics. This kind of ambiguity is not his discovery, but a well-known issue in formal semantics dating back to the original Scott-Strachey formal semantics (Scott and Strachey, 1971). One question might be whether or not these two traditions – the ambiguity of natural language and the ambiguity of formal model-theoretic interpretations – can be brought together. The essay *In Defense of Ambiguity* touches primarily upon ambiguity in model-theoretic interpretations, although examples are deployed from natural language, in laying out a vigorous case against Berners-Lee’s position that a “URI denotes one thing” (Hayes and Halpin, 2008). What is at stake is the Principle of Universality, namely that anything can be identified by a URI. Hayes puts forward the thesis that the word ‘identify’ is simply incoherent, as it has two distinct readings, as explored earlier in Section 4.2.1, that of *access* and *reference*.

While Hayes makes no claim that access to some Web representations via HTTP is not possible, he claims that such access to Web representations is orthogonal to the question of what a URI could refer to, since “the architecture of the Web determines access, but has no direct influence on reference” (Hayes and Halpin, 2008). Furthermore,

and this statement shows where Hayes's logicist understanding of ambiguity radically parts path with natural language understandings of ambiguity: Hayes claims that reference to resources is completely *independent* of whatever Web representations can be accessed, even if those contain logical expressions. However, in natural language, one is not completely unconstrained in one's use of reference, but one is instead bound to the ambiguity given in the shared conventions of the language, a point we will return to when trying to explicitly bring these viewpoints together in Section 8.1.3.

Hayes makes it explicit that he subscribes to the logical atomist epistemology of Russell, as he says that "reference can either be established by either description or ostention" with ostention being defined as the use of Russellian demonstrative (like 'that' or 'this') identifying a particular "patch of sense data" via a statement such as 'that is the Eiffel Tower,' just as Russell used the notion of acquaintance (2006). Since most of the things referred to by names are not accessible, reference can only be determined by description, and these descriptions are inherently ambiguous (Hayes and Halpin, 2008).

The argument over the ambiguity of description is exemplified in not only in logical descriptions, but natural language descriptions. If a person is trying to identify the Eiffel Tower to a friend, then the person may attempt to communicate their thought about the Eiffel Tower by uttering a description such as "the monument in Paris." Yet even the friend may think they are talking about the Arc de Triomphe without further information. If the person tries to give further descriptions, such as "the steel tower," then the hearer might think of the Eiffel Tower, but there are no guarantees. The hearer may also think of the steel dome of Galeries Lafayette. Even if the person said, "the structure made by Gustave Eiffel," the hearer may think of a lesser-known structure like La Ruche. One can imagine that with enough descriptions a person could uniquely pick out the referent for the hearer. Even with an infinite amount of descriptions this may be impossible, since it involves the large presumption that the hearer shares our same metaphysical or perceptual ontology of things in the world. The hearer may simply have no conception that the Eiffel Tower even exists, and so may be unable to grasp the referent – reduce the set of possible referents to a unique thing – regardless of the number of descriptions given.

Even what appears to be a stable reference by description can be easily disrupted by new information. Hayes illustrates this by referring to a famous example about whether "a fitted carpet was 'in' an office or 'part of' the office in which "two competent, intelligent adult native speakers of English each discovered, to their mutual

amazement, that the other would believe what they thought was an obviously false claim" but that "over an hour of discussion it gradually emerged, by a process of induction from many examples, that they understood the meaning of 'office' differently" (Hayes and Halpin, 2008). For one person 'office' referred to "roughly, an inhabitable place" while for the other it referred to "something like a volume of space defined by the architectural walls" (Hayes and Halpin, 2008). These two people had shared the same office for years, and only upon the appearance of a carpet, it seemed that they had different mental meanings for 'office' and more generally, for 'room.' Neither are wrong per se, it's just that different concepts of 'office' were being deployed, concepts whose differences were so subtle that only in rare or 'edge' case were their very real differences revealed.

On the Semantic Web, the negative effects of adding new information also hold. Often simple formal ontologies are more stable, as "if all one wants to say about persons is that they have mailboxes and friends, then one can treat 'person' as a simple category" (Hayes and Halpin, 2008). Even when a stable situation of mutual reference has been reached in some simple formal ontology, it can be upset by the addition of new ontological distinctions, as can be made by so-called "upper ontologies" such as DOLCE (Gangemi et al., 2002). For example, DOLCE claims that the identity of a person continues over time, while other upper-level ontologies do not (Gangemi et al., 2002). Does the Semantic Web distinguish "Tim Berners-Lee the continuant from Tim Berners-Lee the four dimensional history?" (Hayes and Halpin, 2008). For purposes of inference, such a minor distinction can really matter. If one is not careful with one's upper-level ontology, one can produce "immediate logical contradictions, such as inferring that Berners-Lee is both 52 years old and 7 years old" (Hayes and Halpin, 2008).

The situation with descriptions in real life, with the possibility of multiple underlying ontologies and differing interpretations, is thought by Hayes and others to be modeled on the radical model-theoretic pluralism of Tarski-style formal semantics, i.e. for any language "sufficient to express arithmetic" to have many different 'non-standard' models (2008). As our example showed, RDF in general says so little inferentially that many different models can satisfy almost any given RDF statement. Therefore, Hayes considers it essential to ditch the vague word 'identify' as used in URIs, and distinguish between the ability of URIs to access and refer. While access is constrained by Web architecture, according to Hayes, reference is absolutely unconstrained except by formal semantics, and so "the relationship between access and reference is essentially

arbitrary" (Hayes and Halpin, 2008). From this philosophical position, the Identity Crisis dissolves into a pseudo-problem, for the same URI can indeed access a web-page and refer to a person unproblematically, as they no longer have to obey the dictum to identify one thing. Hayes compares this situation to that of *overloading*, using a single name to refer to multiple referents, and instead of being a problem, "it is a way of using names efficiently" and not a problem for communication, as "natural language is rife with lexical ambiguity which does not hinder normal communication," as these ambiguities can almost always be resolved by sufficient context (2008). Overall, the argument of Hayes against Berners-Lee in the Identity Crisis is the position of keeping the formal semantics of reference separate from the Web as given by the Principles of Web architecture.

6.2.4 Logicism Unbound on the Semantic Web

While the logicist position may seem relatively sensible, the logicist position would also hold that the Semantic Web is more or less unremarkable, since "the Semantic Web languages would operate exactly unchanged if the identifiers in them were not URIs at all, and if the Web did not exist" (Hayes, 2006). In this manner, we should be worried, for then the Semantic Web would be no different from the traditional project of knowledge representation in classical artificial intelligence. Indeed, the *first generation* of the Semantic Web was built upon this logicist vision, with a focus on inference, exemplified by the creation of inference programs and hosts of academic papers detailing how description logics could efficiently implement Open World reasoning (Haarslev and Mueller, 2003; Tsarkov and Horrocks, 2003). Given the emphasis on inference, not surprisingly almost all work in producing information for the Semantic Web became focused on the creation of formal ontologies, and while some of the simple ones such as FOAF (Friend-Of-A-Friend) survived, most of these ontologies languish unused (Brickley and Miller, 2000). This complete disregard for the Principles of Web architecture make sense from the logicist perspective, as the referential mechanism of RDF and other Semantic Web languages should have absolutely no relationship with the accessibility of Web representations. While this first generation of the Semantic Web was an academic success story, the Semantic Web nonetheless did not have the tremendous growth of the original hypertext Web. Indeed, its success seems to be confined primarily to becoming a de-facto standard among the knowledge representation community in AI, rather than the more universal vision of Berners-Lee.

There was never a consensus on the first generation Semantic Web about how logical descriptions determine, even ambiguously, the referents of a URI. One implicit viewpoint dominant on the first-generation Semantic Web is a *localist* reading of the scope of URIs; the a URI refers to whatever could satisfy the model of just the current RDF graph given by some Web representation. Yet this makes it difficult, if not impossible, for the Semantic Web to be used for its primary purpose of data integration. One proposal on this point was to assume the localist reading of any Semantic Web statement unless other URIs were explicitly imported via `owl:imports` statements (Parsia and Patel-Schneider, 2006). However, this would put the responsibility for data integration on the server-side hosting of Web representations, not data integration 'on-the-fly' by a user-agent. The second option, the *holist* reading, is that a URI refers to whatever can satisfy the model given by *every* graph that uses the URI on the Semantic Web. Yet this option makes little sense, for as given by the Principle of the Open World, it is impossible to gather all uses of a URI in Semantic Web statements spread throughout the entire Web.

One possibility in combining the Principles of Web architecture with a logicist theory of reference would be to have a URI refer to whatever satisfied all logical descriptions which are accessible from the URI itself, a viewpoint championed by David Booth under the title *URI Declarations* (2008). This particular possibility of using URIs as names would be an almost perfect analogy to Russell's definition of names as a cluster of logical descriptions (Russell, 1905). URI Declarations have a number of advantages over both the localist and holist logicist readings of URIs. First, URI Declarations allow the URI to access "a set of core assertions that are intended to characterize the resource" that can then be determined by the owner of the URI (Booth, 2008). This means that when an agent encounters a previously unseen URI in a Semantic Web statement and the interpretation of the statement itself is not satisfactory, the agent can use the Principle of Self-Description to discover some core assertions. However, the creation of other statements using this URI is not banned, for "different URI users will necessarily wish to make" possibly "mutually incompatible" and so "different sets of assertions involving the URI" (Booth, 2008). According to Booth, these "mandatory core assertions permit the meaning of a URI to be anchored, to prevent it from drifting, and this in turn increases the likelihood that independent assertions made using the URI can be successfully joined" (2008).

While this standpoint makes sense, it is also very limiting for agents and may not encourage re-use, since "if you do not want to accept the core assertions specified by

the URI Declaration, then you should not use that URI to make statements about its denoted resource" (2008). If one doesn't agree with the interpretation of the core assertions in the URI Declaration, then one should mint a new URI. In turn, this violates the strict separation of reference and access that Hayes puts forward as central to the formal semantics of RDF, even though the URI Declaration still maintains a belief in the primacy of logic (Hayes, 2006). Furthermore, it is unclear where the follow-your-nose algorithm should stop in its quest for accessing logical statements. Should an agent follow a HTTP `Link` header, or the `Link` elements in HTML? Should the agent follow HTTP redirect headers, and if so, which ones? These questions are unanswered by the follow-your-nose algorithm. While Rees has developed a more formally specified algorithm called the *URI Documentation Protocol*, there is no W3C standardized follow-your-nose algorithm for logical descriptions associated with a URI, and many other possibilities, such as *Concise Bounded Descriptions* (Rees, 2008; Stickler, 2005). For at least these reasons, URI Declarations have not reached widespread usage.

The inability of a purely descriptivist theory of reference to reach standardization, or even ad-hoc conventional usage, has led the initial first-generation Semantic Web applications to fail. Most of these first generation OWL or RDF(S) ontologies, such as DOLCE, did not in any way re-use URIs and did not let any Web representations be dereferencable from the original URIs (Gangemi et al., 2002). OWL ontologies were stored as one large inaccessible file, difficult to index by search engines and virtually impossible to find by anyone except the creator of the file. This lack of URI re-usage and the inability to communicate about the referents of Semantic URIs have led to the actual possible referents of many Semantic Web URIs to be so drastically underdetermined as to make the URI itself unusable. Strictly speaking, it was impossible to determine a reference except via the relatively weak inference mechanisms of OWL and RDF, which usually did not infer much of interest as predicted by McDermott earlier in 1987. In an attempt to ameliorate the situation, natural language strings were added to describe Semantic Web URIs using properties like `rdfs:label`, but it was left unknown how this information affected the formal semantics. Since an agent could never be clear about the referential status of a Semantic Web URI, rather than trust already-existing Semantic Web URIs, everyone simply created new URIs rather than re-using them. This dire situation has led the first-generation of the Semantic Web to be more like scattered semantic islands rather than vast inter-linked semantic continents, a ghostly web of logical reference separate from the hypertext Web. Yet the failure of this first-generation of the Semantic Web should not be surprising, for it is

not a test of the Semantic Web hypothesis as a knowledge representation language built according to the principles of Web architecture. The first-generation of the Semantic Web has almost *nothing* to do with the Principles of Web architecture besides the Open World Principle, and so is only a decentralized version of knowledge representation as used in classical artificial intelligence with a single logic-based monotonic semantic network language. As such, the failure of the first generation of the Semantic Web is the failure of a decentralized version of the logic-based AI defended by Hayes's *In Defense of Logic* rather than the Semantic Web per se, and this failure should be depressingly familiar (1977).

6.3 The Direct Reference Position and the Causal Theory of Reference

The alternative slogan of Berners-Lee, that "URIs identify one thing," may not be completely untenable after all (2003c). It appears to even be intuitive, for when one says 'I went to visit the Eiffel Tower,' one believes one is talking about a very *particular* thing in the *real* world called the 'Eiffel Tower,' not a cluster of descriptions or model of the world. The direct theory of reference of Berners-Lee has a parallel in philosophy, namely Saul Kripke's 'causal theory of reference,' the classic devastating argument against the descriptivist theory of reference, and so the logicist position of Hayes (Kripke, 1972). In contrast to the descriptivist theory of reference, where the content of any name is determined by ambiguous interpretation of logical descriptions, in the ***causal theory of reference*** *any name refers via some causal chain directly to a referent* (Kripke, 1972).

6.3.1 Kripke's Causal Theory of Proper Names

The causal theory of reference was meant to be an attack on the descriptivist theory of reference attributed to Russell, and its effect in philosophy has been to discredit any neo-Russellian descriptivist theory of reference (Luntley, 1999). Surprisingly, the causal theory of reference also has its origin in logic, since Kripke as a modal logician felt a theory of reference was needed that could make logical statements about things in different logically possible worlds (Kripke, 1972). However, while Kripke did not directly confront the related position of Tarski, his argument does nonetheless attempt to undermine the ambiguity inherent in Tarski's model-theoretic semantics, although

a Tarski-style semantics can merely 'flatten' models of possible words into a singular model (Luntley, 1999). Still, as a response in philosophy of language, it is accepted as a classical refutation of the descriptivist theory of reference.

In Kripke's *Naming and Necessity*, an agent fixes a name to a referent by a process called *baptism*, in which the referent, known through direct acquaintance is associated with a name via some local and causally effective action by the agent (1972). Afterwards, a historical and causal chain between a current user of the name and past users allows the referent of a name to be transmitted unambiguously through time, even in *other possible worlds*. For example, a certain historical personage was given the name 'Gustave Eiffel' via a rather literal baptism, and the name 'Gustave Eiffel' would still refer to that baptized person, even if he had not been the architect of the Eiffel Tower, and so failed to satisfy that definite description. Later, the causal chain of people talking about 'Gustave Eiffel' would identify that very person, even after Gustave Eiffel was dead and gone. In this regard, a name functions much like a representation as given by our representation cycle in Section 3.6, where some baptismal 'input stage' between a name and a thing is necessary to assign the name directly to the referent. Descriptions aren't entirely out of the picture on Kripke's account; they are necessary for disambiguation when the context of use allows more than one interpretation of a name, and they figure in the process by which things actually get their names, if the thing cannot be directly identified. However, this use of descriptions is a mere afterthought with no causal bearing on determining the referent of the name itself, for as Kripke puts it, "let us suppose that we do fix the reference of a name by a description. Even if we do so, we do not then make the name synonymous with the description, but instead we use the name rigidly to refer to the object so named, even in talking about counterfactual situations where the thing named would not satisfy the description in question" (1972). So what is crucial is not satisfying any description, but the act of baptism and the causal transmission of the name.

6.3.2 Putnam's Theory of Natural Kinds

Kripke's examples of the causal theory of reference used proper names, such as 'Cicero' or 'Aristotle,' and he did not extend his analysis to the whole of language in a principled manner. However, Hilary Putnam, in his *The Meaning of 'Meaning,'* extends Kripke's analysis to all sorts of names outside traditional proper names, and in particular Putnam uses for his examples the names of natural kinds (1975). Putnam

was motivated by an attempt to defeat what he believes is the false distinction between intension and extension. The set of logical descriptions, which Putnam identifies with a "psychological state," that something must satisfy to be given a name is the *intension*, while those things in a given interpretation that actually satisfy these descriptions, is the *extension* (1975). Putnam notices that while a single extension can have multiple intensions it satisfies, such as the Eiffel Tower both being "in Paris" and "a monument," a single intension is supposed to have the same extension in a given interpretation. If two people are looking for a "monument in Paris," the Eiffel Tower should satisfy them both, even though the Eiffel Tower can also have many other possible descriptions.

Putnam's analysis can be summarized as follows: Imagine that there is a world "very much like Earth" called 'Twin Earth.' On Twin Earth "the liquid called 'water' is not H_20 but a different liquid" whose chemical formula is abbreviated as XYZ, and that this XYZ is "indistinguishable from water at normal temperatures and pressures", since it "tastes like water and quenches thirst like water" (Putnam, 1975). A person from Earth would *incorrectly* identify XYZ for their normal referent of water, as it would satisfy all their descriptions. In this regard, this shows that meanings "ain't in the head" but are in fact determined, not by individual language use or descriptions, but by some indexical relationship to "stuff that is like water around here" normally. That "stuff" *should* get its name and meaning from *experts*, since "probably every adult speaker even knows the necessary and sufficient condition 'water is H_20,' but only a few adult speakers could distinguish water from liquids which superficially resembled water...in case of doubt, other speakers would rely on the judgment of these 'expert' speakers" who would ideally test XYZ and determine that it was indeed, not water (Putnam, 1975). Indeed, less outlandish examples, such as the difference between "beech trees" and "elm trees" are trotted out by Putnam to show that a large amount of our names for things, perhaps even extending beyond natural kinds, are actually determined by expert knowledge (1975). In this way, Kripke's baptism can extend to almost all languages, and scientists can be considered a special sort of naming authority capable of baptizing all sorts of things with a greater authority than everyone else. As even Putnam explicitly acknowledges "Kripke's doctrine that natural-kind words are rigid designators and our doctrine that they are indexical are but two ways of making the same point" (1975).

6.3.3 Direct Reference on the Web

This expert-ruled causal theory of reference is naturally close to the direct reference position of Berners-Lee, whose background is in expert-created databases. He naturally assumes the causal theory of reference is uncontroversial, for in database schemas, what a term *refers to* is a matter best left to the expert designer of the database. So Kripke and Putnam's account of unambiguous names can then be transposed to the Web with a few minor variations in order to obey Berner-Lee's "crazy" dictum that "URIs identify one thing" regardless of interpretation or even accessible Web representations (2003c). While it may be a surprise to find Berners-Lee to be a closet Kripkean, Berners-Lee says as much, "that the Web is not the final arbiter of meaning, because URI ownership is primary, and the look-up system of HTTP is...secondary" (Berners-Lee, 2003c). There is also an element of Grice in the direct theory of reference, for the *intended* interpretation and perhaps even purpose of the owner is the one that really matters to Berners-Lee, not any publicly accessible particular Web representation (1957). However, ultimately Berners-Lee has far more in common with the causal theory of reference, since although the URI owner's intention determines the referent, after the minting of the new URI for the resource, the intended interpretation is somehow never supposed to vary (Berners-Lee, 1998a).

To apply the causal theory of reference as to URIs, baptism is given by the registration of the domain name, which gives a legally binding owner to a URI. The referent of a URI is established by fiat by the owner, and then optionally can be communicated to others in a causal chain in the form of publishing Web representations accessible from the URI or by creating Semantic Web statements about the URI. This causal theory of reference for URIs is illustrated in Figure 6.2.

In this manner, the owner of the URI can thereby determine the referent of the URI and communicate it to others, but ultimately the act of baptism and so the determination of the referent are in the hands of the owner of the URI, the self-professed 'expert' in the new vocabulary term introduced to the Semantic Web by his URI, and the owner has no real responsibility to host any Web representations at the URI. Since the owner can causally establish a name for a non-Web accessible thing via simply minting a new URI without hosting *any* representation, under the causal theory of reference the Semantic Web can be treated as having a giant translation manual mapping URIs directly to referents, where the URIs refer directly to objects in the world outside of the Web. In this manual, one could look up the URI *http://www.example.org/Gustave_Eiffel* and

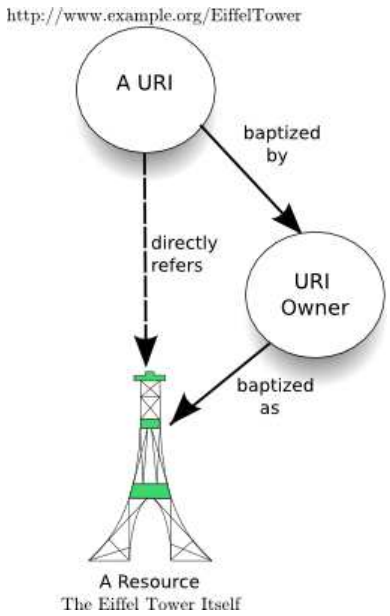

Figure 6.2: The causal theory of reference for URIs

get back Gustave Eiffel himself. From the direct reference position, if an agent got a URI like `http://www.example.org/Gustave_Eiffel` and one wanted to know what the URI referred to, one could use a service such as `whois` to look up the owner of the URI, and then call them over the telephone to ask them what the URI referred to if there was any doubt in the matter. Since obviously such URIs cannot access things outside the Web, what kinds of Web representations, if any, could this giant Semantic Web dictionary return? If it returns no Web representation, how can a user-agent distinguish a URI for a referent outside the Web from that of a URI for some Web-accessible resource? This question is partially answered by `303` redirection, but it is far from satisfactory, as it only allows one to recognize when a URI *may* not refer to an information resource, a very weak promise indeed.

6.3.4 Linked Data: The Second-Generation Semantic Web

While some recognized that the purely logicist first-generation Semantic Web of ontologies is a failure, lately the Semantic Web seems to be taking off under a new name, 'Linked Data' (Bizer et al., 2007). ***Linked Data*** is an *application of the principles of Web architecture to the Semantic Web*. Due to its logicist heritage in classical AI, the first-generation Semantic Web neglected to host accessible Web representations or even use HTTP URIs, as URIs were just regarded as a weird sort of name, with ref-

erence and meaning being taken care of by the Tarski-style formal semantics of the Semantic Web. In contrast, Linked Data recommends that HTTP URIs should be used for everything, and that for any non-information resource, one is likely to have to mint not two, but *three* URIs, "an identifier for the resource" as well as two associated descriptions that can be accessed by content negotiation. The first associated description should be a human-readable HTML-based associated description. The second associated description should allow access to RDF (Bizer et al., 2007). Furthermore, Linked Data encourages the Principles of Linking by encouraging interlinking between data-sets. Following the Principle of Self-Description, Linked Data vocabularies are to allow the retrieval of associated descriptions in both RDF and HTML via 303 redirection for non-information resources. Vocabularies used in Linked Data are encouraged to have accessible namespace documents that describe the Semantic Web terms used in the vocabulary. Lastly, in the spirit of the Principle of the Open World and Least Power, the use of simple Semantic Web languages like RDF and RDF(S) are encouraged over more complex languages like OWL.

A few large data-sets, such as a transformation of Wikipedia to RDF called *DBpedia*, as well as geographical data in *Geonames* and biomedical knowledge to RDF in the *Bio2RDF* project were released as Linked Data (Bizer et al., 2008). This Linked Data initiative is the *second generation* of the Semantic Web. Unlike the *first generation* approach, it also implemented a Kripkean distinction between non-information resources and mere representations of these non-information resources. The distinction is Kripkean insofar as the difference between a non-information resource and its associated description (or any other resource) is assumed to be determined absolutely by the owner of the URI. In marked contrast to its predecessor, the second-generation Semantic Web ignored almost all inference, and focused on producing as much Semantic Web information as possible, even if the published data was inconsistent. The growth of Linked Data has so far been astounding, as it grew from a few million to over a 100 million reusable RDF documents, containing possibly billions of triples (Oren et al., 2008).

While the Linked Data initiative created URIs for many things, such as those things referred to by Wikipedia URIs, it has not created URIs for *everything*, such as the local pub and the proper names of people not famous enough to be on Wikipedia. For any real-world Semantic Web application, it is precisely these types of URIs that are necessary for data integration over something as simple as a Semantic Web-enabled review aggregation site like Revyu (Heath and Motta, 2007). Where are these URIs to

come from, especially if the relevant things or owners of things aren't going to mint the URI themselves? However, if every single application creates these not-so-well-known URIs themselves, then each application will create its own distinct URIs, so that these URIs cannot be used for a graph merger or any other sort of information integration. The Linked Data method has so far been to ignore these issues, although in practice the massive export of Wikipedia into DBPedia, a Linked Data-enabled version of the structured data in Wikipedia, seems to have led the way in minting many useful URIs (Auer et al., 2007).

Besides the Linked Data initiative, another outcome of the Kripkean analysis of creating URIs for non-information resources is the OKKAM project, which declares as its motto the famous principle of Occam's Razor, namely rephrasing the famous maxim to "entity identifiers should not be multiplied beyond necessity" (Bouquet et al., 2007a). The goal of this ambitious project is to provide HTTP URIs for every conceivable 'entity,' where an entity is taken to be some concrete 'thing' such as "electronic documents to bound books, from people to cars, from conferences to unicorns" as opposed to a more 'abstract concept' such as "predicates, relations, assertions" (Bouquet et al., 2007a). Roughly speaking, the distinction is equivalent to the distinction in description logics between 'entities' as individuals in an *ABox* and 'concepts' in a *TBox* which assertions can use, so that an OWL reasoner can use the formal ontology (or terminology) of the *TBox* to classify and make assertions about the entities (Horrocks, 1998). Following Hayes's insight that high-level ontological distinctions are *more* likely to produce ambiguity, OKKAM puts forward the thesis that "while any attempt at 'forcing' the use of the same URIs for 'logical resources' [abstract concepts] is likely to fail (as every application context has its own peculiarities, and people tend to have different views even about the same domain), the same does not hold for entities" (Bouquet et al., 2007a). Everyone is likely to disagree about the concept of justice or even personhood but OKKAM supposes there is unlikely to be disagreement about physical entities like Gustave Eiffel or the Eiffel Tower. However, in a decidedly Kripkean move, instead of building a huge database that contains logical descriptions of the entities, OKKAM merely will construct an enormous and open-ended list of Semantic Web URIs to serve as names for referents. OKKAM can be thought of as the reverse of URI declarations, the only documentation resources to be attached to these OKKAM Semantic Web URIs will be non-logical: collections of pictures, text from other webpages which mentions the same referent, and the like. OKKAM stores "untyped data for the reason that typing an entity's attributes would require us to classify the entity"

because any logical description could lead to disagreement and thus harm re-use of the URIs (Bouquet et al., 2007a). OKKAM so hopes to concretely realize the dream of the Semantic Web as a giant manual that can translate URIs for non-information resources to referents, but without logical descriptions at all.

6.4 Conclusion

The direct reference position attempts to philosophically justify a 'common-sense' notion of reference without sense, and thus it is unsurprising that an autodidact like Berners-Lee has his intuitions about reference fall in line with Kripke and Putnam, even if he is not personally familiar with their work. The logicist position hopes to replace sense with the semantic value of being either true or false, as Frege himself did for mathematical statements. So, while the descriptivist and the causal theories of reference may appear to be contradictory, in reality both of these theories of reference attempt to exterminate a rich notion of 'sense' from a theory of meaning. In this way, the logicist and direct reference position, although they approach getting rid of sense in different manners, on an abstract level are guilty of the same maneuver.

It is precisely the Fregean distinction between 'sense' and 'reference' that provoked both Russell and Kripke's intellectual projects to build an entire theory of meaning on top of only reference, since the notion of 'sense' was thought of by both Russell and Kripke as vague and unnecessary. Therefore, the only way forward seems to be to move from the primacy of reference over sense to the primacy of a more all-encompassing notion of sense over reference. As defined earlier, URIs identify resources, which are objective senses. Therefore URIs don't directly refer to anything, they only refer through mediation of a sense. A theory of meaning that takes into account the objective notion of sense needs to be rehabilitated. A hint of the path to be taken ahead is given currently, but in Chapter 8 we present in full this third position based on Wittgenstein's understanding of sense and reference. This follows naturally from our division of content and encoding, as well as the identification of informational content with a Fregean sense. As Dummett put it, "Frege's thesis that sense is objective is thus implicitly an anticipation of Wittgenstein's doctrine that meaning is use" (1993).

However, before moving to a third position on sense and reference, we need to determine whether or not the direct reference position and its realization in Linked Data is actually empirically triumphant or not? While it may seem so due to the large amount

of data being released as Linked Data, already there are problems arising. On the level of theory, Berners-Lee's Kripkean vision of the Semantic Web as a giant database that maps from URIs to referents is immediately beset by Quine's famous thesis on the *indeterminacy of translation* (1960). The application of the argument of radical translation and interpretation is not explored in detail in the thesis, the interested reader can consult Hayes and Halpin (2008). However, before criticizing the direct reference position purely on theoretical grounds, an empirical examination of the second-generation Semantic Web and Linked Data, needs to be undertaken. Perhaps people on Liked Data do indeed have each URI refer to a unique thing, and that they really are re-using URIs. Unlike the earlier logicist Semantic Web, this possibility cannot be dismissed but needs to be investigated empirically, as the Linked Data Web actually exists in the wild. This empirical work is done in the next chapter, Chapter 7.

Chapter 7

An Empirical Analysis of the Semantic Web

The Database of Intentions is simply this: the aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result. **John Batelle** (2003)

Are there too many URIs for the same thing on the Semantic Web? Or do most things not have a URI on the Semantic Web? Only a large-scale sampling and statistical analysis of the Semantic Web can answer this question. As an added benefit, such a statistical analysis can prove or disprove some widely held assumptions, such as determining if there is an endemic over-use of constructs like `owl:sameAs`, which states that two URIs 'identify the same thing,' and whether the W3C TAG's recommendation of `303` redirection is being followed. Furthermore, such an analysis can quantify the contrast between the direct reference position and the logicist position on the Semantic Web. This can be partially measured by inspecting the deployment (or lack thereof) of constructs in RDF(S) and OWL needed for inference. Only with an empirical analysis of the Semantic Web in hand can we determine the success or failure of Berner-Lee's direct reference position that a URI should identify 'one thing.'

Our methodology is to analyze an hypertext Web search query log to discover a number of non-information resources that *actual* users are attempting to find information about. In particular, we will use a sample of Microsoft's *Live.com* query log to sample the second-generation Semantic Web, the Linked Data Web. Furthermore, our methodology of using a query log leads us to pose and answer the question: Is there anything ordinary users are actually interested in on the Semantic Web?

7.1 Previous Work

For the first-generation of the Semantic Web, there was very little data-driven analysis of the ontologies, primarily because so few were actually in existence. Even in the domains where Semantic Web ontologies existed, due to a lack of following the principles of Web architecture, these ontologies could not easily be discovered. With the advent of Semantic Web search engines such as Swoogle, an empirical analysis of the actual deployment of the Semantic Web became possible (Ding et al., 2004).

The first large-scale analysis of the Semantic Web was done via an inspection of the index of Swoogle by Ding and Finin (2006). In 2006, Ding and Finin first estimated the size of the Semantic Web to be 4.91 million Semantic Web documents via searching Google for the media type `application/rdf+xml` (2006). As this might not include data that is hosted using the wrong media type, using Google to include all FOAF files served as HTML and RSS 1.0 files, Ding and Finin estimated the size of the Semantic Web would optimistically be increased by two magnitudes. By inspecting the index of Swoogle, consisting of 3.7 million URIs with 1.4 million Semantic Web documents, they determined that by far the most popular Semantic Web vocabulary was FOAF (Ding and Finin, 2006). Of the remaining top ten sources of Semantic Web information, the rest consisted of Dublin Core, Proof Markup Language, and RSS 1.0 documents. Both the number of domains hosting Semantic Web documents and the number of distinct URIs in triples were found to exhibit a 'power-law' distribution by visual inspection (Ding and Finin, 2006). As regards the number of sites hosting RDF files, the 'top' of the distribution was found to be `www.livejournal.com`, followed by other social networking sites releasing FOAF files. The most popular Semantic Web term was the `rdf:type` property, followed by FOAF, and then RSS 1.0 (Ding and Finin, 2006).

Although the study of Ding and Finin was of great importance as it was the first empirical study of the Semantic Web, their work has a number of limitations (2006). Its primary limitation was it was unknown if any of the Semantic Web documents contained information that anyone would want to actually re-use. Intuitively, most of the data on this first-generation Semantic Web was likely to be of limited value. For example, the vast majority of data on the Semantic Web in 2006 was caused by Livejournal exporting every user's profile as FOAF – usually without the user's knowledge – without linking to other Semantic Web URIs, serving with the correct MIME type, and deploying `303` re-direction. The second main source of data in Ding and Finin's study,

RSS 1.0, is also of limited value. RSS, originally an XML-based protocol generally used for newsfeeds, was given a RDF-compatible syntax, creating RSS 1.0 (Beged-Dov et al., 2001). First, its use has been surpassed by the non-RDF based Atom and the continued use of XML-based RSS feeds. Second, the very application of RDF in RSS 1.0 is questionable, as the data is primarily information about site updates, and so RSS 1.0 data is rarely merged, re-used, or even linked to in a manner that takes advantage of RDF. Due to the idiosyncratic nature of the data sources of the first generation Semantic Web, it is not surprising that the majority of the data contained little information that could *satisfy the information need* of the average user of the Web.

The principles of Web architecture were finally applied to the Semantic Web in the form of the Linked Data initiative (Bizer et al., 2007). To summarize, the Linked Data initiative required that RDF data actually be accessible from a Semantic Web URI in response to HTTP GET. Furthermore, URIs for non-information resources like entities and concepts were required to use `303` redirection and employ content-negotiation to make both human and machine-readable versions of the information accessible. Other Linked Data good practices are the re-use of URIs, or at least the use of `owl:sameAs` to identify when two URIs identify the 'same' thing, and the interlinking of diverse data-sets. Due to the Linked Data initiative, the size of the Semantic Web has recently increased in size by several magnitudes due to the conversion of a large number of high-quality databases into RDF (Bizer and Seaborne, 2004). Since the study by Ding and Finin missed the rise of Linked Data, the time is ripe for more empirical studies of the Semantic Web. It is unclear how the dynamics of the Semantic Web are changing. While the number of URIs indexed by Linked Data search engines like Sindice shows that the general trend of the number of URIs on the Semantic Web visually follows a 'power-law,' the correct mathematical analysis has not been done to show this to be the case (Oren et al., 2008). The only large-scale study of Linked Data Web at this time has been by Hausenblas et al., and it estimated the size of the Linked Data Web at approximately 2 billion triples (2008). The focus of that study was only on interlinking between data-sets, and it estimated that there were approximately 3 million interlinks between the various data-sets. The most popular interlinking property by far was `dbpedia:hasPhotoCollection`, with approximately 2 million occurrences, most likely due to the term being used by a Linked Data exporter around the popular photo-hosting service Flickr (Auer et al., 2007). In summary, the Linked Data phenomenon is huge, much larger than the first-generation Semantic Web, and its properties have not been fully studied. In particular, there has been little work on determining how the

issues of the reference of URIs play out in the wild given by Linked Data.

7.2 Sampling the Semantic Web via Query Logs

The main problem facing any empirical analysis of the Semantic Web is one of *sampling*. As almost any database can easily be exported to RDF, any sample of the Semantic Web can be biased by the automated release of large, if ultimately useless, data-sets. This was demonstrated in an exemplary fashion by the release of RSS 1.0 data. RDF vocabulary terms that have little content, such as `rss:item`, quickly bias the statistical analysis. With the advent of Linked Data, this has to some extent already happened with large numbers of databases being released as Linked Data ranging from the BBC's John Peel recordings to the MusicBrainz audio CD collection (Hausenblas et al., 2008). How much of the Linked Data Web is aimed for general use? Obviously, components like DBpedia, the export of Wikipedia to Linked Data, could be very useful (Auer et al., 2007). The vast majority of data released into the Semantic Web is of appeal only to a niche audience, such as the great appeal of Bio2RDF to health care and life-sciences. Just as RSS 1.0 and the Livejournal export of FOAF biased sampling of the first-generation Semantic Web, the release of a large Linked Data sets such as the Bio2RDF, containing approximately 65 million triples and so rivaling the size of DBpedia, can bias any sampling of the second-generation Semantic Web (Belleau et al., 2008; Auer et al., 2007). For example, if one just counted the number of URIs used on the Semantic Web, one would quickly find that `bio2rdf:xProteinLinks` would prove to be, in sheer number, a very popular term despite its relative lack of use outside the biomedical community. It is a small step then to imagine 'semantic spamming' that releases large amounts of bogus URIs into the Semantic Web. Furthermore, due to the Open World Principle, it is impossible to determine how many actual separate providers of Semantic Web data there are, so a priori choosing seed samples or to 'weight' any sample is difficult to do in a principled manner. Unlike the original Web, which grew at least in an organic fashion for its first few years, the second-generation Semantic Web progresses in very noticeable 'fits and starts' as large data-sets are released, so each data-set can vastly alter any empirical analysis. The question is not how to avoid bias in sampling, but *to choose the kind of bias one wants*. We are aiming for a bias towards the ordinary user of the Web.

What information is available on the Semantic Web that ordinary users are actually interested in, and how do we sample this data? The obvious candidate for exploring

this would be to look at a major search engine query log, as it gives a sample of the interests of many users in aggregate. Since Semantic Web search engines are currently used mostly by Semantic Web developers and not by ordinary users, the query log of a popular hypertext search engine should be sampled as opposed to a more specialized search engine. Furthermore, the query log should be from a general purpose search engine, not one that puts some constraints on the search such as searching only within bibliographies, as that would prematurely restrict the kinds of queries. The entire bet of the Semantic Web is that it will contain information that many ordinary users will want to re-use and merge via Semantic-Web enabled applications, and that this information will primarily be about non-information resources such as entities like people and places and abstract concepts. Thus, the ideal sampling of the Semantic Web would be to extract query terms referring to physical entities and abstract concepts from a hypertext search engine query log, and then by virtue of a Semantic Web search engine we can determine precisely how much information the Semantic Web contains on these subjects.

7.2.1 The Live.com Query Log

There has been much work on query log analysis in order to discover how to best satisfy the information needs of users on the Web. Since most search query logs of any size belong to search engine companies, it is often difficult for researchers outside those companies to analyze these query logs, and therefore most research in search query logs deal with small or special-purpose query logs, such as the Web track in the TREC competition (Hawking et al., 2000). A few employees of large search corporations have released detailed studies of their search engine query logs. In particular Silverstein et al.'s analysis of a billion queries in the Altavista query log is considered to be a large 'gold-standard' study of query logs (Silverstein et al., 1999).

In order to extract concepts and entities, we analyze the query log of approximately 15 million distinct queries from Microsoft Live Search, and all references to the 'query log' are to this Microsoft query log, as provided by Microsoft due to a 2007 'Beyond Search' award. This query log contains 14,921,285 queries. Of these queries, 7,095,302 (48%) were unique. Corrected for capitalization, 4,465,912 (30%) were unique. Of all queries, only 228,593 (2%) queries used some form of advanced keywords, while 709,102 (5%) used boolean operators and 266,308 (2%) used quotation, leading to a total of 1,204,003 (17%) queries using some advanced techniques

provided by the search engine. The average number of terms per query was 1.76. Note that these extremely brief queries are normal for hypertext Web search engines, with an average query length of 2.35 being reported by Silverstein et al. for the Altavista query log (Silverstein et al., 1999).

7.2.2 Kinds of Queries

Search engine query studies show generally three distinct kinds of user querying behavior: *navigational* queries, *transactional* queries, and *informational* queries (Broder, 2002). For ***navigational*** queries, *the query serves as an abbreviated URI,* such as when the query `Google` is used to access `http://www.google.com`. For ***transactional*** queries, *the query is an attempt to perform a certain transaction*, such as the purchase of a plane ticket. ***Informational queries*** *express the information need of the user for some unknown information*. The query analysis of Broder estimated that informational queries account for 48% of all queries, while transactional queries account for 30% and navigational for 20%, with 2% unclassified (2002). However, studies have shown only a 70-80% confidence in categorizing queries (Jansen et al., 2008). Also, informational queries may *not* be the most important kinds of queries on the Web, since the top ten queries of the *Live.com* query log are *all* navigational queries, as shown in Table 7.1. These distinctions between types of queries are important since only a subset of all queries, *informational* queries, will deal with information that could be found on the Semantic Web. In order for there to be a fair analysis of the Semantic Web, transactional and navigational queries should be removed if possible from the query log.

In an attempt to remove at least a subset of the navigational queries, any query containing a top-level domain (also known as 'TLD,' such as `.com`) was removed from the query log. While this would have removed `google.com` it fails to remove just `google`, so this was augmented by removing the non-TLD form of the top 500 websites as provided by Alexa.[1] Combined, this removed 953,720 (6%) queries from the query log.

The top ten queries of the *Live.com* query log, with navigational and transactional queries manually removed, are given in Table 7.2. When navigational queries are removed, a second trend is that popular queries on the Web are heavily dependent on time. Obviously, these queries are mostly related to either well-known people and

[1] A service that ranks popular websites, available at `http://www.alexa.com/`.

154398	google
132652	yahoo
85664	myspace
72992	ebay
37675	mapquest
27353	my space
23452	aol
20703	american idol
20313	yahoo mail
16060	map quest

Table 7.1: Top 10 queries in query log

11383	weather
7311	david blaine
5279	games
5085	nascar
4815	lyrics
4814	videotaped killing
4418	maps
4039	kelly blue book
2950	dracula castle
2939	ohio bear attack

Table 7.2: Top 10 queries filtered for entities and concepts

events in the news at the time of the query log collection. At the time of query log collection, David Blaine attempted to break various world-records on live television, and there was a high-profile video-taped killing in Kansas. Some of the queries are genuinely general purpose queries, such as 'people' and 'lyrics.' Due to the fact that the top queries tend to be navigational queries *and* that the most popular queries are driven by current events, a sampling regime that is not biased towards the usually transient popularity of a query is necessary.

It should be clear that queries for information about entities and concepts (i.e. non-information resources) will be a sub-category of the much wider class of informational queries. For example, an informational query might be for the 'weather report for Paris,' perhaps phrased as the query `weather Paris`, while the types of queries for physical entities like the Eiffel Tower could be the precise `when was the Eiffel Tower built?` or the foreshortened `Eiffel Tower`. While the distinction between informational queries for an information resource as opposed to informational queries about a non-information resource is fuzzy, this is due to the use of varying levels of abstraction that can be used in terms of interpreting the information need expressed by the query. This problem is made especially difficult given the small number of words used in Web search queries. Due to this problem, it should be expected that any sampling of the query log should be overly vigilant in the attempted deletion of transactional and navigational queries, while at the same time liberal in the acceptance of possible informational queries, not trying to distinguish a query for a weather report from a query about the weather itself.

7.2.3 Extracting Queries for Entities and Concepts

Automatically classifying informational queries is difficult. Rule-based approaches that claim to work over entire query logs like those of Jansen et al. are dubious at best, since they work by applying very loose specifications such as "query length greater than 2" and "any query using natural language terms" (2008). More promising work has applied both supervised and unsupervised machine-learning to discover informational queries, but only achieved an accuracy of 50% when examined by human judges (Baeza-Yates et al., 2006). A number of machine-learning algorithms could be employed to learn named entities, but the sparse amount of linguistic context in query logs makes identifying a named entity difficult in an unsupervised manner, and there is virtually no labeled data for supervised learning (Whitelaw et al., 2008). Even most

rule-based approaches for named entity recognition rely heavily upon capitalization and punctuation, such as 'I.B.M.' and 'Gustave Eiffel,' features that are lacking from query logs (Mikheev et al., 1998).

We call *queries that are automatically identified to be about physical entities in the query log* ***entity queries***. For the discovery of entity queries, people and places are obvious places to begin. An updated version of the system that was the highest performer at MUC-7 (Mikheev et al., 1998), a straightforward gazetteer-based and rule-based named entity recognizer, was employed to discover the names of people and places. The gazetteer for names was based on a list of names maintained by the Social Security Administration and the gazetteer for place names was based on the gazetteer provided by the Alexandria Digital Library Project. Although it could be possible to separate out people and places, this was not done. First, both of these are types of entities. Second, the names of many locations such as 'Paris' can also be used as a name, such as the proper name 'Paris Hilton.' This gazetteer-based approach was chosen to provide high precision, even at the cost of a dramatically reduced recall. This is an acceptable trade-off as we are attempting only to sample the number of queries that would be likely to have URIs on the Semantic Web. A high-quality sample of the query log is more important than a large one for this purpose. Of a random sample of 100 entity queries, a judge considered 94% to be correctly categorized as entities such as people or places.

From the unique queries in the query log, totaling 4,465,912 queries, a total of 509,659 queries (11%) were identified as either people or places by the named-entity recognizer. The top 10 *entity queries* are given in Table 7.3. Some transactional and navigational queries, despite their relatively lower frequency overall in the query log, are highly clustered towards the top of the entity query distribution. These navigational queries such as `chase` and `office max` have clearly snuck into the top ten due to their use of common names in their website names. Furthermore, a number of queries for brands that use names, like 'harley davidson' or 'nick' are present. Still, a number of legitimate real proper names for entities, such as 'jessica alba' and 'marcus vick' were discovered.

A method for discovering abstract concepts in the query log is more challenging. These queries are called ***concept queries***, *queries that are automatically identified to be about abstract concepts in a query log*. Previous attempts at discovering abstract concepts have employed machine-learning over truly massive query logs and document collections from Google (Paşca, 2007). Since this massive amount of data was not

7311	david blaine
4039	kelly blue book
3053	chase
2997	jessica alba
2100	nick
1415	office max
1280	michael hayden
1139	harley davidson
1098	marcus vick
1092	keith urban

Table 7.3: Top 10 entity queries in query log

available, we employed WordNet instead. WordNet consists of approximately 207,000 words with unique synsets. Our algorithm for discovering abstract concepts in query logs using WordNet was straightforward: we only chose queries of length one where the query had a hyponym and hypernym, due to the difficulty of WordNet dealing with some multi-word queries. This assured that the query was for a class that was suitably abstract (having a hyponym) but not so abstract as to be virtually meaningless (had a hypernym). This resulted in a more restricted 16,698 concept queries (.004% of total queries in the query log). The top 10 concept queries are given in Table 7.4. Again, a number of clearly transactional queries have managed to find themselves among the concept queries, such as `chase` and `drudge`, as well as a number of queries where the sense of a word has been taken over by a proper name, such as `sprint` and `aim`. Again, this is due to the preponderance of navigational names towards the top of the query distribution. Of a random sample of 100 concept queries, a judge considered 98% to be classified correctly as concepts. The top ten concept queries are presented in Table 7.4. While some of the queries could be considered somewhat navigational (such as those for maps and dictionaries), they could all be considered informational queries about some abstract concept.

7.2.4 Power-Law Detection

when rank-ordered, the frequency of queries follows what is known as a ‘power-law’ distribution, with a relatively small number of very popular queries and a long-tail of

11383	weather
10321	dictionary
3675	people
3217	music
2192	autism
1468	map
1198	travel
1191	pregnancy
1104	news
1052	charter

Table 7.4: Top 10 concept queries in query log

queries only occurring once or twice, where most of the mass of the distribution is in the long tail and the 'top' of the distribution exponentially decreases. Since this distribution is common on the Web, we will define it precisely: A ***power-law*** is a relationship between two scalar quantities x and y of the form:

$$y = cx^{\alpha} + b \tag{7.1}$$

where α and c are constants characterizing the given power-law, and b being some constant or variable dependent on x that becomes constant asymptotically. Typically it is applied to rank-ordered frequency diagrams, where the frequency of some measurement is given on the vertical axis while the rank order of the measurements in terms of their frequency is given on the horizontal axis. The α exponent is the scaling exponent that determines the slope of the top of the distribution and provides the remarkable property of scale-invariance, such that if a true power-law is observed, as more samples are added to the distribution, the α remains constant, i.e. the distribution is 'scale-free' (Watts and Strogatz, 1998). It is crucial to note that a power-law distribution violates the assumptions of the normal Gaussian distribution, such that routine statistics such as averages and standard deviations can be and *usually* are misleading. In fact, one of the most positive signs of a non-normal distribution like a power-law distribution is a very large standard deviation.

One of the most common power law distributions is known as Zipf's Law, which was originally observed in word frequency estimates. Zipf's Law states that given a finite sample of a natural language of adequate size, the frequency of a word is inversely

proportional to the word's rank in a ranked frequency distribution (Zipf, 1949). In other words, the most frequent word 'the' will have twice as many occurrences as the next most frequent word, 'of.' This sort of distribution seems to be apparent in many evolved systems (Cancho and Sole, 2003), from the link structure of the hypertext Web (Barabasi et al., 2000) to financial systems (May et al., 2008). Is such a distribution evident from Linked Data? One important question is how to detect power-law distributions in actual data. Equation 7.1 can also be written as:

$$\log y = \alpha \log x + \log c \tag{7.2}$$

When written in this form, a fundamental property of power-laws becomes apparent: when plotted in log-log space, power-laws are 'straight' lines. Thus, the most widely used method to check whether a distribution follows a power-law is to apply a logarithmic transformation, and then perform linear regression, estimating the slope of the function in logarithmic space to be α, as done by Ding and Finin (2006). However, standard least-square regression has been shown to produce systematic bias, in particular due to fluctuations of the long tail (Clauset et al., 2007). To determine a power-law accurately requires minimizing the bias in the value of the scaling exponent and the beginning of the long tail via maximum likelihood estimation. See Newman (2005) and Clauset et al. (2007) for the technical details.

Determining whether a particular distribution is a 'good fit' for a power-law is difficult, as most 'goodness-of-fit' tests employ normal Gaussian assumptions violated by potential power-law distributions. Luckily, the non-parametric Kolmogorov-Smirnov test can be employed for any distribution and so is ideal for measuring 'goodness-of-fit' of a given finite distribution to a power-law function. While the details are given at length in Clauset et al. (2007), intuitively the Kolmogorov-Smirnov test can be thought of as follows: Given a reference distribution P, such as an ideal power-law distribution generating function, and a sample distribution Q of size n suspected of being a power-law, where one is testing the hypothesis that Q is not drawn from P, then the Kolmogorov-Smirnov test compares the cumulative frequency of both P and Q to discover the greatest discrepancy (the D-statistic) between the two distributions. This D-statistic is then tested against the critical value of p-statistic at n, which varies per function. The Kolmogorov-Smirnov test is valid even for power-law distributions since Q's cumulative density function is asymptotically normally distributed and this can be compared to the cumulative density function of P.

For a power-law distribution generating function, we can get a critical p-value by

generating artificial data using the scaling exponent α and lower-bound equal to those found in the supposed fitted power-law distribution. A power-law is fit to this artificial data, and then the Kolmogorov-Smirnov test is then done for each distribution that was artificially generated comparing it to its *own* fitted power-law. The *p*-value is then just the fraction of the amount of times the *D*-statistic is larger for the artificially-generated distribution than the *D*-statistic of the empirically-found distribution. Therefore, the *larger* the *p*-value, the more likely a genuine power-law has been found in the empirical data. According to Clauset, "once we have calculated our *p*-value, we need to make a decision about whether it is *small enough to rule out* the power-law hypothesis" (emphasis added) (Clauset et al., 2007). The power-law hypothesis is simply that the distribution was generated by a power-law generating function. The null hypothesis is that by chance a function would generate the power-law distribution observed in the empirical data.

The null hypothesis is rejected if the *D* statistic is *more* than the critical *p*-value for *n*, *p* being the probability that the distribution was drawn from a power-law generating function given the estimated parameters. In order to determine how well the power-law method fits, whenever a power-law is reported, the *D*-statistic is also reported, and we will determine whether or not the fit was significant according to the liberal $p > .1$.

The query frequencies for entity and concept queries are plotted in logarithmic space in Figure 7.1. Both entity and concept queries appear to be linear in log-space, and so can be considered candidates for power-laws. Using the method described above, the α of the queries for entities was calculated to be 2.31, with long tail behavior starting around a frequency of 17 and a Kolmogorov-Smirnov *D*-statistic of .0241 ($p > .1$), indicating a significant good fit. The α of the queries for concept queries was calculated to be 2.12, with long tail behavior starting around a frequency of 36 with a Kolmogorov-Smirnov *D*-statistic of .0170 ($p > .1$), also indicating a significant good fit for the power law. Given their two remarkably similar α statistics and high goodness of fit, one can safely conclude that these query logs do indeed follow power-law distributions. This indicates our sample of entities and concepts are representative of the larger query log, which is well-known to follow power-law distributions (Baeza-Yates and Ribeiro-Neto, 1999).

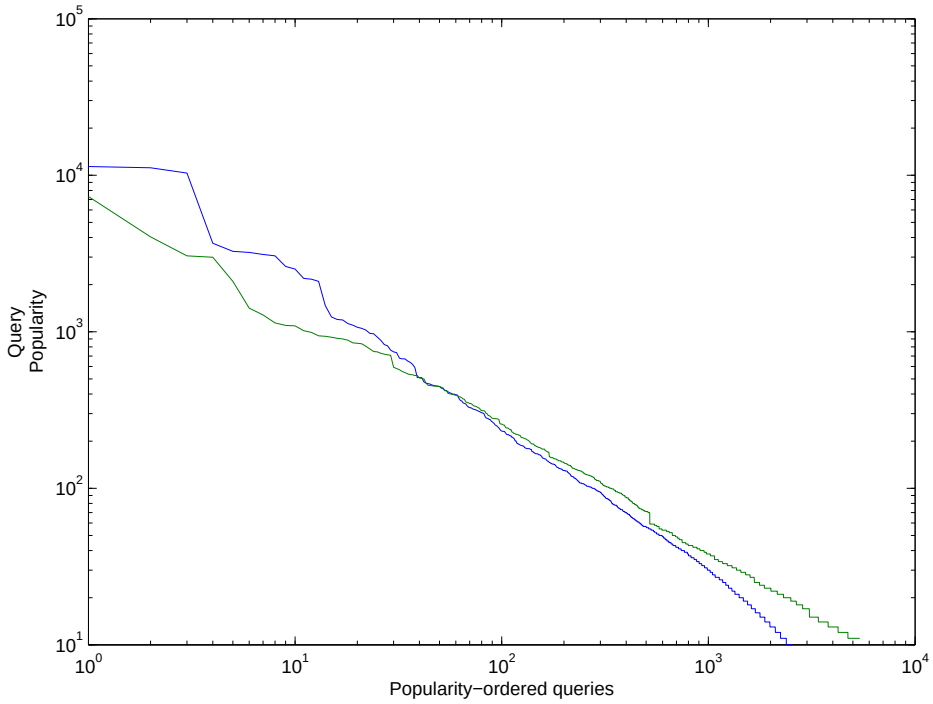

Figure 7.1: The rank-ordered frequency distribution of extracted entity and concept queries, with the entity queries given by green and the concept queries by blue.

7.2.5 Querying the Semantic Web

Both the concept queries and the entity queries are used to query the Semantic Web. Since our goal was to discover how much of interest for ordinary users was present on the Semantic Web, one problem with using the entire query log was that it would contain a vast amount of unique queries that would be unlikely to be repeated. So, we excluded a portion of the long tail from the study by removing all queries of less than frequency 10. The parameter 10 was chosen as it was the number that could reduce both entity and concept queries to the same magnitude. Due to the power-law behavior of both entity and concept queries, this truncation consists of 'removing' a large amount of the long tail, while maintaining the entire 'top' of the power-law distribution, as well as some significant component of the long tail. This procedure is justified insofar as the 'long-tail' likely consists of queries that are never or very rarely repeated, while the remaining queries represents queries that are likely to be repeated. This pruning of low-frequency queries from our sampling does likely exclude many 'difficult' or 'specialist' queries, but we are aiming for queries that are general-purpose

and popular. We call these *queries with more than 10 URIs returned from the Semantic Web* the ***crawled queries*** to distinguish them from the greater query log. Likewise, ***crawled entity queries*** are *entity queries with more than 10 URIs returned from the Semantic Web,* and similarly for ***crawled concept queries***.

This truncation reduced the amount of queries significantly, from 587,283 to 7,848 queries, removing 99% of the queries. It reduced the number of entity queries from 570,585 to 5,308 (a 91% reduction) and the amount of concept queries from 16,698 to 2,540 (an 85% reduction). This gap in the result of pruning off the 'long tail' is interesting, as it shows that while there is a lower amount of concept queries than entity queries overall, concept queries are repeated by a magnitude or so more often than entity queries. The only caveat is that our identification of concept queries via WordNet is likely more stringent than our identification of entity queries, and thus leads to fewer concept queries overall. Furthermore, the vast majority of entity queries, as opposed to concept queries, appear to be queries that are only made once or a very few times. This would make a certain amount of sense, as many queries for people and places are *not* for famous people and places, but for infrequently-mentioned people and places, such as `wayne way`, `san mateo` and `sara matthews`. Some concepts were as diverse as `gastropod` and `accolade`. Still, the crawled queries are still biased significantly in favour of entity queries, with 68% being entity queries and only 32% being concept queries.

The FALCON-S Object Semantic Web search engine (Cheng et al., 2008) was used to query the Semantic Web for selected entity and concept queries between August 3rd and 4th 2008. The results of running the crawled queries against a Semantic Web search engine were surprisingly fruitful, although varying immensely. For entity queries, there was an average of 1,339 URIs (S.D. 8,000) returned per query. On the other hand, for concept queries, there were an average of 26,294 URIs (S.D. 14,1580) returned per query, with no queries returning zero documents. Given the high standard deviation of these results, it is likely that there is either a power-law in the resulting Semantic Web URIs for the queries, or some other non-normal distribution. As shown in Figure 7.2, when plotted in logarithmic space, both entity queries and concept queries show a distribution that is heavily skewed towards a very large number of high-frequency results, with a steep drop-off to almost zero results instead of the characteristic long tail of a power law. Far from having no information that might be relevant to ordinary user queries, the Semantic Web search engines returned either too many URIs possibly relevant to the query or none at all.

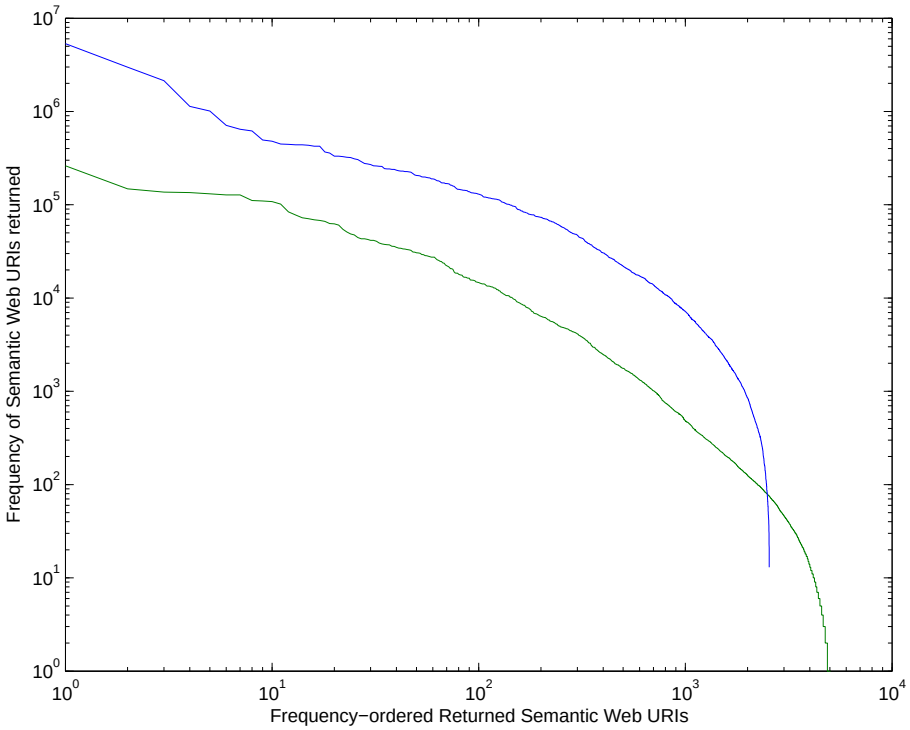

Figure 7.2: The rank-ordered frequency distribution of the number of URIs returned from entity and concept queries, with the entity queries given by green and the concept queries by blue.

Another question is whether or not there is any correlation between the amount of URIs returned from the Semantic Web and the frequency of the query. As shown by Figure 7.3, there is *no* correlation between the amount of URIs returned from the Semantic Web and the popularity of the query. For entity queries, the Spearman's rank correlation statistic was an insignificant .0077 ($p > .05$), while for concept queries, the correlation was still insignificant at .0125 ($p > .05$). Just because a query is popular or unpopular does not mean the Semantic Web will be more or less likely to satisfy the information need of the query. This makes sense, as the vast majority of queries are heavily dependent on current events and fashion, and the Semantic Web is not updated often enough to deal with this kind of information, so there is an inevitable temporal lag between the time information appears in the world outside the Semantic Web and its digitization on the Semantic Web. Yet as shown by Figure 7.2, the amount of *possibly* useful information for the vast majority of queries is still surprisingly large, although how many of the returned Semantic Web URIs are actually relevant to human users is not yet known.

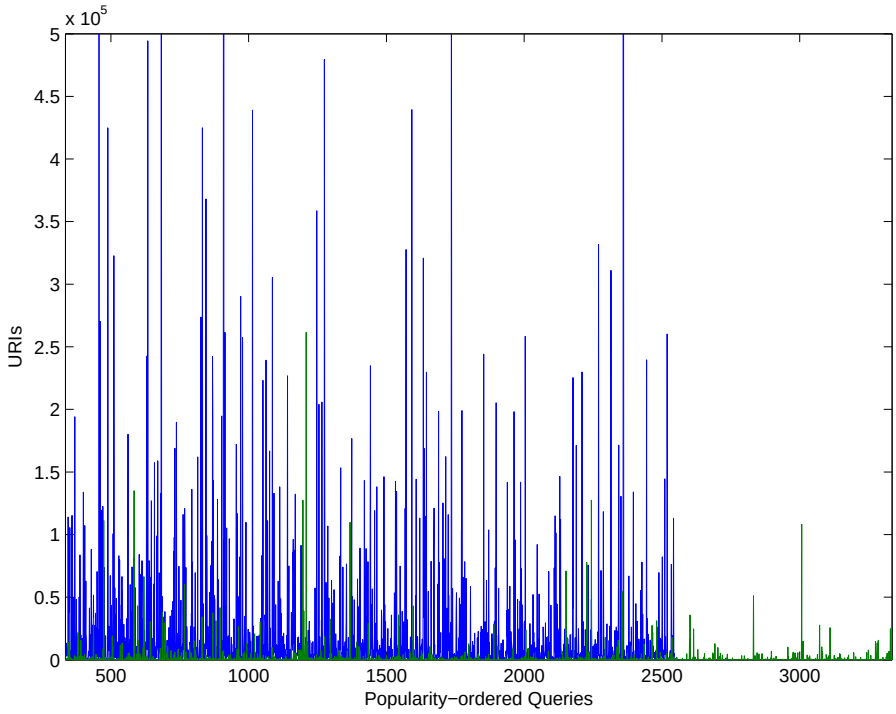

Figure 7.3: The rank-ordered popularity of entity and concept queries is on the *x*-axis, with the *y* axis displaying the number of Semantic Web URIs returned, with the entity queries given by green and the concept queries by blue.

7.3 Empirical Analysis of the Semantic Web

A number of statistics associated with the results of running each query against the Semantic Web are analyzed. First, we investigate statistics about these Semantic Web URIs and their resources themselves, such as their associated status codes and media types. In particular, we focus on the relative prominence of `303` redirection and the hash convention. Then we statistically inspect the URIs actually conveyed by the Semantic Web documents accessible from these URIs.

Surprisingly, there is a deluge of possible Semantic Web URIs for any given query. Due to the high number of results for each query, we restricted our analysis to *the top 10 Semantic Web URI results for each query* as given by FALCON-S's Page-ranking based algorithm and distinguish this subset from all the URIs returned by the Semantic Web, by calling this subset the ***crawled URIs***. ***Concept URIs*** are *crawled URIs from the crawled concept queries* while ***entity URIs*** are *crawled URIs from the crawled entity queries*. Although crawled URIs are a small subset of the total URIs retrieved, given

that user behavior in general inspects the first ten URIs returned by this search (Granka et al., 2004), it makes more sense to sample these ten URIs per query than to sample every URI retrieved. The crawled URIs totaled 70,128 URIs, composed of 25,400 (36%) concept URIs and 44,728 (63.78%) entity URIs. These URIs were crawled using HTTP GET with a preference for application-type of `application+rdf/xml` in order to prefer RDF files served by content negotiation, and any `303` redirection was followed.

Of all crawled queries, a total of 6,673 (85%) had at least 10 crawled URIs. All concept queries had at least 10 crawled URIs and only 4,133 of the entity queries (12%) did not have 10 URIs. Inspecting just the set of queries that did not have 10 crawled URIs, the average number of URIs when 10 URIs were not returned was 2.89 (S.D. 2.88). So, the trend observed earlier was repeated in this smaller data-set, namely that while most of the time too many URIs are retrieved from the Semantic Web, sometimes there are *no* URIs retrieved from the Semantic Web for certain entity queries. Looking at the data more closely, 357 (30%) of the crawled queries with less than 10 results returned *no* URIs, while 138 (12%) returned a single URI and 113 returned two URIs (10%). These queries with zero results seem to be mainly for not well-known places such as `playa linda` (a hotel in Majorica), fairly unknown people such as `william ravies`, misspellings, or popular truncations of names for people such as `steven colbertbush`. This observation helps to explain the sudden drop in Semantic Web URIs returned for queries in Figure 7.1. There was little overlap between the the crawled URIs retrieved by different queries, with an overlap in entity queries of 546 URIs (1%) and an overlap in concept queries of 1031 URIs (4%). In other words, the various queries weren't just retrieving the same small group of URIs over and over again.

7.3.1 URI-based Statistics

In this section, we inspect the various kinds of statistics we can detect on the 'macro-level' of the crawled URIs without actually accessing any Semantic Web documents from the URIs. For all crawled URIs, Web representations were found to be served with 12 different media types. In the event of any forwarding (such as use of the `303` or hash convention), the media-type of the retrieved file was reported. The vast majority of Web representations retrieved from crawled URIs (93%) used the correct media-type (`application/rdf+xml`), although the amount of URIs returned with the

56,893	93.31%	application/rdf+xml
2,410	3.44%	text/plain
1,246	2.04%	text/html
167	.27%	image/jpeg
147	.24%	application/xml
54	.09%	text/xml
31	.05%	image/png
14	.02%	image/svg+xml
3	.00%	image/gif
2	.00%	application/rdf-xml

Table 7.5: Top 10 media types

`text/plain` is large, followed by `text/html` and `application/xml`. This is likely a side effect of being unable to access or being unable to override the default media-types given by the Web server.

The HTTP status returned by attempting to access the various crawled URIs is given in Table 7.6. In particular, the most revealing statistic is that the majority of the Semantic Web sampled by the crawled URIs is served using the `303` convention, not the hash convention. In fact, a total of 51,762 (73%) of crawled URIs use the 303 convention, while only 1,662 (2%) of the crawled URIs use the hash convention. Of these URIs returning the hash convention, manual inspection showed many to be FOAF files. This shows the vast majority of the second-generation Semantic Web is following the `303` convention and so obeying the W3C and the guide to publishing Linked Data (Bizer et al., 2007). Thus, Berners-Lee's vision is to some extent coming true: The second generation of the Semantic Web is taking off, and is at least implicitly endorsing Berners-Lee's direct reference position. Yet this statistic as regards usage of the `303` convention is misleading in the broad sense, as most of the URIs are from a single source, DBpedia, as shown later in Table 7.7.

The majority of URIs, 51,873 (74%), served a Web representation via `303` redirection, and so returned the `200` status code when the Web representation was accessed after the redirection. `200` status codes without `303` redirection still form a substantial fraction of Semantic Web URIs. There are several reasons for this; all hash convention URIs would by default still technically commit a redirect to be served by a `200` status

51,873	73.97%	303
6,061	8.65%	200
4,517	6.44%	404
4,257	6.07%	500
3,147	4.49%	300
246	0.35%	406
20	0.03%	403
4	0.00%	302
3	0.00%	502

Table 7.6: Top 10 HTTP status codes for crawled URIs

code. However, this is only a minority (27%) of those URIs returning a `200` status code. The rest are likely caused by people serving RDF that do not have the access to the Web server configuration needed to serve RDF using the `303` redirection, while many others may have started serving RDF before the TAG decision was made or are not aware of the TAG decision. For example, some earlier RDF-enabled repositories like W3C WordNet did redirection by `300` redirection. A small percentage may be ordinary web-pages, perhaps containing some meta-data as enabled by GRDDL, that just happened to be indexed by the Semantic Web search engine (Connolly, 2007). Furthermore, of these crawled URIs, 9,156 (13%) URIs had no Web representation that was accessible via HTTP, shown by the use of a `4xx` or a `5xx`-level status code.

The top 10 hosts of Semantic Web data in the crawled URIs are given by Table 7.7. DBpedia, the export of Wikipedia to RDF, dominates the results with 83% of all URIs coming from either Wikipedia or DBpedia (Auer et al., 2007). The W3C itself is the third largest exporter of RDF with a share of 5%. Upon closer inspection, most of the URIs crawled from the W3C derive from the W3C-hosted export of the linguistic database Wordnet. The domain of the Frei Universität Berlin has a significant 2% of all RDF data, which is due primarily to its Flickr photo export to RDF. An RDF-version of Cyc and the biomedical data hosting site Bio2RDF also host small but significant amounts of Semantic Web data (Lenat, 1990; Belleau et al., 2008). The Russian-blog hosting site `Liveinternet.ru` carries on the tradition of FOAF exporting of Livejournal. Truesense is another export of WordNet to RDF, although not as frequently used as W3C Wordnet. Towards the end of the distribution there is the RDF version of Uni-

veristät Trier's widely used DBLP academic citation database and `Ontoworld.org`, a RDF-enabled wiki for the Semantic Web research community (Völkel et al., 2006).

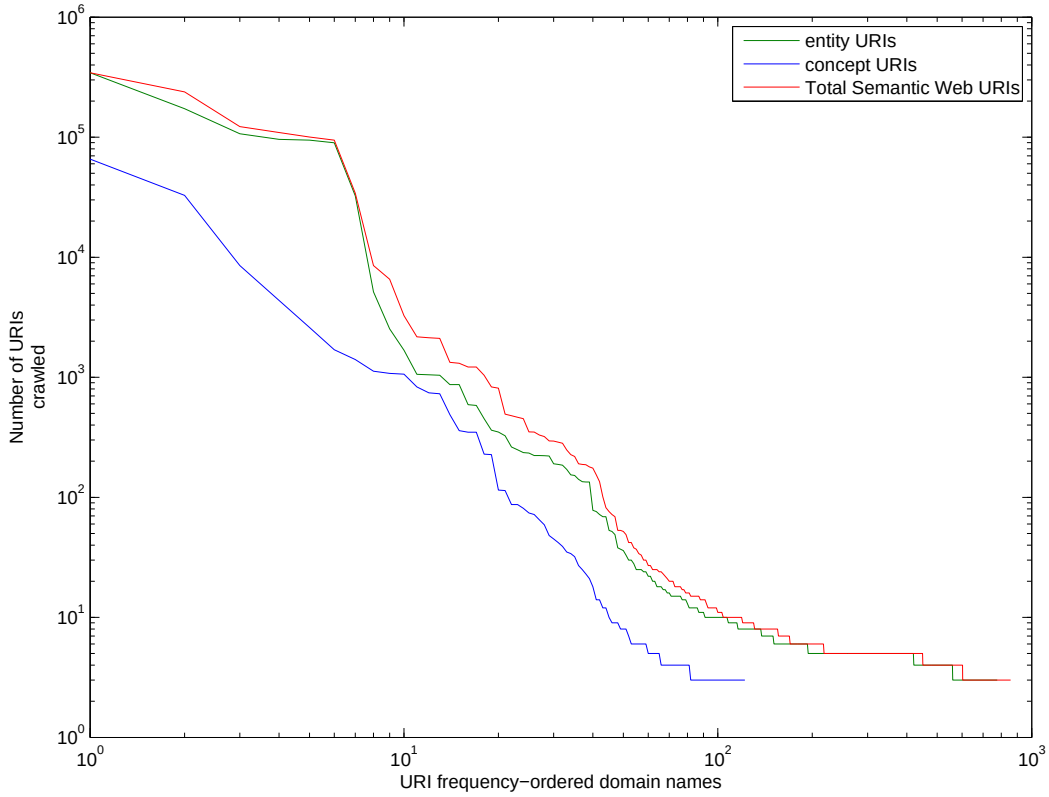

Figure 7.4: The rank-ordered distribution of the domain names hosting Semantic Web data from the crawled URIs ordered by number of URIs hosted.

The average number of URIs hosted by any domain name was 1,268 (S.D. 16,060), with the average number of entity URIs hosted by any domain being 1,236 (S.D. 15,458) and the average number of concept URIs hosted by any domain being 1,0327 URIs (S.D. 6,650). The very high standard deviations are usually a sign of power-law distribution, as shown in Figure 7.4. Attempting to fit a power-law distribution, the α of the rank-ordered domain list frequency distribution is 1.53, with long tail behavior starting around 175 and a Kolmogorov-Smirnov D-statistic of .1414 ($p < .1$), indicating insignificant fit for the power-law distribution. In other words, while a few sources like DBpedia dominate the crawled URIs, with a rapidly decreasing number of smaller sites such as Cyc and the W3C, the long-tail of individual URIs hosting their FOAF files on their personal websites are still rather insignificant compared to the 'top' major

54,698	78.00%	dbpedia.org
3,584	5.11%	wikipedia.org
3,448	4.92%	w3.org
1,704	2.43%	fuberlin.de
811	1.16%	cyc.com
701	1.00%	bio2rdf.org
599	0.85%	liveinternet.ru
417	0.59%	truesense.net
322	0.46%	dblp.unitrier.de
314	0.47%	ontoworld.org

Table 7.7: Top 10 domain names for URIs for crawled URIs

sites hosting Semantic Web data. This is likely because the Linked Data is being artificially generated in large 'chunks' by projects like W3C Wordnet and DBpedia, and so do not organically form the power-law distribution characteristic of naturally-evolving complex systems.

There is some interesting variation in domain names between querying for entities and concepts. While DBpedia dominates both entities and concept URIs, both WordNet and Cyc show themselves to be useful for retrieving information about concepts. This is not surprising, as one of the primary claims of projects like Cyc and WordNet are to encode abstract 'common-sense' knowledge and lexical knowledge respectively, and this would naturally fall more under the domain of abstract concepts than physical entities.

The top ten domains of crawled URIs for entity queries are given in Table 7.8 and are noticeably different from the top crawled URIs for concept queries, which are given in Table 7.9. This data-set is even more overwhelmingly dominated by DBpedia, and to a lesser extent, ordinary Wikipedia URIs that were crawled due to their interlinking with DBpedia. Furthermore, the rest of the domain distribution is more or less the same, although towards the end there is another DBLP bibliographic database and `openlinksw.com`, the site of a commercial Semantic Web and database company. The semi-automatically constructed TAP database of named entities, the oldest large-scale RDF source of data, also appears towards the end (Guha et al., 2003).

More noticeable by its absence than presence is the absence of WordNet and Cyc

18,831	74.14%	dbpedia.org
3,031	11.93%	w3.org
709	2.79%	cyc.com
555	2.19%	bio2rdf.org
243	0.96%	fuberlin.de
169	0.67%	ontoworld.org
222	0.87%	wikipedia.org
132	0.52%	liveinternet.ru
103	0.41%	semanticweb.org

Table 7.8: Top crawled concept URIs

in the list of top sources for entity URIs. Previously in work on lexical resources like WordNet and even machine-readable dictionaries like the Oxford English Dictionary, there has been much focus on the level of terms in the language and on the level of nouns for abstract concepts, and related adjectives, verbs, and adverbs. Many frequently used words, especially those that are of interest to those searching the Web, may not be found so easily among terms in lexical resources like WordNet, since these centrally-curated dictionaries do not include many popular people and places in current events and fashion, such as particular musicians that capture the passing fancy of the moment or particular hotels in popular tourist destinations. Yet collectively-edited databases like Wikipedia do contain such trivial information on current events and fashion, and it is precisely this information that composes much of the information need of Web searches and likely even larger discourse outside the Web.

7.3.2 Triple-based Statistics

In this section, we move our analysis down from the level of URIs to the level of the triples accessible from the URIs. Since a number of crawled URIs were inaccessible (returning some HTTP error code when accessed), this reduced the total number of ***accessible crawled URIs*** to 60,972, a reduction of (13%) from the crawled URIs. The accessible crawled URIs contained 24,074 accessible crawled concept URIs (95% of all crawled concept URIs) and 36,898 accessible crawled entity URIs (82% of all crawled entity URIs). Thus, the accessible crawled URIs maintained a bias towards entity URIs (61% of all accessible crawled URIs) compared to concept URIs (39% of

35,867	80.19%	dbpedia.org
3,362	7.52%	wikipedia.org
1,461	3.26%	fuberlin.de
467	1.04%	www.liveinternet.ru
417	0.93 %	www.w3.org
261	0.58 %	dblp.unitrier.de
171	0.38%	openlinksw.com
145	0.32%	ontoworld.org
139	0.31%	dblp.l3s.de
127	0.28%	tap.stanford.edu

Table 7.9: Top crawled entity URIs

all accessible crawled URIs). Each of the crawled accessible URIs was accessed, and this resulted in a total of 59,228 Web representations with only 48 URIs not allowing access to a Semantic Web document. These non-Semantic Web documents were usually ordinary web-pages from which RDF triples could be extracted via GRDDL or RDFa (Connolly, 2007; Adida et al., 2008). These crawled Semantic Web Documents we will call the ***crawled Semantic Web documents***, and the total sum of triples in these documents are called the ***crawled triples***.

There were a total of 411,574 RDF triples in the crawled triples, with 242,829 (59%) triples for concepts and 168,745 (41%) triples for entity URIs. Concepts seem to require more triples to describe than entities. There were a total of 814,222 URIs in the triples. The internal structure of these triples is of surprising interest. Of these triples, there were a total of 1,051 blank nodes, a measly .25% of all triples in the corpus, of which 772 (73%) were subjects and only 279 (27%) were in the object position. This means that the use of blank nodes, whose purpose is as syntactic placeholders in URIs for objects like lists and in representing *n*-ary arguments in RDF, is almost non-existent in our sample. Of the non-blank node triples, the composition was split between URI nodes (66%) and a surprisingly large minority of RDF literals nodes (34%). These literals contain some form of information in either ‘unstructured’ natural language or some form of structured information in a formal language, such as integer values.

Of the literals, a total of 403,119 were RDF string literals, while only 2% were of

403,119	97.95%	RDF plain literal
3,103	0.75%	http://www.w3.org/2001/XMLSchema#integer
2,789	0.68%	http://www.w3.org/2001/XMLSchema#string
1,185	0.29%	http://www.w3.org/2001/XMLSchema#double
522	0.13%	http://www.w3.org/2001/XMLSchema#date
248	0.06%	http://www.w3.org/2001/XMLSchema#float
136	0.03%	http://www.w3.org/2001/XMLSchema#gYear
65	0.02%	http://www.w3.org/2001/XMLSchema#gYearMonth
59	0.01%	http://dbpedia.org/units/Rank
46	0.01%	http://dbpedia.org/units/Dollar
14	0.00%	http://www.w3.org/2001/XMLSchema#int
9	0.00%	http://dbpedia.org/units/Percent

Table 7.10: Common data types in crawled triples

some other data type, with the top 10 frequent data-types given in Table 7.10. The most frequent data-types are from XML Schema (Biron and Malhotra, 2004), while others are customized for DBpedia. It appears that the vast majority of RDF in the Semantic Web of interest to average users are simple URI-based triples with rich information in natural language. This also goes against the intuition of Berners-Lee that the vast majority of Semantic Web data that is of interest to ordinary users would be the highly structured data of exported databases (1998c) and against the logicist programme for complex ontologies that enable rich inference. Instead, what is of interest on the Semantic Web is stored mainly in natural language, with RDF adding only a minimal structure to essentially fragments of natural language. While it could be argued that this particular finding is merely an artifact of DBpedia, it should be acknowledged that DBpedia *is* most of Linked Data, at least in our query-based sample. We are not studying the Semantic Web as some of its designers would *like* to have it, but as it *actually* exists, and part of its existence is that DBpedia forms a huge central cluster that for ordinary users is the most interesting and useful part of Linked Data. However, it is very possible that this is also an artifact of the indexing of FALCON-S, which also concentrates on DBpedia.

One interesting question is the predominance of the various kinds of Semantic Web knowledge representation terms on the Semantic Web, since this would show what

kinds of inference could actually be deployed on the Semantic Web. First, of the total 1,093,212 URIs in triples harvested from the crawled accessible URIs, only 243,776 (22%) were from one of the primary W3C Semantic Web knowledge representation languages, either RDF, RDF(S), or OWL. Of these, the RDF vocabulary itself was the most popular, with 109,300 URIs (45%), followed fairly closely by the RDF(S) vocabulary with 100,340 URIs (41%), and OWL being dwarfed by RDF and RDF(S) with only 34,136 URIs (14%). This does not mean that OWL is irrelevant to the other corpus, as ontologies constructed with OWL could be deployed to model the concepts and entities employed in 'instance' data. Yet while OWL has been an academic success story, as regards practical deployment, RDF terms and RDF(S)-based inference seems to be the foundation of the Semantic Web in practice.

What precise URI-based terms are used in these knowledge representation languages? The top constructs in either RDF, RDF(S), or OWL in crawled triples are given in Table 7.11. To summarize, RDF(S) class and sub-class reasoning is very popular, with this construction consisting of nearly half (48%) of knowledge representation use of the Semantic Web. The second most popular use of knowledge representation (22%) is for natural language annotation, describing a particular Semantic Web resource using natural language and connecting this natural language description to the URI via the use of `rdfs:comment` or `rdfs:label`. There are surprisingly few (4%) actual ontologies in the crawled Semantic Web resources. Furthermore, non-traditional features of RDF(S), such as the use of `rdfs:property`, frequently occur. Even reification of RDF triples, officially discouraged by the Semantic Web community, accounts for only 95 triples, and there is also fairly heavy use of discouraged RDF constructs to represent different kinds of lists, such as `rdf:Alt` (349 occurrences) and `rdf:Bag` (344 occurrences). Lastly, while many Semantic Web researchers originally hoped that the use of inverse functional properties would allow the merger of Semantic Web data, there were zero explicitly declared usages of `owl:inverseFunctionalProperty`. Overall, the usage of OWL, RDF(S), and RDF terms in the corpus also follows to some degree a power-law like distribution, where α equal to 1.5, with long tail behaviour starting around 90, although the Kolmogorov-Smirnov D-statistic of .1911 ($p < .1$) reveals this to be insignificant. This is because while a few terms vastly dominate, the vast majority of other terms are *not used at all*. This has repercussions for both Semantic Web implementers and vocabulary specification within the W3C, since obviously some level of concentration of effort upon the most frequently-deployed terms would be reasonable.

One of the most popular OWL constructs is indeed the controversial `owl:sameAs`

term, which is used to declare some sort of global equivalence between two URIs. While a tiny portion (.47%) of overall Semantic Web language term usage, it is far from insignificant, with 1,157 occurrences. The use of `owl:sameAs` in the wild is rather different from the role it plays in popular debate than one would suppose. Logicians hold that `owl:sameAs` is only for what is properly considered individuals in description logic, so that classes and properties should use the more restricted and semantically correct `owl:equivalentClass` and `owl:equivalentProperty`. Yet this best practice in logic hasn't reached the Linked Data community, as `owl:equivalentClass` has only 2 occurrences and there are none of `owl:equivalentProperty`. Instead, the Linked Data movement uses `owl:sameAs` to simply "state that another data source also provides information about a specific non-information resource," so leading `owl:sameAs` to tend to mean 'more-or-less the same thing as' (Bizer et al., 2007). This practice leads to the fear that the use of `owl:sameAs` would propagate too far, such that many URIs for perhaps differing referents would be declared identical (Ginsberg, 2006).

Both critiques of `owl:sameAs` appear to be wrong. Given the amount of Semantic Web URIs returned by the queries, while there is considerable use of `owl:sameAs`, it appears that the manual discovery and publication of co-referential URIs using `owl:sameAs` falls far behind the actual growth of the Semantic Web. One could even say that `owl:sameAs` is not being used enough. The real problem is not that distinct things are being given the same URI, but the *reverse*; namely that it appears endemic that the same thing has multiple URIs. Berners-Lee's hypothesis appears to be wrong: A single thing is likely to be identified by more than a single URI on the Semantic Web.

The top 10 Semantic Web vocabularies used in the crawled triples, including those terms outside of the W3C-approved Semantic Web knowledge representation languages, are shown in Table 7.12. The results should not be surprising, in particular the vast dominance of DBpedia. Perhaps surprising is the high frequency of Cyc terms, as well as terms from SKOS, the Simple Knowledge Organization System of the W3C, whose primary source of deployment is the W3C's export of WordNet to RDF (Miles and Bechhofer, 2008). FOAF is also significant, although not nearly as dominant as was found earlier by Ding and Finin (2006). Also popular is YAGO (Yet Another Global Ontology), a merger of WordNet and Wikipedia category hierarchies employed by DBpedia (Suchanek et al., 2007).

There are significant differences in the vocabulary level between entities and concepts. DBpedia URIs occur more often in entity triples than concept triples: 267,323

73,451	30.31%	rdfs:Class
47,044	19.30%	rdfs:comment
44,113	18.10%	rdfs:subClassOf
8,630	3.54%	owl:Ontology
7,256	2.97%	rdfs:label
6,618	2.14%	rdf:Subject
5,107	2.09%	owl:ObjectProperty
3,642	1.49%	rdfs:subPropertyOf
1,157	0.47%	owl:sameAs
535	0.29%	rdfs:range

Table 7.11: RDF and OWL constructs in crawled triples

URIs for entities compared to 66,325 URIs for concepts. There are also far more FOAF URIs in entity triples, ranging from 2,531 FOAF triples as opposed to 732 for concept triples. In contrast, there are 1,105 WordNet URIs in concept triples compared to 731 URIs in entity triples. In general, it seems that the pattern for vocabularies found in URIs holds for vocabularies on the triple-level, and that concepts have a slightly more diverse range of sources than entities.

What URIs are the most popular in the triples themselves? An analysis of the top ten most frequent URIs in *any* position in Semantic Web triples is given in Table 7.13, and the results are of interest. The first triple is the ubiquitous `rdf:type` term that separates predicates, subjects, and objects. Further triples from Cyc, RDF(S), and OWL are also very popular. Yet one very popular URI resource is actually just a Semantic Web version of a Wikipedia redirection, `dbpedia:redirect`. Since most of these URIs are obviously being hosted by `303` redirection, this shows that one crucial error in exporting a database into RDF is the lack of URI re-usage, because these types of large-scale exports simply mint new URIs for everything in the database. For example, it would be far better to have a single URI for these Wikipedia redirections with a single `303` redirection rather than numerous redirections done using a specialized DBpedia vocabulary term inside a Semantic Web document. More surprisingly, bizarre hubs of entities emerge, mainly large lists of entities with common names indexed by Wikipedia, such as a list of Harvard graduates and people who have Dallas, Texas as a hometown. The emergence of these URIs as highly frequent on a popularity list is the

366,849	33.55%	DBpedia URIs
109,300	9.99%	RDF URIs
100,340	9.17%	RDF(S) URIs
94,520	8.65%	Cyc URIs
34,136	3.12%	OWL URIs
6,563	0.60%	SKOS URIs
4,728	0.43%	dblp.l3s.de
3,263	0.29%	FOAF URIS
2,170	0.20%	YAGO URIs
1,836	0.16%	WordNet URI

Table 7.12: Top vocabulary URIs in crawled triples

Semantic Web equivalent, albeit non-malicious, of a link farm on hypertext search engines. Since many people with common names are in these documents, they are heavily linked to, so the employment of algorithms like PageRank over the Semantic Web cannot discriminate these lists of links from the more information-rich Semantic Web documents (Brin and Page, 1998). While the top of the distribution of URIs in triples is a strange mixture of the reassuring and odd, the distribution of URIs in Linked Data follows a power-law distribution, as observed visually by Oren et al. (2008) and shown again in Figure 7.5. Using the maximum likelihood method advocated by Clauset et al., for the first time the actual parameters of this power-law can be given: the α of the power law is 2.00, with long-tail behavior commencing around a frequency of 32, and a Kolomogorov-Smirnov D-statistic of .0157 ($p > .1$), demonstrating an exceptionally good fit (Clauset et al., 2007).

7.4 Conclusion

The empirical analysis of the Semantic Web presented in this study is by no means complete, for it is only a moderately small sample, although it is an important one as this sample is driven by Web search queries by actual users. The results of this empirical analysis show a transformation from the first-generation logicist Semantic Web to the second-generation Web of Linked Data. The Semantic Web as it existed in the first-generation was a motley collection of RDF triples, heavily dominated by a few

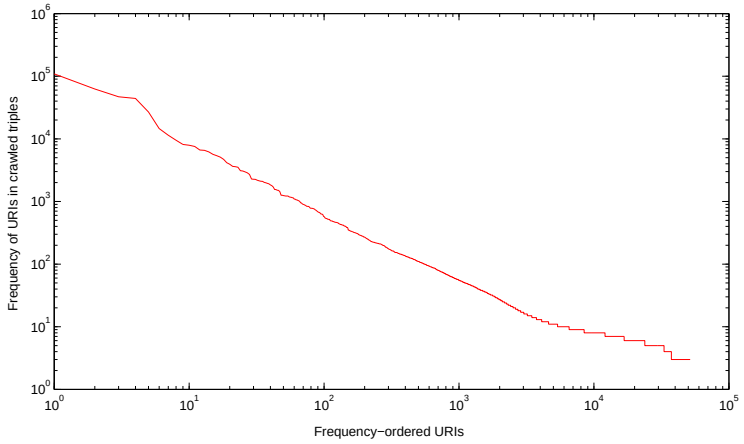

Figure 7.5: The rank-ordered frequency distribution of all distinct URIs in crawled Semantic Web triples.

exports of social networking data into FOAF and a long-tail of complex academically-produced ontologies. Linked Data - at least the section of it that is of interest to users querying the Web for information - is dominated heavily by DBpedia and consists primarily of collections of triples that provide a minimal structure to natural language (Ding and Finin, 2006). While the logicist Semantic Web can be acknowledged as a failure as regards practical deployment, the second-generation Web of Linked Data, heavily inspired by the Principles of Web architecture, has taken off. We have shown that for a wide-range of queries by ordinary users, relevant information may very well be on the Semantic Web. Furthermore, the success of the Linked Data Web points to what appears to be a practical victory for Berners-Lee's direct reference position, as almost all of the Linked Data Web consists of exports of databases and almost all of it employs the `303` redirection convention.

One could argue that these results are more characteristic of FALCON-S and DBpedia than the second-generation 'Linked Data' Semantic Web as a whole. However, we would respond that it is natural in decentralized information systems for power law distributions, where one source of data massively outweighs others in weight to evolve, and the 'giant component' of Linked Data is DBpedia (Barabasi et al., 2000). In fact, if such a 'giant component' and long tail were not observed, it would be cause for suspicion. Furthermore, the results *should* be checked against other Semantic Web search engines besides FALCON-S, and future work with different Semantic Web search engines will be done for future work.

108,909	13.37%	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
62,469	7.67%	http://www.cyc.com/2004/06/04/cyc#guid
47,021	5.77%	http://www.w3.org/2000/01/rdf-schema#comment
44,113	5.42%	http://www.w3.org/2000/01/rdf-schema#subClassOf
26,789	3.29%	http://www.w3.org/2002/07/owl#Class
14,615	1.79%	http://dbpedia.org/property/wikiPageUsesTemplate
11,402	1.40%	http://www.cyc.com/2004/06/04/cyc#EnglishWord
9,492	1.17%	http://dbpedia.org/resource/List_of_Harvard_University_people
8,149	1.17%	http://dbpedia.org/property/redirect
7,918	1.00%	http://dbpedia.org/resource/Dallas%2C_Texas

Table 7.13: Top 10 URIs in crawled triples

On the level of triples, there are some surprising conclusions. The triples on the Semantic Web contain a vast range of data, and the exact kinds of URIs used in the triples are somewhat unpredictable. However, the kinds of vocabularies actually deployed are almost entirely from a few large vocabularies, such as DBpedia, DBLP, WordNet, YAGO, and FOAF. This again points to a victory of Berners-Lee's idea that a few large vocabularies with well-defined terms could dominate the Semantic Web (Berners-Lee and Kagal, 2004). In a further defeat for the logicist position, the kinds of triples that structured this data do not contain many OWL terms optimized for inference, but consist almost entirely of relatively straightforward RDF(S) expressions for sub-class relationships and for annotations in natural language. Overall, the Semantic Web is primarily being used to provide structured relationships between fragments of natural language, and *not* for inference. Given the lack of use of inference and the widespread use of the `303 convention`, the vision of Berners-Lee and the direct reference position is the victor in practice over the earlier logicist Semantic Web championed by Hayes.

All is possibly not well with Berners-Lee's vision of a Semantic Web where reference is established by fiat. The entire purpose of the Semantic Web is supposedly decentralized data integration via the re-use of public identifiers. While the number of RDF properties, or kinds of links, are dominated by a few large vocabularies, as regards re-using URIs to identify things in the world, Linked Data may not be faring well. The most noticeable result of this keyword-driven analysis of the Semantic Web is that a

truly huge list of URIs have been returned for each keyword. At first glance, this has far more in common with the hypertext Web than the Semantic Web, as normal hypertext search engines usually respond with a long list of URIs in response to a hypertext search. However, it should be remembered that there is a crucial difference between the work done in this chapter and hypertext Web search: that we were searching for Semantic Web URIs for concepts and entities, not just relevant web-pages. In the ideal Semantic Search scenario, for every reasonable unambiguous query for an entity or concept, the single 'best' URI for that entity or concept would be returned. So does this mean that necessarily there are many URIs for a single referent being returned? Indeed, at this point, we cannot determine that too many URIs for a single concept or entity exist on the Semantic Web from the experiment results given here without further analysis. A number of alternative hypotheses are possible. As these URIs were returned by common information retrieval techniques, it is very possible that every Semantic Web document that mentioned the term is returned, and this would naturally overgenerate URIs, even if like a golden needle in a haystack, *somewhere* in the list of returned URIs was the one and only one Semantic Web URI for the concept or entity. Second, it could very well be that query itself is ambiguous, and thus naturally there would be more than one URI for an entity or concept returned, as the query term would retrieve at least one URI for every sense. It could even be that each URI denotes a slightly different sense of the term of query term, so that none of these URIs can be thought of as the 'best' Semantic Web URI for that concept or entity. With so many possible hypotheses, at this moment we cannot judge whether or not Berners-Lee's direct reference position to use only one URI for a concept or entity is being followed. What is needed is for humans to inspect at least a subset of these queries to see if any of the returned URIs genuinely do refer to the same entity or concept, as is done in Chapter 9. Yet, first we need to determine how to reduce the ambiguity of the queries themselves, so we can be sure that the returned URIs are genuinely about what referents the agent was trying to express with the keywords. In order to capture the phenomenon of reference in relationship to natural language in a more sophisticated manner than done in Chapter 6, we outline both a new position on sense and reference, and a practical system for capturing the sense of keyword searches in Chapter 8.

Chapter 8

A Solution to the Identity Crisis: From Wittgenstein to Search Engines

The solution to any problem in AI may be found in the writings of Wittgenstein, though the details of implementations are rather sketchy **R.M. Duck-Lewis** (Hirst, 2000), as quoted in Wilks (2008a).

It appears we are at an impasse at the Identity Crisis. First, both the positions championed by Berners-Lee and Hayes seem to ground out in some theory where meaning is determined by reference. While the failure of the first-generation Semantic Web shows that reference via logical descriptions is not enough, the growth of the Linked Data project shows that the application of the principles of Web architecture to knowledge representation works. This in turn seems to have implicitly validated Berners-Lee's direct reference position. Yet that is far from true; what is apparent from our analysis of Linked Data in Chapter 7 is primarily that the Identity Crisis persists in a new form on the Linked Data Web; there appear to be *too many* URIs for some things, while *no* URIs for other things. Having someone declare a URI to refer somehow directly to some referent by *fiat* does not work in a decentralized system like the Web. As differing users export differing representations to the Web in a decentralized manner, new URIs are always minted, leading each Linked Data source to be fairly closed, and so running the risk of fracturing the Semantic Web into isolated 'semantic' islands instead of becoming a unified 'semantic continent.' The critical missing element of the Semantic Web is some mechanism that allows users to come to agreement on URIs and then share and re-use them, a problem ignored both by the logicist and direct reference positions. In this chapter, we outline a third position, the *public language position*.

Rather surprisingly, the way forward is to be found in a footnote of Kripke's in *Naming and Necessity*, where Kripke says that "a name refers to an object if there exists a chain of communication, stretching back to baptism, at each stage of which there was a successful intention to preserve reference" (1972). Yet Kripke stipulates an unreasonable condition. It is almost impossible to determine with certainty if reference has been preserved in of and itself due to the inherit ambiguity in natural language. More importantly, Kripke admits even the *causal* theory of reference is not a *purely* causal story, at least in the way the term 'causal' has been defined in Chapter 3, where causal is exemplified as a purely physical story dependent on local connections, such as when a ball on a billiard table hits another ball and *causes* the latter ball to move. Kripke admits there must be a chain of *communication*, and this communication must exist in the form of information encoded in a *language*, which for distal and so representational content, this language must be phrased in terms of descriptions and depictions. The language responsible for naming conventions that Kripke hints at is not a private language, or a logical language, but a *social* language capable of having causal effects upon the world and its users, and so being "objective" as was required of the concept of sense by Frege (1892). So, in our pursuit of a theory of reference and meaning for the Semantic Web, we are drawn into the waiting arms of Wittgenstein.

8.1 Wittgenstein and the Public Language Position

It is precisely the *social* notion of language that has been strangely missing from the debates on reference and meaning on the Semantic Web so far. One of the hidden presumptions of the logicist position, as promoted from Carnap to Hayes, is the tradition that language can be a *private* phenomenon, that it can be possessed and used by a *single* idealized agent to accurately describe and refer to the world. Wittgenstein, whose *Tractatus* was the original inspiration for this position, returned to refute this point in his *Philosophical Investigations* (1953). In this later work, Wittgenstein gives a forceful argument against private language and logicism, whose defenders he believed had misinterpreted his outlook in the *Tractatus*. This 'late' Wittgenstein opens up the way for a new conception of language based on the *public* use of language.[1] To briefly outline Wittgenstein's arguments in *Philosophical Investigations* is an impossible task, due to both the density of his thought, his brief aphoristic style, and the vast

[1]This adjective 'late' is used to distinguish his philosophy from his earlier work on the *Tractatus*, although the rupture between these two periods may be exaggerated by his interpreters.

range of topics he covers. To complicate matters, the 'late' Wittgenstein has produced a massive secondary literature, which due to space constraints we will ignore, focusing only via direct quotation from Wittgenstein himself on a few aspects of his work with consequences for computational implementations on the Semantic Web.

The purpose of this section is to clarify a few key concepts of Wittgenstein, in particular, his analysis of 'the form of life' and 'language games,' the dictum 'meaning-is-use' and the status of reference in a Wittgensteinian theory of language. From this exposition we will attempt to determine what a Wittgensteinian response to the Identity Crisis would be, a position we call the *public language position*. From this position we will determine the design requirements for a practical implementation for helping to solve the Identity Crisis.

8.1.1 Language Games and Data Integration

When Wittgenstein was arguing with Piero Sraffa that everything in the world must be expressible by the grammar of logic, Sraffa made a flicking of his fingers underneath his chin, asking Wittgenstein, "what was the grammar of that?" (Monk, 1991). Realizing that no logical grammar did justice to Sraffa's act, Wittgenstein abandoned his view of language as logic and rephrased it in terms of a "language game" (1953). The term 'language-game' is "meant to bring into prominence the fact that the speaking of language is part of an activity, or of a form of life" (Wittgenstein, 1953). So, languages are composed of *actions in the world*. Earlier in Chapter 3 we defined the 'meaning' of a term to be the concrete activity of the agent that encounters or uses the term, and so encompasses communicative actions like Sraffa's flicking of the fingers as meaningful. Wittgenstein also points out that all the terms in a language derive their meaning from this interwoven web of action and words, so that the words compose a language in virtue of their relationships to other words and actions, for "these phenomena have no one thing in common which makes us use the same word for all – but that they are related to one another in many different ways. And it is because of this relationship, or these relationships, that we call them all 'language'..." (1953). However, there is no *one* monolithic language, but a variety of different language-games that represent the multiplicity of uses in which language can be applied; "the functions of words" are as diverse as the purposes of "tools in a tool-box" as "there are countless different kinds of use of what we call 'symbols', 'words', 'sentences"' (Wittgenstein, 1953). The purpose of a particular language-game is not the transmission of subjective and inner

intentions from one agent to another in some sort Gricean manner, but the creation of co-ordinated action driven by a purpose (Grice, 1957).

Wittgenstein says "to invent a language could mean to invent an instrument for a particular purpose" (Wittgenstein, 1953). The purposes of evolved natural languages are incredibly varied, but new formal languages are invented for a purpose, at least as we defined the term in Chapter 3. What is the purpose of the Semantic Web? Why would anyone participate in this particular language game rather than the language game of the hypertext Web, or some other language game altogether? On this point, the Semantic Web is positively schizophrenic, vacillating between a *first-generation* vision of classical artificial intelligence replete with inference-driven agents, and *second-generation* vision of opening databases according to the Principles of Web architecture for applications that cannot yet be imagined. Obviously, these purposes have only been successful at attracting artificial intelligence researchers and true believers in Berners-Lee to the fold of the Semantic Web.

What the Semantic Web needs is a convincing purpose that will attract large numbers of users: "the Semantic Web is a solution in need of a problem" (Halpin and Thompson, 2006). The best way to understand the purpose of a language, including a formal language, is not to inspect what the language specification *says* it does, but to observe what it *actually* does in operation. In this, the only benefit of RDF over traditional semantic networks is the use of URIs, which allows differing graphs that share the same RDF to automatically merge. So regardless of what its proponents say, the purpose of the Semantic Web is *data integration*. However, as there is almost no re-use of URIs on the Semantic Web, as a language-game for data-integration the Semantic Web also seems to be a failure. The first-generation of the Semantic Web ignored the re-usage of URIs due to its logicist position that held URIs to be merely an odd sort of symbol, no better or worse than any other. The second-generation of the Semantic Web tends to mint new URIs for everything in order to preserve the unique and particular meaningful use of a term in each database. What is lacking from the Semantic Web is obvious: *agents should be able to easily discover and re-use URIs for things outside the Web like concepts and entities*.

8.1.2 Against Private Language

Wittgenstein attacks the very idea of a *private* language, a language that is somehow only understood by a single person and hence untranslatable to other languages, where

"the individual words of this language are to refer to what can only be known to the person speaking; to his immediate private sensations. So another person cannot understand the language" (1953). His primary example is the use of a language to describe sensations of pain. Wittgenstein argues that such a language is absurd, as there would be no "right" way to use the private word for the sensation, for "whatever is going to seem right to me is right" (Wittgenstein, 1953). In his second famous attack on private language, Wittgenstein phrases an attack on private codes of behavior in the infamous example of rule-following in a game like chess, stating that, "It is not possible that there should have been only one occasion on which only one person obeyed a rule" (Wittgenstein, 1953). There can be no norms for behavior, and therefore no meaning, in a private language game. This follows from the insight that norms ultimately involve *others*, where the norm is repeated in different circumstances and mediates the collective behavior of multiple agents.

On the Semantic Web, the logicist and direct reference positions *both* conceive language as a private language. The causal theory of reference of Kripke, Putnam,and Berners-Lee believes that a name is established by fiat by an individual or some approved authority, such as science or the domain name registry, and so is dependent on some notion of what the individual or science wants a name to 'really' mean. In contrast, the descriptivist theory of reference of Hayes, Russell, and Tarski holds that the referent is established by the use of logical descriptions regardless of what any individual 'means' by the term. However, the descriptivist theory of reference *also* ignores any public or social aspect of the descriptions: the descriptions can be created by an individual without regard to any social convention and the satisfaction of the descriptive terms is given by either objective features of the world or satisfaction of the model. Furthermore, both the causal and descriptivist theory of reference crucially depend on some notion of 'sense-data' that can be assigned a name, either by a description or direct acquaintance.

Strangely enough, there is a deep affinity between both the descriptivist and causal theories of reference, for a Kripkean baptism is just some sort of *causal* relationship between sense data and a name, exemplified by the act of saying 'the name of that is the Eiffel Tower.' This account of baptism is *precisely the same as* Russell's account of the use of names via direct acquaintance with 'sense-data,' given a slightly more modern update with Hayes's account of ostention for naming on the Semantic Web (Hayes and Halpin, 2008). Furthermore, there is in no difference in establishing a name via baptism-acquaintance than there is establishing a name by the use of descriptive terms.

A Russellian descriptivist would simply have some 'sense-data' that they could label with 'that is an iron tower' and then generalize to other sets of 'sense data' to which one can apply the terms 'iron' and 'tower' via more complex logical statements involving towers and their descriptions. Likewise, the idea of direct acquaintance with sense data equally underpins both Putnam and Berners-Lee. Both think that reference should be determined by some "guardians of meaning," for instead of just labeling a patch of sense-data with the term 'iron tower,' the scientists would label the sense-data with the use of a name like 'iron tower' only after it successfully passed some authoritative test, such as a test for the chemical composition of iron (Wilks, 1975).

Using the famous example of the 'duck-rabbit', Wittgenstein undermines the very idea of establishing a referent via direct acquaintance and baptism (1953). After all, if one can not determine that a simple sketch is of a 'duck' or a 'rabbit,' then how can *anyone* objectively and without ambiguity attach a name to some data? The indeterminacy of the infamous 'duck-rabbit' shows that at least in some cases there is no determinate nature of our phenomenological 'sense-data.' Having disposed of the notion of ostention somehow providing direct access to sense-data, baptism of even indeterminate sense-data – by either Kripkean baptism or Russellian descriptions – is attacked next. Wittgenstein holds that any act of baptism is incapable of assigning a name if the act is done by a private individual, "naming appears as a *queer* connection of a world with an object – and you really get such a queer connection of a word when a philosopher tries to bring out the relations between name and thing by staring at an object in front of him and repeating a name or even the word 'this' innumerable times" (Wittgenstein, 1953). Only in the very rarefied form of life known as academic philosophy does this happen even *in theory*. This is because "naming is so far not a move in the language-game any more than putting a piece in its place on a board is a move in chess. We may say: *nothing* has so far been done, when a thing has been named. It has not even *got* a name except in the language game. This is what Frege meant too, when he said that a word has meaning only as part of a sentence" (Wittgenstein, 1953). Indeed, naming as a purely private convention serves no purpose. It is only as part of a wider language-game that anything can have a name in the first place. Even what appears to be the most private of sensory experiences is both determined and expressed by a public language.

8.1.3 The Public Language Position

In order to escape the philosophical quagmire of private language, Wittgenstein points out, "Do not ask yourself 'how does it work with me?' – ask 'What do I know about someone else?' (1953). A language is *public* and inexorably *social*, involving more than one agent. A ***community*** can be defined sparsely as *a group of agents that use the same language*. Note that languages are not monoliths, as an agent may use many languages, and may only share certain intersections of names in various languages with other agents. As a public language-game is used by more than a single agent involved, it is proper to say that a *community uses a language* rather than an individual agent. So a third position, in contrast to both the logicist and direct reference positions, can now be staked. The ***public language position*** states that since *the Semantic Web is a form of language* then as *a language exists as a mechanism for co-ordination among multiple agents, then the meaning of a URI is the use of the URI by a community of agents*.

To contrast this position with the direct reference position, the meaning of a URI is not determined by whatever referent is assigned to it by its owner, unless the owner and other agents actually can come to an agreement on its meaning. The public language position does not give the owner of a URI any particular privilege, except for the obvious asymmetric technical privilege of having the ability to influence the use of the URI through hosting an accessible Web representation or redirecting to another URI.

Unlike the causal theory of reference and the descriptivist theory of reference, Wittgenstein does not equate the meaning of a sentence with 'truth' or the satisfaction of a model as something *outside* the language-game. Wittgenstein retorts that only "in our language" can "we apply the calculus of truth" (1953). From Frege to Tarski, the logicist camp's reduction of meaning to truth-conditions only makes sense in terms of *their* particular language-game of logic, which while useful in the realm of mathematics, fails when the wider social aspects of meaning come into play. The model(s) that satisfy the descriptions are only interesting insofar as the inferences they allow to play meaningful roles within a wider language-game. In the case where the inferences and the use of the URI are at odds, an agent using the URI can just *ignore* the inferences in determining the meaning of the URI.

Ambiguity is built into a Wittgensteinian public language position, and the kind of ambiguity that Wittgenstein is concerned with is not the logicist kind of ambiguity resulting from entailments failing to constrain interpretations. Earlier in Section 6.2,

Hayes defended the notion that names were fundamentally ambiguous. While this is common-sense, he put forward the thesis that reference could not be determined at all by external factors, but is instead determined purely by the individual using the name, who can assign to it any interpretation they wish. This is to some extent similar to Kripke and Berners-Lee's assignment of a referent via baptism as explored in Section 6.3, and as such is also a private language argument. Yet unlike their direct reference position, Hayes holds that the reference given in an interpretation happens to be incommunicable unambiguously via description, as "there will always be some slack, some possible doubt about what exactly is being referred to" (Hayes and Halpin, 2008). Again, the ambiguity in the logicist position is much wider than Wittgensteinian ambiguity. For Wittgenstein, ambiguity is naturally constrained by the conventions of the language game and the form of life, which are restricted in turn by the external world. While the Wittgensteinian public language position would note that there is always some ambiguity in language, worrying about this ambiguity misses the point, as the point of a language game is not to pin down names to referents exactly, but instead to share enough of a convention to accomplish some task or solve some problem. Ambiguity is usually solved by the embodied or implicit context given in the language-game – it is not without reason that Wittgenstein begins the *Philosophical Investigations* contrasting the Augustinian approach of assigning the builders to objects with the language game of builders moving slabs or rock around. For the builders, their task at hand determines their meaning of the word. Thus, some ambiguity may be necessary for successful communication. The role of descriptions and inference is not in determining referents, but only when the various agents in a language-game are not clear about the role of a name in a language game, so that "an explanation may indeed rest on another one that has been given, but none in need of another – unless *we* require it to prevent a misunderstanding" (Wittgenstein, 1953). In this manner, inference and entailments that restrict interpretations, as defended by Hayes, are only a primitive logical analogue to the real-world context that both constrains ambiguity in a language game while usually never dispelling it. While some inferential mechanisms can be useful when errors are made in a language game, in general inference can not express the constraints and even the world given by the contextual use of name in a language game.

From the perspective of the public language position, when a new URI comes into play on the Semantic Web, the agents do not have to specify the referents of the URI to use it meaningfully. This justifies the earlier observation of Hayes that attempts to

over-specify reference can in fact lead to disagreement (Hayes and Halpin, 2008). If the referent of a name has to be specified for the name to be used, it only has to be specified to the minimal conditions necessary to co-ordinate actions between agents. Contra Berners-Lee's direct reference position, only in very rare language games does the referent of some representation have to be specified in an 'unambiguous' manner.

How does the public language position actually play out on the Semantic Web? To apply Wittgenstein to the Semantic Web, the first observation is then that the Semantic Web *is* a *new* language-game. There is no reason why language-games in a Wittgensteinian sense have to be restricted to natural languages, for Wittgenstein himself notes that "new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten" (1953). The struggle over the Identity Crisis within the Semantic Web is precisely the struggle over the conventions of reference needed for a new language. Remember that we have defined earlier in Chapter 3 the term 'language' and 'sense' to be *neutral* between formal languages for computers and natural languages. Formal languages are often mistakenly assumed to be meaningless due to their not taking into account the concrete activity that occurs as a result of their use but instead to be pure "syntax churning" (Harnad, 1990). Given that agents can be computers as much as humans, with their own norms for behavior – such as protocols – there seems to be no reason why computers, or combinations of computers and humans, cannot create and use new language-games. After all, the moving around of voltage-driven bits by a computer is just as real and meaningful as a human moving their body around and uttering sounds. It is just that what is meaningful for a computer may be meaningless to a human observer! Still, with the Semantic Web, we are hoping to create a language to mediate data integration between various human-created sources of data, and so one criterion of the Semantic Web is that it should *both* be meaningful for computers and humans.

Are URIs somehow different from names in natural language? The answer to this goes back to the notion of the Semantic Web being a game where *new* names can be created primarily for *machines* rather than humans to use. While Wittgenstein himself does not give an adequate treatment of the creation of new language-games, other philosophers like Searle have pursued this line of inquiry. Unlike names in natural language based on what Anscombe termed "brute facts," such as 'trees', 'forests', and 'leaves,' Searle points out that some names exist *only* due to social conventions (Anscombe, 1958; Searle, 1995). The existence of a name – which Searle classifies as one kind of "institutional fact" – only exists in the context of some social phenomenon,

just as the name 'money' and its concrete referents only exist in the context of commercial exchange (1995). Both the name and the actual use of money are not based on any regularities of the physical things, but instead depend on a collective agreement that bestows a certain function upon the use of the name and associated activities of the language-game. There is no reason certain sound-waves or even bits of paper are associated with the linguistic term 'money'. Also, names of institutional facts can refer to classes or kinds of things: There is nothing in the *realization* of money, such as a piece of thin paper, that would necessitate it being an all-purpose-mechanism to indicate value; it is precisely this fluidity of encodings that allows money to have manifold realizations from encodings in stock-market databases to bars of gold. This agreed upon purpose of a new name and its referent in a language gives the name and referent its *status function* (Searle, 1995). In order to convey the status function, the referent of the name can be given some additional physical mark(s), called a *status indicator* that demarcates the special role the realization is playing in some language-game, such as a seal and writing which were attached to money. Once some community has accepted that particular status function, then the status function impacts on the activity of that community, but "the object is no different...that function is manifested only in actual transactions; hence our interest is not in the object but in the processes and events where the functions are manifested" (Searle, 1995). The agreement on status functions does not have to be *conscious*. We simply use money to exchange commodities and expect other agents to value the nature of our agreement, and are not even overtly conscious of the agreement; the language-game is simply accepted as *given*. However, a name *only* has this status function because agents collectively agree that the name does at some point, and convey the usage of this name in a language-game to others. If people refused to believe that there was a class of institutional facts called money, money itself would return to being worthless paper overnight. One feature of language-games is brought into the clear by institutional facts: most institutional facts only exist due to the existence of other institutional facts and associated activities. For example, the collective agreement that is money comes along with many debts and obligations, such as the agreement that the money can be exchanged for goods in proportion to its value, and it also comes with a cluster of other names, such as 'bank' and 'interest' that it cannot exist without. The same even goes for proper names such as the 'Eiffel Tower,' which exists in a cluster with Paris, France and Gustave Eiffel.

The parallel of URIs with names in natural language for institutional facts should be straightforward. The Semantic Web needs URIs to be accepted as names for things,

in particular entities and concepts that cannot be realized as some encodings transferable over the Web. In order for URIs to be used as names for these kind of things, URIs need at first to be explicitly and collectively agreed upon by a community, and then as more and more applications use these URIs, this usage of URIs as names will unconsciously actually *become* names for things. The status function of these URIs is their use as identifiers for data merger in RDF triples, and the somewhat unsatisfactory status indicator that separates URIs for things not on the Web from other URIs is their use of the `303` or hash convention. The URI by itself is not special, for to someone outside the language-game of the Semantic Web, the URI for the Eiffel Tower itself would just access another web-page about the Eiffel Tower. So the sheer *assumption* of the use of URIs as some sort of universal naming convention is doomed to failure, as there is no reason a URI, which is just a particular character string in of and itself, is a better name than any other string of characters, like Digital Object Identifiers (DOIs) or just names in natural language (Kahn and Wilensky, 2006). The main reason URIs work for names for certain types of information like hypertext web-pages is that they allow *access* to these web-pages. Of course, this advantage can be lost with regards to using URIs as names for things like entities and concepts, so the principles of providing some accessible Web representation should be followed. Any naming convention cannot be taken for granted, but must be established by explicit or implicit agreement in order to boot-strap its use in the wild.[2]

8.1.4 The Representational Nexus

How can new language-games, like the Semantic Web's language-game of URIs, be created? Searle and Wittgenstein offer us no answer. Worse, for the purpose of the Semantic Web, it is important that these URIs be used referentially, yet Wittgenstein appears to completely dismiss notions of reference by stating that "the meaning of a word is its use in the language" (Wittgenstein, 1953). By throwing the problem of reference out of the window, Wittgenstein is actually in good company, with Quine having argued for the "inscrutability of reference" and Chomsky, who despite his heavy

[2] A parallel may be made to Kripke's examples of the causal theory of reference; one reason that Kripke's argument for unambiguous naming has been so successful was because Kripke employed widely accepted famous names such as "Cicero" in his examples, since Kripke rightfully assumed most of his readers were already in the naming-using community of that particular name (1972). For names of not well-known people like 'Kavita Thomas,' the 'famous name' convention of Kripke's examples does not hold. Furthermore, for people there is a clear and legal process of baptism. This is not obviously the case for URIs like `http://www.example.org/EiffelTower`.

leanings towards a stance close to Carnap in logic in his syntactic theory, has claimed that the existence of reference in semantics is questionable (Chomsky, 2000).

Reference has not been banished from the conceptual landscape quite so easily, but can still be saved even in a neo-Wittgensteinian public language position. Having the reference somehow be attached to a name via a causal chain is also not enough, as that supposed 'causal' chain has nothing to do with the meaningful and co-ordinated behavior of agents. However, we can return to the *referential chain* as given in Section 3.6 to construct a theory of reference compatible with a Wittgensteinian notion of meaning. The referential chain maintains some surface similarity with the causal theory of reference, for the stage of *presentation* is similar to Kripke's *baptism* (1972). The main difference is that in the referential chain the stage of *output* corresponds to local behavior that is in part caused by the representation, and the representation *is* a representation precisely because of the fact that the agent's behavior *depends* on some aspect of the representation that was caused by its initial connection to a referent. So, reference no longer is some ephemeral epiphenomenon that should be disposed of, but something that incarnates itself in the meaningful behavior of an agent. This is precisely where the referential chain inspired by Brian Cantwell Smith and the causal theory of reference of Kripke radically diverge. In contrast, Kripke wants the act of reference to somehow hold in all possible worlds, regardless of the meaningful behavior of agents employing the name (Kripke, 1972). Thus, in our interpretation, while all sorts of names in a language can have no referent but have a sense, at least *some* of those things that have a sense can have a referent. It is in this manner that we can establish the priority of meaning and sense over reference yet simultaneously maintain the existence of reference. Both sense and reference must be understood to operate *simultaneously*.

If we are to take this reading of the concept of reference seriously, then there are serious repercussions for the Semantic Web. In particular, it dethrones the notion of any formal knowledge representation language like RDF or OWL being somehow superior to natural language. A representation in a formal language should be put on the same footing as natural language, or even below. If any information whose distal referent has an effect on the meaningful behavior of an agent is to count as representational, then the space of representations on the Web explodes in size to encompass much of the hypertext Web. If the Semantic Web is fundamentally about extending the Web to those things outside the Web, then we have to acknowledge that *most of the current hypertext Web is already representational*. We call *the multitude of representations*

that share a referent the ***representational nexus*** of the referent, a potentially large collection of representations in a variety of formal, natural, and even iconic languages that all share the same referent. For example, if one uses a search engine to look for the 'Eiffel Tower,' one gets a large number of web-pages that are to some extent all *about* the Eiffel Tower by virtue of having some meaningful relationship with it, ranging from pictures of the Eiffel Tower, maps to the Eiffel Tower, and even possibly even videos of the Eiffel Tower. These would all count as representations of the Eiffel Tower, and so would be part of the representational nexus of the Eiffel Tower.

Since sense walks hand-in-hand with our notion of reference, then it can also be said that multiple representations on the Web, both in hypertext and on the Semantic Web, can share the same sense. It is precisely this point that we so laboriously argued in Chapter 3, where we gave an account of the construction of a robust notion of sense on top of information given in multiple and possibly non-natural language encodings. The sprawling representational nexus of a referent, in which almost anything literally counts as a representational by virtue of its causal and historical relationship with at least some referent or another, can then be subdivided and re-factored into senses. Senses are where referents affect behavior of the agents in the language-game. Unlike the definitions of senses as glosses in dictionaries, these senses on the Web exist as information in a vast array of different encodings. In particular, the *same* sense can be shared between a representation of the Eiffel Tower in a formal knowledge representation language like RDF and in a hypertext web-page that is about the Eiffel Tower in natural language. The classic problems of word-sense disambiguation return as problems of URI-sense disambiguation, where the problem is to identify a *cluster* of representations in various encodings that all embody the same sense. We can imagine this problem being especially difficult, for as argued in Section 3.2, the same sense can be interpreted from many different encodings, ranging from multimedia encodings like video to formal languages.

How can we detect the sense of a URI on the Semantic Web, especially if many agents are *not* using URIs as names with definite senses? In this regard, pure empirical observation of the behavior of Semantic Web enabled-agents does not help, as these kind of agents are still academic curiosities and do not crawl an use the Semantic Web in any real sense now. Also, while the Semantic Web may use URIs as names for things not accessible to the Web, a URI that did not allow access to any representations would be an empty move in a private language-game. In a public language game, a URI should access descriptions or depictions of what it refers to in order for other

agents to determine *how* such a URI can meaningful govern their behavior. The critical role of associated descriptions gives us the crucial clue of how to build the new language-game of the Semantic Web: any new language-game must be boot-strapped from *already-existing language-games,* and the primary language-game on the Web is *natural language text*. Since both associated descriptions in some Semantic Web language like RDF and hypertext web-pages can all share the same sense, the question then becomes one of combining natural language text with information on the Semantic Web. Many efforts in automated ontology creation like those of Brewster et al. are already moving in this direction (2007). However, our question is different: given the tremendous number of Semantic Web URIs found in Chapter 7, how can we associate *already existing* Semantic Web URIs with natural language text? Once a Semantic Web URI has been attached to some sense by having it parasite on natural language (and possibly multimedia and the other forms of information), then agents can detect the sense of a URI even in a decentralized environment like the Web.

The most revolutionary concept of Wittgenstein is the *form of life*, and everything else in his philosophy flows from this. The key to understanding the form-of-life is that the meaning of a word is *not* just in other words, but in the entire activity of the agents that share the language that uses the word. If the Semantic Web is to succeed, it must take into account not only natural language, but the real activity of users on the Web, in order to base a new 'language-game' upon this form of life. Currently, the primary approach is to build Semantic Web ontologies direct from the text in web-pages in natural language (Brewster et al., 2007). We should notice that there is a *particular* use of natural language on the Web that is hegemonic: the searching for information by using brief natural language keywords. While this is far from the only use of the Web, it is by far the most dominant, as shown by various studies of user behavior on the Web (Battelle, 2005). This constant and near obsessive use of Web search engines *is* the de-facto cybernetic form of life on the Web. So, any attempt to 'boot-strap' a new language-game for the Semantic Web will have to take into account that the use of natural language keyword-based Web search is *fundamental* for the Web, a point routinely ignored by both the direct reference and the logicist positions. The foundation to boot-strap the use of URIs as names for things on the Semantic Web is on top of hypertext search engine queries and the resulting hypertext web pages.

8.2 Solving the Identity Crisis Through Web Search

At this turning point we descend from an argument in *theory* to the level of *practice*, a move from the philosophy of engineering to philosophical engineering. By almost any possible metric that takes into account real users of the Web, the Semantic Web seems to be a failure, as virtually no Semantic Web applications have been released that have had an impact outside the academic research community. However, we should remind ourselves that the failure of the first generation Semantic Web so far has merely been the failure of the logicist position of Hayes and other formal ontologists, not an underlying failure of the concept of the Semantic Web itself, which is just the extension of URIs to be used as names for things not accessible on the Web. As our empirical analysis of the Semantic Web in Chapter 7 showed, the direct reference position also seems headed to trouble, as it appears that many things will have multiple URIs, with each new data-set creating its own URI.

If "to understand a language is to be the master of a technique," we must make at least a tentative sketch and implementation that demonstrates how the Semantic Web can be a language in the manner proposed by the public language position (Wittgenstein, 1953). The requirement is straightforward: *the system should allow URIs for non-Web accessible things to be easily found with their meaning shared as broadly as possible*. In our exposition of Wittgenstein, we have determined four desiderata for applying the public language position to creating a system:

- Agents should be able to easily discover and re-use Semantic Web URIs.
- Agents should not have to change their behavior in order to utilize these Semantic Web URIs.
- The selection of appropriate Semantic Web URI should take into account the entire representational nexus of the non-information resource.
- Agents should come to some sort of collective agreement about what URIs for non-information resources refer to.

Our proposed system is to put a *hypertext search system into a feedback-loop with Semantic Web URIs*. The system fulfills the four desiderata. First, it would allow users to easily discover Semantic Web URIs by typing in simple natural language query terms. Both the direct reference and logicist position put forward versions of what a URI means as some sort of private language position which hopes to determine what

a Semantic Web URI means without mentioning the information needs of agents. Instead, we shall seek to incorporate the contextualized information needs of agents into the meaning of a Semantic Web URI by matching queries for information to information in the form of associated descriptions accessible via Semantic Web URIs. Thus, if the user wants to discover information about the Eiffel Tower, it would suffice to type in `eiffel tower` as the query terms to discover a Semantic Web URI for the Eiffel Tower and associated information. Such a system would not be a parallel and separate search engine for the Semantic Web, but can be built on top of current hypertext search engines that operate in conjunction with an index of Semantic Web URIs and associated descriptions. Again, if an agent is looking for information on the Eiffel Tower, the agent would go to an existing hypertext search engine and use it. Our system would let the users do that, but then simultaneously run their query against the Semantic Web, in order to discover if there are any URIs with associated information on the Semantic Web about the information need expressed by their query. As shown by Chapter 7, for many queries about non-information resources such as entities and concepts, there is a high likelihood that there is information on the Semantic Web relevant to such information needs. Thus, this system satisfies the first two criteria.

Also as demonstrated by Chapter 7, there is possibly *too much* relevant information in the Semantic Web that could satisfy these queries, and even possibly multiple Semantic Web URIs for a given entity or concept. A large part of the problem may be that the query itself drastically under-determines the sense of the information need. For example, a query for `eiffel` may equally be for the Eiffel Tower or Gustave Eiffel. It would be unlikely that a normal user would be able to sort through masses of RDF data, which to most human agents is indecipherable, even with the aid of special-purpose Semantic Web browsers like the Tabulator (Berners-Lee et al., 2006a). Fulfilling our third requirement, instead of forcing a human agent to change their form-of-life and to somehow adapt to using RDF natively, our system takes advantage of what every user of hypertext web search engines already does: the selecting and browsing of the web-pages returned by the hypertext search engine. If a user chooses one of the hypertext URIs correctly, this can be *implicit* approval that the web-page represents the intended referent of the search terms. Furthermore, these web-pages are part of the same representational nexus as the Semantic Web URI and so share its sense. Our system can then use these as inputs to an algorithm that compares these selected web-pages to the returned associated descriptions from the Semantic Web URIs, so that the retrieved Semantic Web URIs can then be ranked in order, with the Semantic Web URIs and

the associated description that most closely matches the selected web-pages becoming the top returned URIs. This method is known as *relevance feedback* in information retrieval (Rocchio, 1971). So, if a user of our system clicked on a web-page about the Eiffel Tower in Paris, the Semantic Web URI whose associated description that most closely matched that result – the Semantic Web URI about the Eiffel Tower in Paris as opposed to a Semantic Web URI that denotes Gustave Eiffel – would *also* be returned. The system can then take into account the entire vast representational nexus retrieved by the hypertext search engine as well as the various associated descriptions of Semantic Web URIs in order to determine the appropriate Semantic Web URI for a given set of query terms.

If a human-readable associated description is presented in some usable form to the human agent, the agent can quickly determine if the Semantic Web URI is relevant or not. This relevance feedback from the Semantic Web URI can then be fed back into the hypertext search engine, completing a cycle of feedback. As more and more users use the system, the amount of selected web-pages will increase, and this information can then be used to choose a URI that has an associated description that carries as much of this information as possible. As multiple users use the search-based system, each of them can be considered to have 'voted' on a particular Semantic Web URI via their selection of hypertext web-pages, and the Semantic Web URIs that are collectively chosen rise to the top. So our system takes advantage of the ordinary 'wisdom-of-crowds' of human agents searching the Web in order to reach collective agreement about what the Semantic Web URIs refer to and what they mean. It is this extension of our system that fulfills the fourth requirement of the public language position.

8.3 Justification of System

In a broad stroke, we have reduced the Identity Crisis to be fundamentally an information retrieval problem. We will call this the ***Semantic Search*** paradigm: *the attempt to retrieve Semantic Web URIs and possibly associated descriptions* in response to query words, in order to refer to our particular information retrieval problem. To phrase this paradigm formally, given a query Q, we wish to maximize the likelihood of relevant Semantic Web URIs (U) being retrieved from the Semantic Web. To do this, we will use the URI's associated descriptions, or D. For the particular use-case of the Semantic Web, it would be best to have a single 'best' URI u returned in response to a query Q. However, given the large number of URIs that could be returned in response to a

query as observed in Chapter 7, it seems that it is better to assume that more than a single URI will be retrieved. Due to the foundational *Probability Ranking Principle*, the order in which to rank the documents is by their estimated probability of relevance with respect to the query. As stated by van Rijsbergen, "if a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, the overall ranking will be the best that is obtainable on the basis of that data" (van Rijsbergen, 1979). Given the scenario where the system in penalized if it returns a non-relevant document, then the Probability Ranking is optimal, since it minimizes expected loss. This has been formally proven (Ripley, 1996), although the proof requires that the probabilities for every document D and query Q as well as relevance values are known. Since the Probability Ranking Principle is optimal, it should return the most optimal URI u for the query Q in the first position of the ranking. In this way, the ad-hoc information retrieval paradigm used by Web search engines solves the Semantic Search problem of finding the 'best' URI for a given query in the information retrieval paradigm without any major changes to the general paradigm.

However, one large problem with information retrieval systems lies in the query itself. As observed in our query log in Chapter 7, the average query length is barely two words. This is a result of Belkin's Anomalous State of Knowledge (ASK) hypothesis, namely that "an information need arises from a recognized anomaly in the user's state of knowledge...and, in general, the user is unable to specify precisely what is needed to resolve that anomaly" (1982). Since the agent does not know precisely what information they lack, they have trouble phrasing accurate keywords in natural language to describe the information. This problem is ameliorated somewhat in the Semantic Search paradigm, as the user is generally aware of the natural language name of what entity or concept for which they are seeking a URI. However, even in Semantic Search, the ASK hypothesis still holds, as often the natural language name of the entity or concept is ambiguous by itself. Furthermore, if our system is using as its criteria for the 'best' URI the associated description with the most relevant and complete information, then the retrieved Semantic Web URIs, even if they all refer to the same thing as the query, can still have substantial differences in terms of the 'goodness-of-fit' to a query due to differences in associated descriptions.

In order to deal with these problems, we will employ ***relevance feedback***, the *use*

of explicit relevance judgments from users of a query in order to expand the query. By 'expand the query,' we mean that the usually rather short query is expanded into a much larger query by adding words from the known relevant documents. The hypothesis of relevance feedback, as pioneered by Rocchio in the SMART retrieval system, is that the relevant documents will disambiguate and in general give a better description of the information need of the query than the query itself (1971). This has been shown in general to improve retrieval performance significantly, both in early studies and in later work (Lavrenko et al., 2002).

Our novel solution to the Semantic Search problem is to use hypertext web-pages that share the same sense of the query as the URIs. These can then be retrieved by running the query Q against a normal hypertext Web search engine. Another question is how to get the associated descriptions D, which can then be built on top of current indexes or *Semantic Search* engines like FALCON-S built on top of the Semantic Web (Cheng et al., 2008).

Indeed, one problem that is beyond the scope of this thesis is the general information retrieval problem of building either a better search engine for either RDF or hypertext. Instead, our system is built on top of current hypertext and Semantic Web search engines. The insight of our system is that search engines for *both* the Semantic Web and the hypertext Web, can be put in what Baeza-Yates calls a "virtuous cycle" (2008). While Baeza-Yates wishes to use the Semantic Web in order to "effectively make [hypertext Web] search easier," our system does the *reverse*: We use hypertext search in order to make using the Semantic Web easier. Our system shows how the problem of finding URIs for non-information resources can be built on top of existing search infrastructure with *no* modification to the often delicately parametrized basic hypertext and Semantic Web search engines.

8.3.1 Information Retrieval Components

In this section we will establish our vocabulary in terms of information retrieval, used in this chapter and in Chapter 9. We will use this terminology to give an algorithmic description of our system, and then a detailed description of its operational steps. We can consider a hypertext search engine $HypertextSearch$ to be a function from a query Q to a set of web-pages Z, $HypertextSearch(Q) = Z$, where the relevant web-pages $W \subset Z$. In parallel, we can consider a Semantic Web search engine *SemSearch* to be a function from model of the query Q to URIs U, that due to the transitivity of access, can be sub-

stituted by their associated descriptions D so $SemSearch(Q) = D$. Note that we use D both to mean 'associated descriptions' in our system *and* the more general concept of retrieved documents in information retrieval, which for our system are the same. Thus, our system can then be considered a re-ranking 'feedback' system based on query expansion which starting with $Feedback(Q)$, transforms into $HypertextSearch(Q) = Z$ and $SemSearch(Q)) = U$ and use selected relevant documents R to expand the query Q and re-rank D, and since each associated description D_i has an associated URI U_i, this leads to $Feedback(Q) = U$.

8.3.1.1 Models

In order to explain our system, a description of the general information retrieval problem is necessary, along with the vector-space model of Salton with $tf.idf$ term weighting as a guiding example. Given a set of documents (such as associated descriptions or web-pages) D, we can consider all these documents in some native encoding to be transformed to models, often called u_D. The model D_i of each document in the index of the search engine is the *document model*. So D_i is then just the transformation of the raw terms in each document into an m-dimensional term list, where each m is some parameter, which is at most the number of unique terms in the entire collection C. The set of terms in the entire collection is called the *vocabulary* V. Usually m is parametrized to be some smaller amount, such as the top 30 most frequent terms in each document. Thus, $w \in V$ represents a single term, such as 'tower.' These are referred to as *words* since documents are assumed to be in natural language, although for our system w is also automatically extracted from RDF triples. A certain amount of prepossessing can be done on words in the form of stemming or morphological analysis to reduce terms to a common base term, so that 'tower' and 'towers' or 'going' and 'go' map to the same term. If a term in the vocabulary $w \in V$ is not present in the document D, then it will either have a value of zero or some 'weighted' value if smoothing is employed. We will examine different possible values of m in constructing document models for our system.

8.3.1.2 Weighting

The key question in information retrieval is how to 'weight' the value D_w as to fulfill the Probability Ranking Principle. The simplest option is to use the term frequency (tf), where $D_w = n(w,D)$, where $n(w,D)$ is the number of occurrences of w in D

(Salton et al., 1975). However, this technique also performs poorly in practice, as it does not take into account how frequent a term is over all documents D. So tf can be 'inverted' in order to determine how rare a term is over all documents (idf) (Jones, 1972). For example in English, the term 'the' would have a high frequency in a given document d but would also have a high frequency in all indexed documents D, while the term 'Eiffel' would have a high term-frequency in some relevant documents $R \subset D$, but a low frequency overall in D, leading one to suspect that documents in R might be relevant. Mathematically, given a word w in a document, with the frequency normalized over the size m of the document, the term frequency for word w_i (tf_i) is given by $tf_i = \frac{n(w_i,D)}{\sum_{w \in W} n(w,D)}$. The inverse term frequency takes into account *all* documents O where the term w_i occurs once, so that $idf_i = log\frac{|D|}{|O \subset D|}$ and therefore $tf_i.idf_i = tf_i \cdot idf_i$. The weighting scheme used by $tf.idf$ is only one option of many possible weighting schemes, and we will focus more on the highly parametrized and effective $BM25$ when evaluating our system (Robertson et al., 1994), although other forms of weighting such as language modeling will be explored (Ponte and Croft, 1998).

8.3.1.3 Smoothing

The opposite problem of weighting the occurrence of words in a document is also pernicious to information-systems, namely the problem of *smoothing* words in the document and query models (Zhai and Lafferty, 2001). Intuitively, if a word w is missing in D, then $w = 0$. However, in many calculations that require some form of multiplication or division, the presence of a zero in a weight can factor out otherwise relevant weights from other terms in the vocabulary, or lead to errors. The solution of smoothing is just to add a small non-zero factor ε to each $w = 0$, therefore having $w = \varepsilon$. There are a wide variety of possible smoothing techniques, ranging from the simple setting of ε to a constant, to having it be chosen at random from some particular distribution like the Dirichlet distribution. This smoothing function we will consider part of our transformation of the query or document into a model, and we will use the smoothing function most appropriate to each weighting scheme of our system, as usually the appropriate smoothing function is dependent on the weighting function.

8.3.1.4 Comparing Documents to Queries

Having a set of weighted and smoothed document models D does not in of itself produce a ranking, since the ranking is always in relationship to query Q. However, since

our document models are term vectors where each term is from $w \in V$, and each term q in the query Q is also $q \in V$, the query itself can be transformed to a *query model* u_Q. As most queries are only a few words, many systems produce a very sparse term representation, which then have to be weighted and smoothed. Some probabilistic information retrieval models such as 'relevance models' automatically expand the query Q. Since all the models inhabit the same space V, they can be directly compared to each other using a *ranking function*, so that for every $D \in C$, $Comparison(Q,D) = \gamma_D$, where Q and D are transformed into query models u_Q and u_D respectively, while γ_D is the *relevance score* of D for query Q, which is generally smaller the closer D 'matches' Q. The descending order by γ satisfies Robertson's Probability Ranking Principle (1977), and we will not investigate alternate methods of presenting the results, such as clustering-based methods. Various weightings in different frameworks have their own preferred methods of comparison. For example, vector-space models may be compared via cosine distance, while cross-entropy is more appropriate for comparing the distributions resulting from probabilistic weighting schemes.

8.3.1.5 Relevance Feedback

One immediate problem in almost any comparison of the query model and the document models is the sparseness of the query model. There are many different techniques for incorporating relevance feedback, each based on the differing methods for transforming the relevant documents (Z, given by "selecting' ($Select$) relevant Z from web-pages W) into relevant document models R and then combining or creating a new query model Q from this information. For example, for vector space models the well-known Rocchio algorithm attempts to re-calculate the query model vector to match the centroid of the document (1971) to relevance models that 'automatically' expand the query model into a distribution (Lavrenko and Croft, 2001). Thus, for the relevance feedback function $Q_2 = Relevance(R)$ that produces a new expanded query Q_2 given a set of relevant document models R, we will use the precise relevance function most appropriate for the weighting function.

8.3.2 Detailed Description of System

Using the terminology given above, we can unpack the entire system into the following algorithm given by Figure 8.1. Details of every step are given in the next section, and illustrated in Figure 8.2 (placed at end of chapter).

Algorithm 8.3.1: FEEDBACK(Q)

$U \leftarrow SemSearch(Q)$
$D \leftarrow access(U)$
$Z \leftarrow HypertextSearch(Q)$
$R \leftarrow Select(Z)$
$Q_2 \leftarrow Relevance(Q,R)$
for each $D_i \in D$
$\begin{cases} \\ \gamma_D \leftarrow Compare(Q_2, D_i) \end{cases}$
$Present(D)$

Figure 8.1: Feedback-Driven Semantic Search

To go through the diagram one step at a time, in Step 1 the system presents the agent with a text box where the agent can enter a query (Q). In Step 2, the agent formulates the query in terms of natural language keywords, and thus the *FeedbackQ* algorithm begins. In Step 3, a number of URIs are returned by automatically running the query Q against a Semantic Web Search engine that does not incorporate hypertext-based relevance, such that a number of URIs are returned ($U = SemSearch(Q)$). In Step 4, the system accesses each URI $U_i \in U$ and gets a collection of associated descriptions D in RDF. Each of these associated descriptions D_i is indexed by its URI U_i. For Step 5, the exact same query Q is sent to a hypertext Web Search engine ($HypertextSearch(Q)$), which then returns a series of URIs, which are accessed in order to retrieve hypertext web-pages (Z). Since we are not interested in the URIs of the hypertext web-pages, they are not maintained past Step 6. In Step 6, each of the result web-pages (or snippets thereof) is displayed to the agent, and the agent examines (via clicking on or 'choosing' a hypertext web-page) some subset of web-pages $R \subset Z$, and this subset is given to be the relevant web-pages, $R = Select(Z)$. Optionally, if the query has been repeated in the past, the query may be expanded using the previously discovered relevant web-pages and relevant associated descriptions in RDF from previous usage sessions. Some techniques in information retrieval like relevance modeling, would automatically expand the query at this stage, regarding it as merely a sample from a larger language model. In Step 7, every relevant web-page is transformed into a document model. In order to

do this, the web-page is first normalized to Unicode and then stripped of all HTML. Note that a variant of this algorithm *could* use features of HTML, such as whether or not text is in the title, and factor this in the document model using inference networks (Baeza-Yates and Ribeiro-Neto, 1999). Then, the document undergoes organization and stemming to normalize a set of terms from the vocabulary. Once a set of terms in the vocabulary have been established, terms with non-zero counts are also weighted via some weighting function, and terms with a zero-count are given a smoothing function. Relevance feedback takes place in Step 8, where the various relevant document models are factored into the query model. This can take place in a number of ways, such as forming a single document relevance model (R) or considering each of the relevant document models separately. Regardless, the query is expanded into a less sparse query Q_2 via the use of relevance feedback, leading to $Q_2 = Relevance(Q,R)$.

First, each $D_i \in D$ is transformed into a 'pseudo-document,' a reduction of RDF into a 'bag-of-words.' This is done because associated descriptions are composed of RDF triples. Therefore, a number of questions arise about how to create some sort of representation that can be compared to the expanded query model. Obviously, the only challenge is how to deal with URIs. Instead of discarding them or keeping them (which would be equivalent, since they would not be found in V and thus excluded from any D), URIs must be treated as separate words in natural language if they are to be part of a document model. As a cursory glance at some URIs reveals, there is important information in them, due to the propensity of humans to use natural language terms in their Semantic Web URI, called the 'Fido-FIDO' fallacy in philosophy (Ryle, 1949). For example, the URI `http://www.example.org/ArchitectOf` generally denotes an "architect" relationship, such that the triple `ex:EiffelTower ex:hasArchitect ex:Gustave_Eiffel` could be reduced automatically to the pseudo-natural language terms 'Eiffel Tower has architect of Gustave Eiffel.' This allows URIs to be part of V and so compared to Q. The heuristics we employ in Step 11 to reduce URIs to natural language terms are straightforward:

- Reduce to last rightmost hierarchical component.
- If URI contains a fragment identifier (#), consider all characters right of the fragment the last rightmost hierarchical component.
- Remove non-rightmost hierarchical component.
- Tokenize on space, capitalization, and underscore.

So, the URI `http://www.example.org/hasArchitect` would be reduced to two tokens, 'has' and 'architect,' while `http://www.example.org/Gustave_Eiffel` will be reduced to 'Gustave' and 'Eiffel' respectively. Then, in Step 12, each associated description is given its ranking score γ_D via a ranking function, $\gamma_D \leftarrow Compare(Q_2, D)$ that compares the expanded query to the document. In Step 13, these ranked URIs are then arranged in descending order by their ranking score. At this point, the system looks up the topmost D in its index to discover D_i and therefore U_i, or the URI that allowed access to the associated description D in the first place. Note that the index of URIs and associated descriptions keeps track of which URI is used to get the associated description, so that even if the same encoding of an associated description is given by multiple URIs, the associated description D_i can be tracked down to the URI U_i that originally had a causal role to play in the production of its document model. At this point, the URI and its associated description is presented, possibly in a variety of ways including the direct display of meta-data on the search result bar or use of the RDF triples in an application. Optionally the system may in Step 14 determine if an agent examined (or some other program used) the associated description, and add these to a cache of relevant URIs. Also optionally in Step 15 the relevant hypertext web-pages in the form of the relevant document models and even relevant Semantic Web URIs and their associated descriptions can be cached. Finally, in Step 16 the agent may enter another query.

8.3.3 Other Methods

Our system has a number of advantages over other systems, namely in that it does not require the end user to use a specialized language for discovering URIs or navigating Semantic Web data, but instead lets them use a normal Web-search interface with queries in natural language. Furthermore, the disambiguation and discovery of relevant URIs then happens as a side-effect of their normal behavior of examining web-pages. Lastly, this method helps users discover URIs and re-use them, rather than create new ones for each query. The advantages of our system and difference with other approaches are given in this section.

8.3.3.1 URI Co-reference Resolution with RKBExplorer

Another attempt to automate the discovery of co-reference is to create a *Consistent Reference Service* that automatically finds both explicitly declared equivalences with

`owl:sameAs` and inverse functional properties and then calculates their closure (Jaffri et al., 2008). As implemented in *RKBExplorer.com*, the system stores the result of its closure calculations in its own RDF/XML file using a specialized Semantic Web co-reference vocabulary. They recommend that each Linked Data source maintains their own co-reference server, and have demonstrated their system on bibliographic data and WordNet (Glaser et al., 2008). While they claim that unlike OKKAM, a Consistent Reference Service does not simply create 'new' URIs for things, in reality these co-reference bundles are given their own URI and their own associated descriptions, which in turn are indexed by Semantic Web Search engines like Sindice, so inevitably leading to an explosion of new URIs (Glaser et al., 2008). With *RKBExplorer.com*, each Semantic Web URI now is being 'shadowed' by a URI for a co-reference bundle! Also, the Consistent Reference Services only deal with co-reference at the level of formally declared logical co-reference in OWL, and it neglects the very source that lets human agents detect co-reference: the associated descriptions. The primary advantage of our proposed system over Consistent Reference Services is that our system does not create new URIs, but merely brings the likely correct Semantic Web URI to a user's attention, by taking relevance feedback and associated descriptions into account.

8.3.3.2 Semantic Search

There are many commercial companies like *Hakia.com* and *Powerset.com* now offering what they call 'Semantic Search,' although the exact definition of 'Semantic Search' seems to vary, with the common denominator being the use of some knowledge representation to augment information retrieval, such as the use of natural language processing to discover implicit knowledge representations implicit in queries or documents. Another approach, more related to ours, is to try to connect already-existing and explicit knowledge representations. For example, these knowledge representations could be given by explicit mark-up inside hypertext or by discovering complementary knowledge representations to queries. In this vein, the closest system to ours in spirit in the *Microsearch* system (Mika, 2008). This system has been re-deployed commercially by Yahoo! as *Search Monkey*, and takes a similar approach to ours. Microsearch also retrieves hypertext web-pages based on query terms and displays meta-data in a human-usable fashion on the result list, also using Simile (Huynh et al., 2007). Microsearch is similar to our system insofar as it associates hypertext web-pages with Semantic Web information. However, there are two main practical differences between our system and Microsearch. Microsearch does not attempt to determine authoritative

URIs and associated descriptions based on query terms, but only extracts Semantic Web information *already present on the page* in the form of either associated RDFa or a conversion of microformats to RDF in a manner similar to GRDDL (Adida et al., 2008; Suda, 2006; Connolly, 2007). The Semantic Web information is displayed directly parallel to each individual web-page. Therefore, Microsearch is still basing its information retrieval on the level of web-page, as opposed to attempting to discover the best Semantic Web URI and associated description related to the intended referent of the query. Due to this shortcoming, it *only* extracts Semantic Web information itself, and does not run the query in parallel on the Linked Data Web. So, Microsearch does not attempt to find the best Semantic Web URI that matches the intended referent of the query, and thus does not help resolve the Identity Crisis by encouraging URI re-usage.

Worse still, Microsearch extracts *all* structured data from the web-page, without any regard for the similarity of the query terms. This Semantic Web information (in particular, information related to time and people) is aggregated and displayed in a 'box' near the search results. While this approach seems to work for relatively simple queries about people who only have only a small amount of Semantic Web information about them on the Web, for queries like 'The Eiffel Tower' too much is brought up, and information about events at the Eiffel Tower and movies like 'The Plot to Blow Up the Eiffel Tower' are mixed, leading to a bewildering agglomeration of structured data displayed to the user. Lastly, no relevance feedback is taken into account to refine this Semantic Web data. Instead of pursuing a synchronous relationship with the Semantic Web, the Yahoo! research team behind Microsearch has now moved their focus to the more difficult problem of large-scale Semantic Web information extraction from text, with all the problems that entails, including excessive URI creation (Baeza-Yates et al., 2008).

8.3.3.3 Ontology Creation from Text

Our system is *not* attempting to do information extraction over the Web representations in order to present the users just the relevant web-pages or extracted 'answers,' as is traditional in information extraction frameworks and question-answering systems (Etzioni et al., 2004; Kwok et al., 2001). Unlike question-answering systems, we are not attempting to answer a query for *specific* information, but only to find URIs with appropriate associated descriptions for non-information resources, rather than return specific 'answers' encoded in natural language. Furthermore, we are not employing

any techniques to transform natural language text directly into ontologies for later use, such as the formal-concept analysis method put forward by Cimiano et al. or the dynamic iterative method using knowledge extraction patterns put forward by Brewster et al. (2007; 2005). This problem of learning ontologies is by itself very difficult and outside the scope of this thesis. These text-to-ontology methodologies seek to ground RDF triples in individual natural language sentences or phrases, which could be considered a *microscopic* approach to associating text with RDF, usually with *new* RDF triples generated directly from that text. Instead, we are pursuing a *macroscopic* approach that grounds *already-existing* RDF triples in associated descriptions with entire collections of web-pages. The information in these already-existing associated descriptions may overlap with the text in the web-page, and a central hypothesis of our system is that this will indeed be the case, but we do not attempt to release this information in some *intermediate* form onto the Semantic Web directly. The approach to generating Semantic Web ontologies directly from text was necessitated by the first-generation Semantic Web's lack of usable data. The second-generation Web of Linked Data has the problem is the reverse: There is too much Semantic Web information for a given query! However, it is likely there are many queries for non-information resources for which no relevant Semantic Web URI exists, and in this particular realm ontology construction from text will be vital. This particular problem of discovering queries that have no relevant URIs and then creating new URIs is beyond the scope of this thesis, but is potentially exciting future work.

8.3.3.4 Ontology Alignment

One opposing methodology for the URI re-use problem is some form of *ontology alignment*. In Semantic Web ontology alignment, the various terms in two or more different languages are 'matched' together with other terms that have the same content, for example, matching 'Eiffel Tower' to 'Tour Eiffel.' There is a long history of ontology alignment or 'mapping' research in knowledge representation, and the advent of the Semantic Web has led to a revival of these techniques (Euzenat and Shvaiko, 2007). Ontology alignment employs a number of distinct heuristics, ranging from the syntactic manipulation of the knowledge representation language to methods based on detecting high-level formal semantic similarities (Bundy et al., 2006; Shvaiko and Euzenat, 2005). The advantage of ontology alignment is that it supposedly allows the users of the program to maintain their "semantic autonomy," so as to maintain their own irreducibly unique perspective on the world while still mapping their terms to the

terms of other agents (Zurawski et al., 2008). As attractive as it appears, ontology alignment has proven itself not to work in practice. While on certain selected ontologies, a particular method may claim to get high recall and precision, so far whenever an ontology alignment system is ported to a new domain the method fails to produce an acceptable level of performance (at best approximately 50% recall and 60% precision) and having unacceptable runtimes, ranging from a few minutes to hours (Caracciolo et al., 2008). This has led the general practice within the Semantic Web community to rely on manually created ontology alignments. Some of this may no doubt be due to irreconcilable social distinctions between certain concepts, such as whether 'marriage' has a constraint of one man and one woman (Ginsberg, 2006). Merging all ontologies that mention a term in order to produce an 'ideal' associated description would easily lead to ontologies with an excess of spurious and inconsistent information.

More importantly, ontology alignment may be criticized as simply the wrong technique for the Semantic Web, trying to solve a problem that would otherwise not exist if the correct technical infrastructure were created and URIs could be easily found in the first place. From a philosophical perspective, most of the ontologies created on the Semantic Web are created by lone individuals and often not re-used by anyone else, so mapping between them is the equivalent of mapping between private language games instead of creating a new public language game. The entire point of the Semantic Web is to create URIs for common concepts and physical entities, and only if the URIs are re-used can graph merger take place. Until very recently, with the advent of Semantic Web search engines for ontologies like Swoogle, it was impossible to even find already-created ontologies, thus leading users with no other recourse than to create their own ontologies. One would suspect that once users have an ability to find URIs, they would not have mint new URIs, but instead re-use URIs, much as names are re-used in natural language and code re-used in open source projects. So our system attempts to solve the very problem that creates the need for ontology alignment on the Semantic Web in the first place.

8.3.3.5 Sense Disambiguation

As would naturally follow from the public language position, URI disambiguation is an analogue with word-sense disambiguation, where instead of associating a number of sentences with a distinct sense, we are associating a number of hypertext web-pages with a distinct URI. Furthermore, multiple co-referential Semantic Web URIs can be considered a class of URIs that share the same sense. It is unclear how well humans

can actually label senses, although performance of word-sense disambiguation systems seems to be reasonable (Kilgarriff, 1993). Despite these difficulties, algorithms that make the assumptions that collocated words can discriminate between senses and that a single discourse uses a single sense have been able to produce very accurate results (Yarowsky, 1995). Furthermore, automated word-sense disambiguation techniques have been shown to work over a substantial number of senses gathered from different sources and a large number of texts (Stevenson and Wilks, 1999). Even so, there are a few practical issues with the notion of word-sense. Unlike part-of-speech tagging, there is no clearly delimited set of word-senses, although in practice both finitely-bounded machine-readable dictionaries and manually-created lexical resources like WordNet tend to be used (Miller, 1995). However, in an open-ended domain like text on the Web, the number of senses becomes even more noticeably open-ended, such that word-sense disambiguation becomes difficult yet again (Stevenson and Wilks, 1999). Worse, even natural language is continually evolving, with new senses being introduced, old senses disappearing from use, and senses drifting over time. This is especially noticeable in the world of Web queries as explored in Chapter 7, as these are driven by fashion and current events.

We take inspiration from the successful statistical work on word-sense disambiguation by transforming 'one sense per discourse' into 'one sense per query.' Our algorithm *adds* additional context by associating hypertext web-pages – which are generally more rich in information than Semantic Web documents – with URIs, and then assumes the user, since they are searching for information about a particular sense, will automatically click on web-pages that give a single sense (Yarowsky, 1995). Our algorithm then can be said to determine sense on the *document*-level as opposed to *word*-level. A critic could respond that if the query terms were ambiguous, the ambiguity would be passed on to the search results, which would then be a mixture of web-pages about different referents. If the user meant a little-known sense of a query term, perhaps all the high-ranked search results would refer to another more prominent sense. These criticisms would be true if we did not rely on the user-behavior of accessing URIs to determine a subset of web-pages in the search results that are actually about what the user considers to be the same referent. By manually examining the web-pages, the user sorts this out, so for our purposes, relevance feedback serves as the primary source of URI disambiguation.

8.4 Conclusion

It could be considered ironic that in our system logical knowledge representations are explicitly transformed into a 'bag of natural language words' in order to allow agents to actually discover and use these knowledge representations. The feature of some theories of meaning is that some entity like the dog Fido was represented by some symbol called `FIDO` (Ryle, 1949). This was considered by Ryle to be a defect, and the very label 'Fido-FIDO' was invented as a derogatory term by Ryle to insult theories of meaning such as Carnap's *Meaning and Necessity* that made such a move (Ryle, 1949).

While Ryle was right to point out the ridiculous nature of the 'Fido-FIDO' principle in theories of meaning, the 'Fido-FIDO' principle *also* describes perfectly the common practice of using natural language terms in knowledge representation systems (Wilks, 2008a). This principle returns to the Semantic Web as a crucial advantage for our system! While the 'Fido-FIDO' pattern of URIs breaks the principle of URI Opacity,[3] it crucially allows knowledge representation languages to be put on the same footing as both user queries and web-pages. Once this move of transforming knowledge representation to natural language form is accepted, then the highly optimized methods of information retrieval can be applied to the Identity Crisis. In particular, this move allows the crucial notion of relevance, reformulated for the Semantic Web in terms of being an *accurate* representation of the intended referent of a query, to be applied to the Semantic Web. Then we can take advantage of the vast representational nexus of the hypertext Web to 'boot-strap' through ordinary user behavior a philosophically well-founded notion of URI meaning on the Semantic Web, and so provide a practical application of the Wittgensteinian public language position.

From a purely pragmatic standpoint, given the historical shipwreck of classical artificial intelligence, it may make more sense for the Semantic Web to harness its fortune to the phenomenal success of information retrieval rather than knowledge representation. Yet one could argue that our system's rather ruthless taking advantage of the 'Fido-FIDO' phenomenon on the Semantic Web is purely an artifact of our algorithm, and that the connection from Wittgenstein to Web search engines is far from philosophically well-grounded. On the contrary, the discipline of information retrieval is *directly* descended from Wittgenstein himself via the under-appreciated philosopher

[3] Perhaps it is better termed the 'Fido-`http://www.example.org/FIDO`' theory of meaning on the Semantic Web.

and linguist Margaret Masterman. One of the six students of Wittgenstein's course that became *The Blue Book*, she was exposed directly by Wittgenstein to the conceptual apparatus of the *Philosophical Investigations* (Sowa, 2006). Twenty years later, she founded the Cambridge Language Research Unit. Overall, Masterman was convinced that a scientific theory of computational language based on are a neo-Wittgensteinian 'semantics' could be created, and that this theory could be computational and created from empirical data (Sowa, 2006). As seen by the virtual take-over of artificial intelligence and natural language processing by statistical methods, it is clear that Masterman and Karen Spärck Jones's often implicit neo-Wittgensteinian approach was ahead of their time. Information retrieval, and its data-driven, statistical methodology, *are* neo-Wittgensteinian philosophy of language given computational flesh.

The history of how Wittgenstein, via Masterman, influenced information retrieval and thus search engines is a fascinating trajectory. Wittgenstein's infamous dictum that "meaning is use" seems often itself meaningless upon first glance; how can "meaning is use" possibly be operationalized into a methodology that could form the basis for a science of language (Wittgenstein, 1953)? The answer is obvious: in studying the structure of language empirically, which can be done computationally by the statistical analysis of actual samples of human language. In other words, the building of "language processing programs which had a sound philosophical basis" (Wilks, 2005a). To Masterman, key to this entire effort was the primacy of semantics over syntax, and "the use of a thesaurus as the main vehicle of operations" (Wilks, 2005a). As opposed to the use of logic by Carnap (and later Chomsky) in describing language, Masterman hoped to use lattices and 'fans' to provide a mathematical foundation for the structure of thesauri, a non-logical mathematical theory of language. Her interest in this led to the revision of her colleague Richens's semantic network machine-translation interlingua into a more empirically justified group of open-ended semantic primitives – although this would be an externalized language like any other, not a mere reflection of an internal mental language resembling Fodor's 'Language of Thought' (1975) – that could arise organically and be detected from language use (Wilks, 2005a). This list of semantic primitives and attendant emphasis on the use of semantics in parsing (as opposed to the purely syntactic approach of Chomsky) were first used by Masterman in machine translation (Masterman, 1961), and then influenced heavily any systems in natural language processing, such as the work of Wilks in resolving ambiguities using preference semantics and the work of Schank using conceptual dependency graphs to discover identical sentences regardless of their syntactic form (Schank, 1972; Wilks,

1975). Another student of Braithwaite and Masterman, Yorick Wilks put forward the most explicit linking of Wittgenstein's 'meaning is use' to statistical studies of natural language by noting that for the first time the sheer size of human text on the Web may allow us to quantify the meaning of words as use using statistical techniques such as skip-grams (2008b).

However, our work does not rely only on statistics based in other words to quantify the "meaning is use", but on information retrieval techniques, in particular, relevance feedback. The foundations for information retrieval that we build upon were also influenced by Wittgenstein via another student of Masterman and her husband Richard Braithwaithe, Karen Spärck Jones (Wilks, 2007). Spärck Jones laid the foundations of information retrieval, and even hinted at relevance feedback, in her dissertation *Synonymy and Semantic Classification* (Jones, 1964). Spärck Jones stated that her dissertation proposed "a characterisation of, and a basis for deriving, semantic primitives, i.e. the general concepts under which natural language words and messages are categorized." (1964). She did this by applying the statistical 'Theory of Clumps' of Roger Needham – a theory that was itself one of the first to explicate what Wittgenstein called "family resemblances" – to words themselves, leading her to posit that words could be defined in terms of statistical clumps of other words, a Wittgensteinian insight that contrasts with Needham's more Kripkean attempt to directly connect words to things (Needham, 1962). Also, the first traces of relevance feedback can be found in her thesis, for as noted by Wilks, "these techniques presume that terms which co-occur in documents with query terms are semantically related to query term uses. They rely on the implicit existence of an empirically derived thesaurus, or clump dictionary" (Wilks, 2007). Applying her work over larger and larger sources of data, she later began to abandon using even the open-ended semantic primitives of Masterman. In her later critique of artificial intelligence, she cited that one of the key insights of information retrieval is that programs should take "words as they stand" and not as mere adjuncts to some logical knowledge representation system (1999). In contrast, Wilks points out that statistical techniques from machine-learning have had considerable influence on artificial intelligence, although not via information retrieval, but instead via a general breakdown of disciplinary boundaries in artificial intelligence and the influence of statistics from machine-translation (2005b). In line with the general move towards semantic search we put forward, Wilks maintains that light-weight knowledge representations are becoming increasing crucial to knowledge representation.

It should not be viewed with irony but with a sense of things coming full circle,

that the methods of information retrieval could be considered crucial to the success of the Semantic Web, and even vice versa. Earlier in Section 5.4, we gave an overview of Spärck Jones's critique of both the logicist position of the Semantic Web – which she termed the 'high-end' Semantic Web – and Berners-Lee's direct reference position of the Semantic Web, which Spärck Jones terms the 'middle end' version. She ends up indicating a 'low-end' version that deals with very general 'tags' may work, and while the Semantic Web has not experienced exponential growth, tagging has succeeded (Halpin et al., 2007). However, our work connects something resembling the 'low-end' with a rehabilitated version of the Semantic Web, as our experiment shows that simple queries and statistical information retrieval can be connected to more structured knowledge, and in fact is vital for finding and discovering the quality of such knowledge representations. Indeed, it is clear Spärck Jones's true target is not the Semantic Web as a system of URIs as common names but what she rightly recognized as a logicist approach to reviving classical knowledge representation. In this, she is clearly right, for our system takes advantage of the fact that there is *not* "something better than natural language as a general means of expressing, and hence accessing, information," which is tacitly acknowledged by the presence of natural language terms in URIs (2004). In fact, our system attempts to vindicate a neo-Wittgensteinian public language position primarily by showing that natural language queries work well for describing and finding Semantic Web URIs, and that even the knowledge representations of the Semantic Web ground out in meaningful natural language words that they share with other representations of the same referents on the Web, like web-pages.

The revival of knowledge representation due to the Semantic Web initiative is more of a historical accident than the consequence of any plan, as who but the refugees from the failed knowledge representation projects of classical artificial intelligence would be desperate enough to join in Berners-Lee's efforts to create the Semantic Web? Indeed, there is no objection to the general notion of discovering some sort of open-ended common lexicon of semantic primitives for natural languages, a notion initiated computationally by Masterman (Wilks, 2007). One could simply say that the Semantic Web is the naming of these semantic primitives by URIs rather than abbreviated natural language names, with all the advantages the principles of Web architecture bring. Lastly, RDF triples could then be considered the minimal structure one could attach to these semantic primitives (Masterman, 1961). There is nothing in this URI-based version of the Semantic Web that ties it to any commitment to a single ontology or even single knowledge representation language. The bet of using URIs as a universal nam-

ing scheme for things can just as easily be tied to statistical methods from information retrieval as it can to logic-based knowledge representations. However, as Spärck Jones would remind us, we should proceed next to a test of the system on real users and real data.

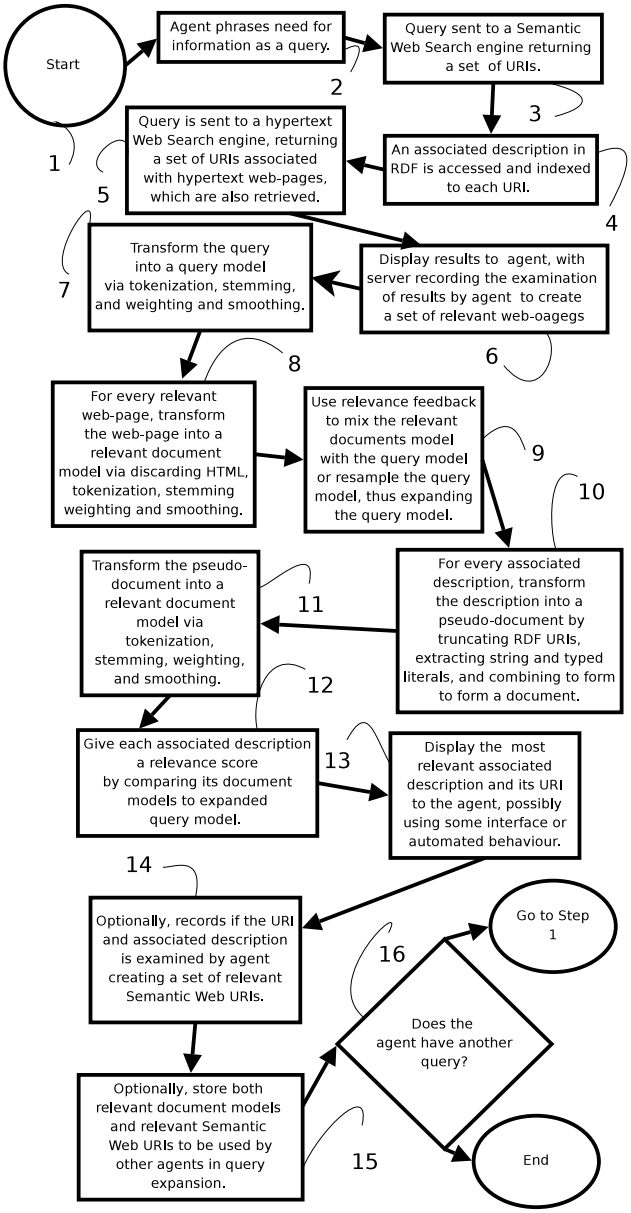

Figure 8.2: Diagram of Feedback-Driven Semantic Search System

Chapter 9

Evaluation

You philosophers ask questions without answers, questions that have to remain unanswered to deserve being called philosophical. According to you, answered questions are only technical matters. That's what they were to begin with. **Jean Lyotard** (1988)

9.1 Experiment

The primary goal of the experiment is to collect what are known as *relevance judgments* of both Semantic Web documents and hypertext web-pages about non-information resources such as concepts and entities, and to determine if these relevance judgments can improve the ranking of the results from search engines operating over both hypertext and Semantic Web information. The criteria for success is that a query in natural language terms to a Semantic Web search engine should return the single best URI for the intended referent of the query. In order to determine if our Wittgenstein-inspired methodology works in practice, an experiment with real human subjects operating over real queries is needed. A random selection of the entire query-driven Semantic Web corpus, as described in Chapter 7, is run against both the hypertext and Semantic Web, and human judges rank both the Semantic Web and hypertext results for relevance. These relevance rankings are then applied to re-rank the results from the Semantic Web and hypertext search engines.

9.1.1 Corpus

For our experimental query corpus, 100 entity queries and 100 concept queries were randomly selected from the crawled URIs from the original corpus for a total experi-

1	ashville north carolina
2	harry potter
3	orlando florida
4	ellis college
5	university of phoenix
6	keith urban
7	carolina
8	el salvador
9	san antonio
10	earl may

Table 9.1: 10 Selected Entity Queries

mental query corpus of 200 queries. Constraints were placed on crawled URIs, such that at least 10 Semantic Web documents were crawled for each query, leading to a total of 1,000 Semantic Web documents about entities and 1,000 Semantic Web documents about concepts, for a total of 2,000 experimental Web representations. Then, the same experimental query corpus was used to crawl the hypertext Web, resulting in a total of 1,000 web-pages about entities and 1,000 web-pages about concepts. The web-pages were retrieved using Yahoo! Search, a commercially deployed hypertext Web search system.[1] While the exact algorithm Yahoo! uses is unknown, it is likely related to PageRank, the original algorithm of their competitor Google, although it is likely both companies have many modifications to the basic PageRank algorithm (Brin and Page, 1998). A random selection of ten queries from the entity corpus is given in Table 9.1 and another random selection of ten queries from the concept corpus is given in Table 9.2. As one can tell, the queries about entities and concepts are spread across quite diverse domains, ranging from entities over locations (El Salvador) and people (both fictional such as Harry Potter and non-fictional such as Earl May) and for concepts over a whole range of abstraction, from sociology to ale.

9.1.2 Defining Relevancy

Since the Web representations were retrieved from search engines, it is entirely possible that the search engine returned irrelevant search results. This is for a number of

[1] Available at `http://www.yahoo.com`.

131	sociology
133	clutch
134	telephone
135	ale
136	pillar
137	sequoia
138	aster
139	bedroom
140	tent
141	cinch

Table 9.2: 10 Selected Concept Queries

reasons, primarily including queries in the natural language that use ambiguous terms and the ability of 'link farms' (web-pages consisting of many links) to manipulate PageRank or other link-based weighing schemes for search engines. For each Web representation, the human judge had to decide whether or not the Web representation was ***relevant*** to the query, where relevance was defined *as whether or not a Web representation is about the same thing as the query, which can be determined if accurate information about the thing is expressed by the Web representation*. By fulfilling these requirements, a particular Web representation can be said to 'satisfy' the information need of a particular user.

Our definition of relevance is considerably stronger than most more informal notions of relevance used in the information retrieval literature (Mizarro, 1997). However, these definitions of relevance are considerably more general-purpose than our notion of relevance because this broader notion of relevance has to deal with not only informational queries, but navigational and transactional queries. Furthermore, our notion of relevance is grounded in the idea of the Web representations actually being *representations* that refer to some sort of entity or concept in the world, and so share the same sense as the referent. Therefore, our definition of relevance encompasses only a subset of all possible informational queries, in particular, those queries where the information is representational. In this manner, we consider the query terms to be descriptions of some referent, where more information is needed by a user about the referent.

A number of types of Web representations that would ordinarily be considered

relevant are therefore excluded. In particular, there is a restriction that the relevant information must be present in the Web representation itself. This excludes possibly relevant information that is accessible via outbound links, even a single link. All manner of Web representations that are collections of links are excluded from relevancy, including both 'link farms' purposely designed to be highly ranked by page-rank based search engines (Brin and Page, 1998), as well as legitimate directories of high-quality links to relevant information. These are excluded precisely because the information, even if it is only a link transversal away, is still not directly present in the retrieved Web representation. By this same principle, Web representations that merely redirect to another resource via some method besides the standardized HTTP `303` method are excluded, since a redirection can be considered a kind of link. They would be considered relevant only if additional information was included in the Web representation besides the redirection itself.

Query terms are astoundingly brief, usually only one or two words, and are so liable to be highly ambiguous, a problem that is unresolvable using statistical natural language processing methods due to there being no context for the query terms besides the query itself. Due to this long-standing problem, there has long been an interest in combining some form of knowledge representation to disambiguate the queries, and recently attempts have been made to use Semantic Web to represent background knowledge (Castells et al., 2007). However, results of disambiguating queries via semantics show that even with some formalized background knowledge, given the vast number of queries possible, it is non-trivial to attach unambiguous semantics to queries reliably, and always more and more queries and relevant documents fall into some 'miscellaneous' category (Lavrenko, 2008).

All hope is not lost. Wittgenstein's emphasis on the form-of-life should remind us that it is not only the linguistic form, but the extra-linguistic activity, that gives meaning to a language. In the case of search terms, the ambiguity can often be resolved by attention to what Web representations have been examined by actual users. In our experiment, a query is considered not only natural-language terms, but also the Web representation clicked on by the user is considered part of the query. Since the queries in the evaluation have been selected from an actual query log from Microsoft *Live.com*, we used a query log to select sample hypertext Web representations that an actual user judged as relevant to the query. If the human judge is in doubt of the intended sense of search terms in the query, then the human judge can use the associated rendered Web representation to determine the intended information need of the query. If the

associated result is itself confusing, the human judges are to assume the most common use of the word in English. If the term is still confusing, the human judge could leave a comment.

The question of what actually defines 'accurate information' is vexing, but can be defined in a satisfactory manner without resorting to any appeal to a heavy-weight logical notion of truth. In a Wittgensteinian manner, the notion of accurate information can be grounded out in the notion of sense, where sense is defined by the use of a term in a language. If a Web representation shares the same sense as the intended referent of the query, then it contains accurate information*about* that referent. However, ascertaining sense is notoriously difficult to do automatically by machine for natural language. However, being proficient at natural language, humans can determine the sense of even limited information. If a Web representation does not contain enough information in it for the human judge to interpret whether or not it shares the same sense as the query, then the Web representation is not relevant. Therefore, many Web representations that merely mention the query terms, but do not provide any information about the referent of the query terms, can be viewed as irrelevant. Given a query for 'Eiffel Tower,' a result entitled 'Monuments in Paris' would likely be relevant if there was information about the Eiffel Tower in the page, but a result entitled 'The Restaurant in the Eiffel Tower' containing only the address and menus of the restaurant would not be relevant.

Following tradition in information retrieval, the human judges are forced to make binary judges of relevance, so each result must be either relevant or irrelevant to the query. Human judges are usually inaccurate when forced to make finely-graded relevance judgments, so users prefer binary relevance judgments (Janes, 1993). Generally, binary relevance judgments have been shown to be statistically stable over time, even if relevance judgments can differ in minor regards both in between judges and in the same judge over time (Baeza-Yates and Ribeiro-Neto, 1999). If the human judge faces any difficulty or has any doubts about their relevance judgment a comment box is given for them to express this difficulty.

9.1.3 Making Relevance Judgments

For each of the 200 experimental queries, 10 hypertext web-pages and 10 Semantic Web documents need to be judged for relevance, leading to a total of 4,000 human judgments for relevance in total for our entire experiment. The human judges each judged 25 queries presented in a randomized order, and were given a total of 3 hours

to test the entire sample for relevancy. No researchers were part of the rating. The judges were each presented first with ten hypertext web-pages and then with ten Semantic Web documents. So for each query, the judge determines relevance for 20 Web representations, leading to a total of 20 judgments per query per judge. Each Web representation therefore judged by three judges, with a total of 30 judges used in the entire experiment. So over a single session, the judges gave judgments to 20 distinct results. The judges were given instructions in line with the definition of relevancy given in Section 9.1.2.

In order to aid the judges, a Web-based interface was created to present the queries and results to the judges. Although an interface that presented the queries and the search interface in a manner similar to search engines was created, human judges preferred an interface that presented them the judgment results one-at-a-time, forcing them to view a rendering of the web-page associated with each URI originally offered by the search engine. For each hypertext web-page, the web-page was rendered using the Firefox Web Browser and PageSaver Pro 2.0. For each Semantic Web document, the result was rendered (i.e. the triples, any associated text in the subject, and any associated depictions) by using the open-source Disco Hyperdata Browser with Firefox.[2] In both cases, the resulting rendering of the Web representation was saved at 469 *x* 631 pixel resolution. The reason that the web-page was rendered instead of a link given directly to the URI is because of the unstable state of the Web, especially the hypertext Web. Even caching the HTML would have risked losing much of the graphic element of the hypertext Web. By creating 'snapshot' renderings, each judge at any given time was guaranteed to be given the same experience in the experiment and to be presented with the web-page in its intended visual form. However, one side-effect of this is that web-pages that heavily depended on non-standardized technologies or plug-ins would not render and were thus presented as blank screen shots to the user.

The judges were each given time to read the instructions as given earlier and were then allowed a test-run on three queries, and these queries were removed from the results. During this training phase, a tutor was allowed to explain why each page was either relevant or irrelevant. Since breaks were allowed for the judges during the judging session, the judges created a login, and were allowed to log-out and re-start the experiment at the beginning of the sub-task they were in. The user-interface broke the evaluation into two steps:

[2]The Disco Hyperdata Browser, a browser that renders Semantic Web data to HTML, is available at `http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/`.

- *Judging relevant results from a hypertext Web search:* The judge was given the query terms created by an actual human user and an example web-page that a user selected, whose full snapshot could be viewed by clicking on it. A full rendering of the retrieved web-page was presented to the user with its title and summary (as produced by Yahoo! Search) easily viewed by the judge as in Figure 9.1. The judge clicked on the check-box if the result was considered relevant. Otherwise, the web-page was by default recorded as not relevant. The web-page results were presented to the judge one at a time, ten times for each query.

- *Judging relevant results from a Semantic Web search:* Next, the judge assessed all the Semantic Web results for relevancy. The judge was given query terms and data from the Semantic Web. A title was displayed by retrieving any literal values from `rdfs:label` properties and a summary by retrieving any literal values from `rdfs:comment` values. Using the same interface as in the judgment of hypertext results, as shown in Figure 9.2, the judge had to determine whether or not the Semantic Web results were relevant.

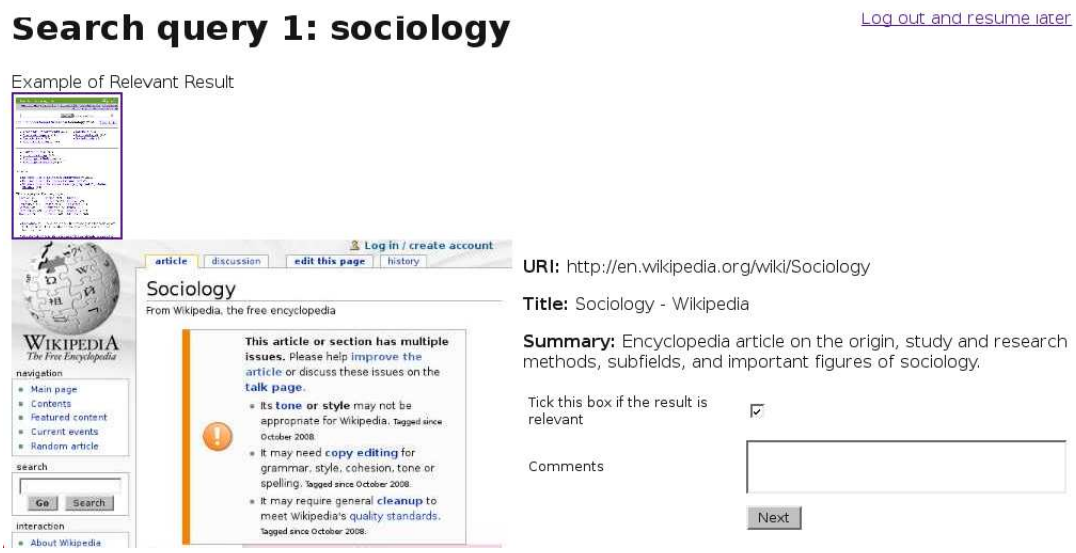

Figure 9.1: The interface used to judge web-page results for relevancy.

After the ratings were completed, Fliess's κ statistic was taken in-order to test the reliability of inter-judge agreement over the relevancy ranking (Fleiss, 1971). Simple percentage agreement is not sufficient, as it does not take into account the likelihood of purely coincidental agreement by the judges: Two judges would naturally have an

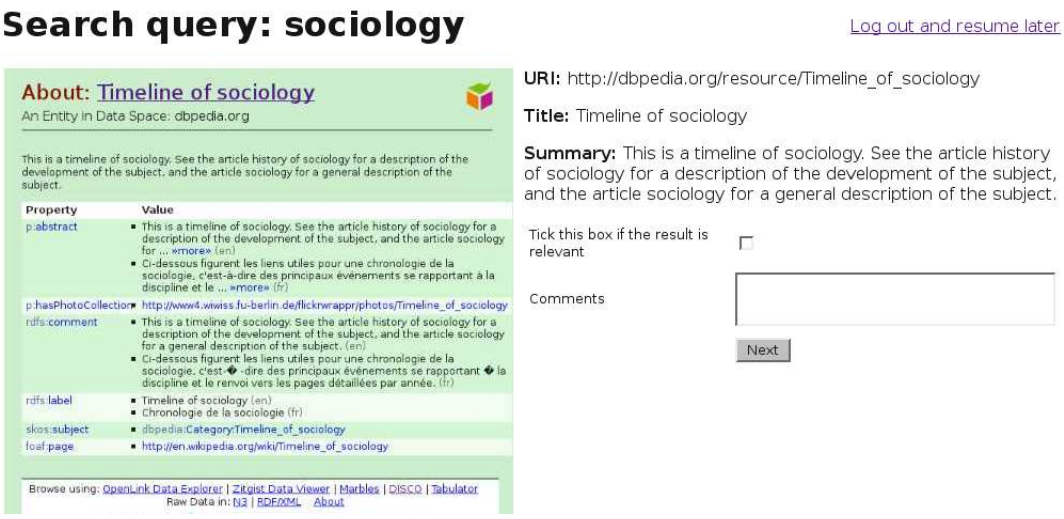

Figure 9.2: The interface used to judge Semantic Web results for relevancy

expected agreement of 50%. While the most common statistic used in assessing inter-judge reliability that corrects for chance agreement is Cohen's κ statistic, Cohen's κ statistic only applies to either two judges per sample or a single judge making two judgments of a single sample (Carletta, 1996). However, the related Fleiss's κ both corrects for chance agreement and can be used for more than two judges (Fleiss, 1971). Fleiss's κ, from here on referred to only as κ, which given O as the observed inter-rater agreement and E as the expected chance agreement between raters, is given in Equation 9.1.

$$\kappa = \frac{O - E}{1 - E} \tag{9.1}$$

The null hypothesis is that the judges cannot distinguish relevant from irrelevant results, and so are judging results randomly. Overall, for both relevance judgments over Semantic Web results and web-page results, $\kappa = 0.5724$ ($p < .05$,95% Confidence interval $[0.5678, 0.5771]$), indicating the rejection of the null hypothesis and moderate agreement. For web-page results only, $\kappa = 0.5216$ ($p < .05$, 95% Confidence interval $[.5150, 0.5282]$), also indicating the rejection of the null hypothesis and moderate agreement. Lastly, for only Semantic Web results, $\kappa = 0.5925$ ($p < .05$, 95% Confidence interval $[0.5859, 0.5991]$), further indicating the null hypothesis is to be rejected and moderate agreement. So, in all cases there is 'moderate' agreement, which is sufficient given the general difficulty of producing perfectly reliable relevancy judgments. Interestingly enough, the difference in κ between the web-page results and Semantic

Web results show that the judges were actually *slightly* more reliable in their relevancy judgments of information from the Semantic Web rather than the hypertext Web. This is likely due to the more widely varying nature of the hypertext results as compared to the more consistent informational nature of Semantic Web results.

Were judges more reliable with entities or concepts? Recalculating the κ for all entity results, $\kappa = 0.5989$ ($p < .05$, 95% Confidence interval $[0.5923, 0.6055]$), while for all results based on concept queries was $\kappa = 0.5447$ ($p < .05$, 95% Confidence interval $[0.5381, 0.5512]$). So it appears that judges are slightly more reliable discovering information about entities rather than concepts, backing the claim made by Hayes et al. that there is more agreement in general about 'less' abstract things like people and places rather than abstract concepts (Hayes and Halpin, 2008). However, agreement is still very similar and moderate for both information about entities and concepts.

However, is this disparity in agreement between entities and concepts affected by media type? For content about entities encoded in hypertext, $\kappa = 0.5112$ ($p < .05$, 95% Confidence interval $[0.5019, 0.5205]$), while for information about concepts encoded in hypertext, $\kappa = 0.5271$ ($p < .05$, 95% Confidence interval $[0.5178, 0.5364]$). Taking confidence intervals into account, there is no significant difference in relevance judgments between entities and concepts in hypertext web-page results. However, relevance judgments of entity information encoded for the Semantic Web led to 'substantial' agreement, as shown by $\kappa = 0.6622$ ($p < .05$, 95% Confidence interval $[0.6528, 0.6715]$), while associated descriptions for concepts on the Semantic Web had substantially less agreement on relevance, with $\kappa = 0.5364$ ($p < .05$, 95% Confidence interval $[0.5271, 0.5457]$). As far as reliability is concerned, information about concepts encoded on the Semantic Web is indistinguishable from concept-based information encoded in hypertext, while information about entities coded on the Semantic Web is much more reliably rated for relevance than concepts and even the very same entities encoded in hypertext. Although this seems unusual, upon consideration it makes considerable sense: Agreement on entity-based information may be hindered rather than helped by multimedia and the lack of a structured focus of web-pages, while the more lean and information-rich Semantic Web languages leave less doubt about the primary referent. For example, there may be disagreement among judges about whether a page selling an Earl May jazz album was 'about' Earl May the musician, but the Semantic Web would ideally separate these two things clearly, having distinct representations for Earl May and his music. Also, this is even a stronger validation on the hypothesis that on the Semantic Web, agreement on entities will be higher than abstract concepts

(Hayes and Halpin, 2008).

Given the (at least) moderate agreement across relevance judgments for both webpage and Semantic Web results, pooled voting was used to assess binary relevance scores for each result. For each result, if at least two of the three judges scored the result as relevant to the query, the result itself were considered relevant for the rest of the evaluation. Otherwise, the result was considered to be irrelevant, even if one of the judges found it relevant. After this pooled voting procedure was completed to test for relevancy, a number of statistics can be gleaned from the relevancy judgments. The ***Semantic Web relevancy corpus*** is *the 200 judged queries and 2000 results derived from searching the Semantic Web using FALCON-S* while the ***hypertext relevancy corpus*** is *the 200 judged queries and 2000 results derived from searching the hypertext Web using Yahoo! Web search*. Both the hypertext and Semantic corpus can be combined to create the ***total relevancy corpus***, the *corpus of 400 judged queries and 4000 results*. In the total relevancy corpus each query given is presented twice, so there are only 200 unique queries for the 400 judged results. This was done to allow us to compare the four corpora conditions (Semantic Web, hypertext, entity, and concept) fairly, and each condition had its presentation randomized. However, as we are interested primarily in the differing roles of entity and concept queries on the Semantic Web, we will focus on this condition only in the context of the Semantic Web and not the hypertext Web.

For the queries, much of the data is summarized in Table 9.3. 'Hypertext' means that the result was taken only over the hypertext relevancy corpus and 'Semantic Web' indicates the same for the Semantic Web relevancy corpus. Results for 'Entity (SW)' and 'Concept (SW)' were calculated only over the Semantic Web relevancy corpus and percentages were taken over the results from the Semantic Web relevancy corpus. This is because we are primarily concerned with how entities and concepts differ over the Semantic Web, not the hypertext Web. The percentages for resolved and unresolved for 'hypertext' and 'Semantic Web' were taken over the hypertext and Semantic Web relevancy corpora in order to allow direct comparison of the Semantic Web and hypertext search results. However, the percentages for 'Top Relevant' (a relevant result at the top ranking) and 'Top Non-Relevant' (a non-relevant result at the top ranking) were computed as percentages over all relevant queries, and so excludes unresolved queries. For ease of reference, a pie-chart for the hypertext relevancy corpus is given in Figure 9.3 and for the Semantic Web relevancy corpus in Figure 9.4.

Resolved queries are *queries that return at least one relevant result* in the top 10 results, while ***unresolved*** are *queries that return no relevant queries in the top 10*

Results:	Hypertext	Semantic Web	Entity (SW)	Concepts (SW)
Resolved:	197 (98%)	132 (66%)	70 (53%)	62 (47%)
Unresolved:	3 (2%)	68 (34%)	42 (62%)	26 (38%)
Top Relevant:	121 (61%)	76 (58%)	47 (62%)	29 (38%)
Top Non-Relevant:	76 (39%)	56 (42%)	23 (41%)	33 (59%)

Table 9.3: Results of Hypertext and Semantic Web Relevance Judgments

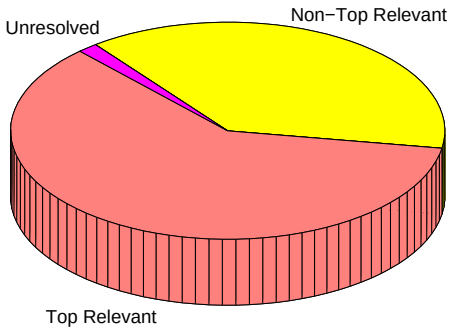

Figure 9.3: Results of Querying the Hypertext Web.

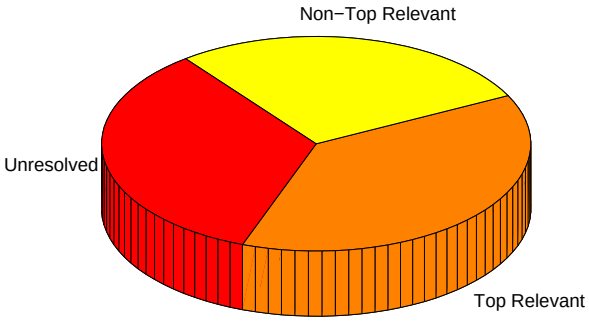

Figure 9.4: Results of Querying the Semantic Web.

results. For the total relevancy corpus, there were 71 (18%) unresolved queries that did not have any results. For the hypertext relevancy corpus, only 3 (2%) queries were unresolved, while 68 (34%) of the queries were unresolved for the Semantic Web. This simply means that the hypertext search engines almost always returned at least one relevant result in the top 10, but for the Semantic Web almost a third of all queries did not return any relevant result. This only means there is much that is still to be represented on the Semantic Web. There was no intersection between those few queries that were unresolved for the hypertext search engine and the numerous queries that did not produce any results on the Semantic Web. Queries that gave the hypertext search engines difficulty were those like 'fable,' since a query for the definition of a 'fable' was over-run by results about a video-game that used the same name. For the Semantic Web, entity queries about specific places with very common names like 'willow ridge' or not-so-well known people like 'monica james' led to no results in the top 10, while concepts like 'doctor' and 'tv' caused problems as well. The reason some concept queries were hard to satisfy was because the Semantic Web simply had information that was too specific for the particular concept, such as information *only* on a few particular television shows in the top 10.

Another endemic problem was the take-over of common conceptual names by popular products (like video-games, novels, or even housing detergents) and companies and music bands. Overall, on the Semantic Web it is far more difficult to locate relevant results about entities than concepts. Of the unresolved queries for the Semantic Web relevancy corpus, there were 47 (58%) entity queries and 33 (42%) unresolved concept queries. Apparently, there are quite a few entities people are interested in, such as the 'Wilson County News,' that do not have a URI yet on the Semantic Web, and so this to some extent validates the OKKAM hypothesis of Bouquet et al. that there are many entities that were still in need of a URI (2007a). However, it appears these entities were only about one-quarter of what users were searching for.

Another question is how many queries had a relevant result as their top result? In general, 197 queries (50%) had top-ranked relevant results for the total relevancy corpus. However, while the hypertext relevancy corpus had 121 (61%) top-ranked relevant results, the Semantic Web relevancy corpus only had 76 (58%) top-ranked relevant results. A lack of top-ranked relevant results becomes particularly acute on the Semantic Web for queries about concepts. For the Semantic Web relevancy corpus, there were 47 (63%) top-ranked relevant queries about entities and only 29 (38%) about concepts. It appears that while search terms often directly take the user to information

about a relevant entity, for concepts this happens less often. This is likely due to there being many concepts whose natural language term is being used as the name of some other entity (such as the term ‘fable’ being used as a company name), and the fact that many concepts are ambiguous and have multiple senses even in natural language.

What makes a more compelling case for relevance feedback is the number of times a *non*-relevant result was the top-ranked (top non-relevant) result in response to a query. For the entire relevancy corpus, there were 132 (33.0%) queries where a non-relevant result was in the top position of the returned results. For the hypertext Web relevance corpus there were 76 (39%) queries with a non-relevant top result, while for the Semantic Web relevance corpus, 56 (42%) of all queries had a non-relevant top result. While queries on the Semantic Web are more likely to turn up no relevant results, when a relevant query is returned, both for the hypertext Web and the Semantic Web it is quite likely that a non-relevant result will be in the top position of the result list. For the Semantic Web top non-relevant results, 23 (41%) of the queries about entities had a top non-relevant result, while there were 33 (59%) queries about concepts that had a top non-relevant result. In particular, this means that concepts were overall more likely to have a top non-relevant result in response to a query, in line with our earlier insights about the different behavior of concepts and entities on the Semantic Web.

Excluding unresolved queries, there is an average of 3.97 (S.D. 2.14) relevant results per query in the hypertext Web relevancy corpus and an average of 1.93 (S.D. 2.2) relevant results per query for the Semantic Web relevancy corpus. While having more than one relevant result in the top 10 for a hypertext search engine is an advantage, having more than one co-referential URI on the Semantic Web is a problem, and with most queries producing about two relevant URIs seems to support the hypothesis that on the Linked Data Web, multiple people are actually producing multiple URIs for the same thing. There were 80 queries that had more than one relevant result, with an average of 3.36 (S.D. 2.14) relevant results per query. With regards to differences between entities and concepts, there were substantial differences. From the 80 queries with more than one relevant result in the Semantic Web relevance corpus, entity queries have an average of 2.79 (S.D. 1.59) relevant results, while concept queries have an average of 4 (S.D. 2.50) relevant results. This means that abstract concepts on the Semantic Web often have *many* shared URIs, while in the case of an entity being mentioned on the Semantic Web, it usually has two URIs. From inspection of entities with many relevant results, it appears the usual case is that DBpedia and WordNet have a substantial amount of overlap in the concepts to which they give URIs. For example, they have dis-

tinct URIs for such concepts as ‘violin’ (`http://dbpedia.org/resource/Violin` vs. `http://www.w3.org/2006/03/wn/wn20/instances/synset-violin-noun-1`). Likewise, most repetition of entity URIs comes from WordNet and DBpedia, both of which have distinct URIs for famous people like ‘Charles Darwin’ (`http://dbpedia.org/resource/Charles_Darwin` and `http://www.w3.org/2006/03/wn/wn20/instances/synset-Darwin-noun-1`).

How is a user supposed to choose between equally authoritative URIs from W3C WordNet or DBpedia? Our information-retrieval based system discovers which Semantic Web URI better ‘matches’ the information in the relevant hypertext web-pages.

9.2 Information Retrieval Framework

In our experiment we tested two general kinds of information retrieval frameworks: vector-space models and language models. In the *vector-space model*, document models are considered to be vectors of terms (usually called ‘words’ as they are usually, although not exclusively, from natural language) where the weighing function and query expansion has no principled basis besides empirical results. Ranking is usually done via a comparison using the cosine distance, a natural comparison metric between vectors. The key to success with vector-space models tends to be the tuning of the parameters of their weighing function. While fine-turning these parameters has led to much practical success in information retrieval, the parameters have little formally-proven basis but are instead based on common-sense heuristics like document length and average document length.

Another approach, the *language model* approach, takes a formally principled and probabilistic approach to determining the ranking and weighting function. Instead of each document being considered some parametrized word-frequency vector, the documents are each considered to be samples from an underlying probabilistic language model M_D, of which D itself is only a single observation. In this manner, the query Q can itself also be considered a sample from a language model. In early language modeling efforts (Ponte and Croft, 1998), the probability that the language model of a document would generate the query was the comparison function of the document. A more sophisticated approach to language models considers that the query was a sample from an underlying *relevance model* of unknown relevant documents, but that the model could be estimated by computing the co-occurrence of the query terms with every term in the vocabulary. In this way, the query itself was just considered a lim-

ited sample, so the it is automatically expanded before the search has even begun by re-sampling the underlying relevance model.

In detail, we will now inspect the various weighting and ranking functions of the two frameworks. A number of different options for the parameters of each weighting function, and the appropriate ranking function, will be considered.

9.2.1 Vector Space Models

9.2.1.1 Representation

Each vector-space model has as a parameter the factor *m*, the maximum *window size*, which is the number of words, ranked in descending order of frequency, that are used in the document models. In other words, the size of the vectors in the vector-space model is *m*. Words with a zero frequency are excluded from the document model.

9.2.1.2 Weighting Function: BM25

The current state of the art weighting function for vector-space models is $BM25$, one of a family of weighting functions explored by Roberson (Robertson et al., 1998) and a descendant of the *tf.idf* weighting scheme pioneered by Spärck Jones and Robertson (Robertson and Spärck Jones, 1976). In particular, we will use a version of *BM25* with the slight performance-enhancing modifications used in the InQuery system (Allan et al., 2000). This weighting scheme has been carefully optimized and routinely shows excellent performance in TREC competitions (Craswell et al., 2005). The InQuery BM25 function assigns the following weight to a word q occurring in a document D:

$$D_q = \frac{n(q,D)}{n(q,D)+0.5+1.5\frac{dl}{avg(dl)}} \frac{\log(0.5+N/df(q))}{\log(1.0+\log N)} \tag{9.2}$$

The $BM25$ weighting function is summed for every term $q \in Q$. For every q, $BM25$ calculates the number of occurrences of a term q from the query in the document D, $n(q,D)$, and then weighs this by the length of document dl of document D in comparison to the average document length $avg(dl)$. This is in essence the equivalent of term frequency in $tf.idf$. The $BM25$ weighting function then takes into account the total number of documents N and the document frequencies $df(q)$ of the query term. This second component is the idf component of classical $tf.idf$.

9.2.1.3 Comparison Function: Cosine and InQuery

The vector-space models have an intuitive comparison function in the form of cosine measurements. In particular, the cosine comparison function is given by Equation 9.3, for a document D with query Q, where both D and Q contain q words, iterating over all words.

$$cos(D,Q) = \frac{D \cdot Q}{|D||Q|} = \frac{\sum_q Q_q D_q}{\sqrt{\sum_q Q_q^2}\sqrt{\sum_q D_q^2}} \tag{9.3}$$

The only question is whether or not the vectors should be normalized to have a Euclidean weight of 1, and whether or not the query terms themselves should be weighted. We investigate both options. The classical cosine is given as *cosine*, which normalizes the vector lengths and then proceeds to weight both the query terms and the vector terms by $BM25$. The version without normalization is called *inquery* after the *InQuery* system (Allan et al., 2000). The *inquery* comparison function is the same as *cosine* except without normalization each word in the query can be considered to have uniform weighing.

9.2.1.4 Relevance: Okapi, LCA, and Ponte

There are quite a few options on how to expand queries in a vector-space model. One popular and straightforward method, first proposed by *Rocchio* (Rocchio, 1971) and at one point used by the *Okapi* system (Robertson et al., 1994), is to expand the query by taking the average of the j total relevant document models R, with a document $D \in R$, and then simply replacing the query Q with the top m words from averaged relevant document models. This process is given by Equation 9.4 and is referred to as $okapi$:

$$okapi(Q) = \frac{1}{j}\sum_{D \in R} D \tag{9.4}$$

Another state of the art query expansion technique is known as *Local Content Analysis* (*lca*) (Xu and Croft, 1996). Given a query Q with query terms $q_1 \ldots q_k$ and a set of results D and a set of relevant documents R, then lca ranks every $w \in V$ by Equation 9.5, where n is the size of the relevant documents R, idf_w is the inverse document frequency of word w, and D_q and D_w are the frequencies of the words w and $q \in Q$ in relevant document $D \in R$.

$$lca(w;Q) = \prod_{q \in Q}\left(0.1 + \frac{1/\log n}{1/idf_w}\log\sum_{r \in R} D_q D_w\right)^{idf_q} \tag{9.5}$$

After each word $w \in V$ has been ranked by *lca*, then the query expanded by LCA is just the top m words given by *lca*. Local Content Analysis attempts to select words from relevant documents to expand the query that have limited ambiguity, and so it does extra processing compared to the *okapi* method that simply averages the most frequent words in the relevant documents. In comparison, Local Content Analysis performs an operation similar in effect to $tf.idf$ on the possibly relevant terms, and so attempting by virtue of weighing to select only words w that both appear frequently with terms in query q but have a low overall frequency (idf_w) in all the results.

The final method we will use is the heuristic method developed by Ponte (1998), which we call *ponte*. Like *lca*, *ponte* ranks each word $w \in V$, but it does so differently. Instead of taking a heuristic-approach like *Okapi* or *LCA*, it takes a probabilistic approach. Given a set of relevant documents $R \in D$, Ponte's approach estimates the probability of each word $w \in V$ being in the relevant document, $P(w|D)$, divided by its overall probability of the word to occur in the results $P(w)$. Then the *Ponte* approach gives each $w \in V$ a score as given in Equation 9.6 and then expands the query by using the m most relevant words as ranked by their scores.

$$Ponte(w;R) = \sum_{D \in R} log\left(\frac{P(w|D)}{P(w)}\right) \tag{9.6}$$

9.2.2 Language Models

9.2.2.1 Representation

Language modeling frameworks in information retrieval represent each document as a language model given by an underlying multinomial probability distribution of word occurrences. Thus, for each word $w \in V$ there is a value that gives how likely an observation of word w is given D, i.e. $P(w|u_D(v))$ (Ponte and Croft, 1998). The document model distribution $u_D(v)$ is then estimated using the parameter λ_D, which allows a linear interpolation that takes into account the background probability of observing w in the entire collection C. This is given in Equation 9.7.

$$u_D(w) = \lambda_D \frac{n(w,D)}{|D|} + (1-\lambda_D)\frac{n(w,C)}{\sum_{v \in V} n(v,C)} \tag{9.7}$$

The parameter λ_D just takes into account the relative likelihood of the word as observed in the given document D compared to the word given the entire collection of documents C. $|D|$ is the total number of words in document D, while $n(w,D)$ is the frequency of word d in document D. Further, $n(w,C)$ is the frequency of occurrence

of the word w in the entire collection C divided by the occurrence of all words v in collection C.

9.2.2.2 Language Modeling Baseline

When no relevance judgments are available, the language modeling approach ranks documents D by the probability that the query Q could be observed during repeated random sampling from the distribution $u_D(\cdot)$. The typical sampling process assumes that words are drawn independently, with replacement, leading to the following retrieval score being assigned to document D:

$$P(Q|D) = \prod_{q \in Q} u_D(Q) \tag{9.8}$$

The ranking function in Equation 9.8 is called *query-likelihood* ranking and is used as a baseline for our language-modeling experiments.

9.2.2.3 Language Models and Relevance Feedback

The classical language-modeling approach to IR does not provide a natural mechanism to perform relevance feedback. However, a popular extension of the approach involves estimating a relevance-based model u_R in addition to the document-based model u_D, and comparing the resulting language models using information-theoretic measures. Estimation of u_D has been described above, so this section will describe two ways of estimating the relevance model u_R, and a way of measuring distance between u_Q and u_D for the purposes of document ranking.

Let $R = r_1 \ldots r_k$ be the set of k relevant documents, identified during the feedback process. One way of constructing a language model of R is to average the document models of each document in the set:

$$u_{R,avg}(w) = \frac{1}{k} \sum_{i=1}^{k} u_{r_i}(w) = \frac{1}{k} \sum_{i=1}^{k} \frac{n(w, r_i)}{|r_i|} \tag{9.9}$$

Here $n(w, r_i)$ is the number of times the word w occurs in the $i'th$ relevant document, and $|r_i|$ is the length of that document. This model is abbreviated as *rm* for relevance model.

Another way to estimate the same distribution would be to *concatenate* all relevant documents into one long string of text, and count word frequencies in that string:

$$u_{R,con}(w) = \frac{\sum_{i=1}^{k} n(w, r_i)}{\sum_{i=1}^{k} |r_i|} \tag{9.10}$$

Here the numerator $\sum_{i=1}^{k} n(w, r_i)$ represents the total number of times the word w occurs in the concatenated string, and the denominator is the length of the concatenated string. The difference between Equations 9.9 and 9.10 is that the former treats every document equally, regardless of its length, whereas the latter favors longer documents (they are not individually penalized by dividing their contributing frequencies $n(w, r_i)$ by their length $|r_i|$). This model is abbreviated as tf from hereon.

9.2.2.4 Comparison Function: Cross Entropy

We now want to re-compute the retrieval score of document D based on the estimated language model of the relevant class u_R. What is needed is a principled way of comparing a relevance model u_R against a document language model u_D. One way of comparing probability that has shown the best performance in empirical information retrieval research (Lavrenko, 2008) is cross entropy. Intuitively, cross entropy is an information-theoretic measure that measures the average number of bits needed to identify the probability of distribution p being generated if p was encoded using given probability distribution p rather than q itself. For the discrete case this is defined as:

$$H(p,q) = -\sum_{x} p(x) log(q(x)) \tag{9.11}$$

If one considers that the $u_R = p$ and that document model distribution $u_D = q$, then the two models can be compared directly using cross-entropy, as shown in Equation 9.12. This use of cross entropy also fulfills the Probability Ranking Principle and so is directly comparable to vector-space ranking via cosine (Lavrenko, 2008).

$$-H(u_R||u_D) = \sum_{w \in V} u_R(w) \log u_D(w) \tag{9.12}$$

Note that either the *averaged* relevance model $u_{R,avg}$ or the *concatenated* relevance model $u_{R,con}$ can be used in Equation 9.12. We refer to the former as *rm* and to the latter as tf in the following experiments.

9.3 Evaluation Metrics

The two most popular measures for determining system performance, *recall* and *precision*, were originally introduced to compare information retrieval systems. Given that 'positive' (R) is a relevant result and every document in the collection C, then $Recall = \frac{|R \cap C|}{|R|}$ and $Precision = \frac{|R \cap C|}{|C|}$. In this way, a search engine with perfect recall

would require that it retrieve *all the relevant results* while a perfectly precise information retrieval engine retrieved only ten relevant results. However, note that recall is not penalized by retrieving both relevant and irrelevant documents, so that perfect recall could be achieved by retrieving *all* documents, relevant and irrelevant, and presenting them to the user. Due to this, precision is usually regarded as the most important statistic, particularly as the Open World Principle states that it is *impossible* for evaluations to categorize *all* relevant results on the Web for a given query. Also, standard notions of recall and precision have no clear cut way of dealing with ranked results in ad-hoc information retrieval, such that a relevant result at the first rank is more important than a relevant result at the last rank. Due to these features of information retrieval systems, the metrics of *mean average precision* and an accompanying significance test known as the *Wilcoxon signed-rank test* have been developed, which are the ones we employ to evaluate our system.

9.3.1 Mean Average Precision

In order to deal with ranked data, precision is modified to be *precision at rank* ρ. Note that this measure takes into account recall as well, as if precision at one rank is greater than precision at another rank, the first rank will *also* have greater recall than the second rank. With our system, given that users only look at the top ten results (Baeza-Yates and Ribeiro-Neto, 1999), we will focus on precision at rank 10 or less.

To give an intuitive example of ranked precision, a quick example is given. Assume our search engine had returned 6 out of 10 relevant results, then the precision at rank 10 would be 0.6. If the first three results were relevant and then only the last three results were relevant, precision at rank 1 would be 1.0, precision at rank 3 would still be 1.0, precision at rank 5 would be 0.6, precision at rank 8 would be 0.5, and precision at rank 10 would return to 0.6. In order to calculate a single evaluation, the precision at each rank with a relevant result can then be averaged by the number of relevant results. So, in our example, $\frac{1.0+1.0+1.0+.5+.56+.6}{6.0}$ results in an average precision of 0.78. However, as information retrieval systems generally need to be evaluated across many different queries, then for each query, the average precision across all queries is averaged, producing the *mean average precision* (MAP), the standard single digit method for evaluating information retrieval systems. When comparing systems over multiple queries, often the term 'mean average precision' is just shortened *average precision*, a convention we shall employ since we do not perform any per-query analysis. When

combined with average precision at various ranks, this provides an overview of a system's performance over a large number of queries. In order to comprehensively test the effectiveness of various parameters, we will test over mean average precision at rank 10, and for the best performing parameters, we will inspect mean average precision at rank 1. This tests the ability of the system to return a relevant Semantic Web URI at the 'top' rank.

9.3.2 Wilcoxon Sign Test

Another problem in evaluating information retrieval systems is evaluating the significance of the results. In particular, standard significant tests like the *t-test* do not apply to information retrieval. First, it is generally thought that the retrieved data is not sampled from a normal distribution. We have shown in Chapter 7 that the amount and kinds of data on the Semantic Web generally follow the non-normal power law distribution. Second, the *t-test* makes the assumption that the underlying scale is an *interval scale*, such that the differences between the rank of each result are some meaningful constant, such that a precision at rank 2 is precisely three times as precise as precision at rank 6. However, it has been found that users value highly ranked results, but not in any absolute manner, so that search engine rankings are better thought of as an *ordinal scale* where the magnitudes of differences do not matter (Baeza-Yates and Ribeiro-Neto, 1999). One test that allows significance testing but only assumes an ordinal scale and does not assume the data has been sampled from a normal distribution is the *Wilcoxon signed rank test*, as given by Equation 9.13 (Baeza-Yates and Ribeiro-Neto, 1999).

$$w = \sum_{i}^{m} R_i \tag{9.13}$$

In this equation, there are m samples to be compared, where each i is a non-zero difference. R_i is then the signed (positive or negative) difference between the two systems. So a system whose parameters gave it a mean average precision of .50 compared to another set of parameters that had a mean average precision of .70 would then have a signed difference of .20, while the reverse comparison would have a signed difference of $-.20$. Once the w has been calculated from a Wilcoxon test, a p-value for rejecting the null hypothesis (that the two sets of parameters were the same) at some significance level can be calculated. We shall use the significance level of $\alpha = 0.05$, and unless explicitly otherwise stated, the Wilcoxon test will always be comparing whatever parameters or results are under scrutiny to the *best* performing parameters. If

the result is the *best* result, then the test is with respect to the *best baseline* parameters. For every group of tests, the best baseline will be explicitly denoted as such.

9.4 Feedback Evaluation

9.4.1 Hypertext to Semantic Web Feedback

9.4.1.1 Results

A number of parameters for our system were evaluated to determine which parameters provide the best results. For each of the parameter combinations, we compared the use of relevance feedback to a baseline system which did not use relevance feedback, yet used the same parameters with the exception of any relevance feedback-related parameters. The baseline system without feedback can also be considered an unsupervised algorithm, while a relevance feedback system can be thought of as a supervised algorithm. For example, the relevant hypertext web-pages R can be considered to be training data, while the Semantic Web data D we wish to re-rank can be considered to be test data. The hypertext web-pages and Semantic Web data are disjoint sets ($D \cap R = \emptyset$). For evaluation we used mean average precision (MAP) with the standard Wilcoxon sign-test, which we will often just call 'average precision.'

For vector-space models, the *okapi*, *lca*, and *ponte* relevance weighting functions were all run, each trying both the *inquery* and *cosine* comparison functions. The primary parameter to be varied was the *window size* (m), the number of top frequency words to be used in the vectors for both the query model and the document models. Baselines for both *cosine* and *inquery* were run with no relevance feedback. The parameter m was varied over $5, 10, 20, 50, 100, 300, 1000, 3000$. The results in terms of mean average precision are given in Figure 9.5.

Interestingly enough, *okapi* relevance feedback weighting with a window size of 100 and an *inquery* comparison was the best, with a mean average precision of 0.8914 ($p < .05$). It outperformed the baseline of *inquery*, which has an average precision of 0.5595 ($p < .05$). Overall, *lca* did not perform as well, often performing below the baseline, although its performance increased as the window size increased, reaching an average precision of 0.6262 with $m = 3000$ ($p < .05$). However, given that a window size of 10,000 covered most documents, increasing the window size will not likely result in better performance from *lca*. The *ponte* relevance feedback performed very well, reaching a maximum MAP 0.8756 with a window size of 300 using *inquery*

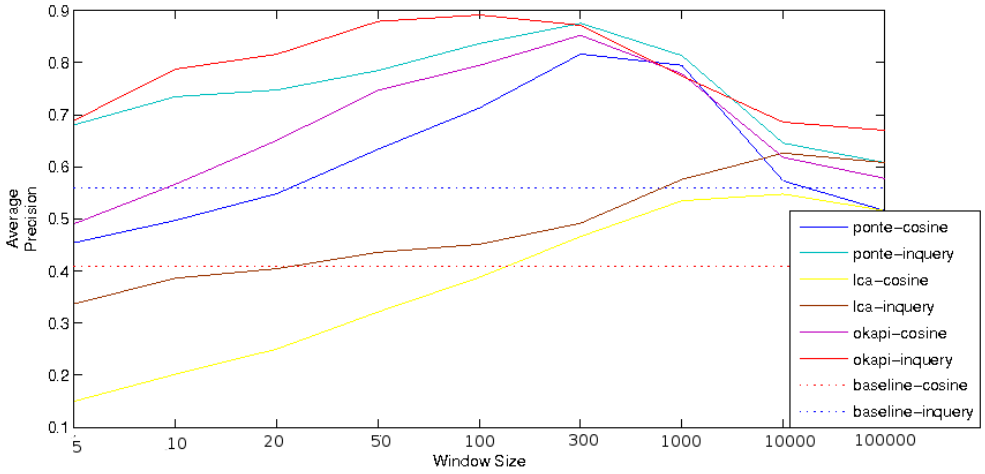

Figure 9.5: Average Precision Scores for Vector-space Model Parameters: Relevance Feedback From Hypertext to Semantic Web

weighing, and so was insignificantly different from *inquery* ($p > .05$). Lastly, both *ponte* and *okapi* experienced a significant decrease in performance as m was increased, so it appears that the window sizes of 300 and 100 are indeed optimal. Also, as regards comparing baselines, *inquery* outperformed *cosine* ($p < .05$).

For language models, both averaged relevance models *rm* and concatenated relevance models tf were investigated, with the primary parameter being m, the number of non-zero probability words used in the relevance model. The parameter m was varied between 100, 300, 1000, 3000,and 10000. Remember that the query model *is* the relevance model for the language model-based frameworks. As is best practice in relevance modeling, the relevance models were not smoothed, but a number of different smoothing parameters for ε were investigated for the cross entropy comparison function, ranging from ε between .01, .1, .2, .5, .8, .9, and 0.99. The results are given in Figure 9.6.

The highest performing language model was tf with a cross-entropy ε of .2 and a m of 10,000, which produced an average precision of 0.8611, which was significantly higher than the language model baseline of 0.5043 ($p < .05$) using again an m of 10,000 for document models and with a cross entropy ε of .99). Rather interestingly, tf always outperformed *rm*, and *rm*'s best performance had a MAP of 0.7223 using an ε of .1 and a m of 10,000.

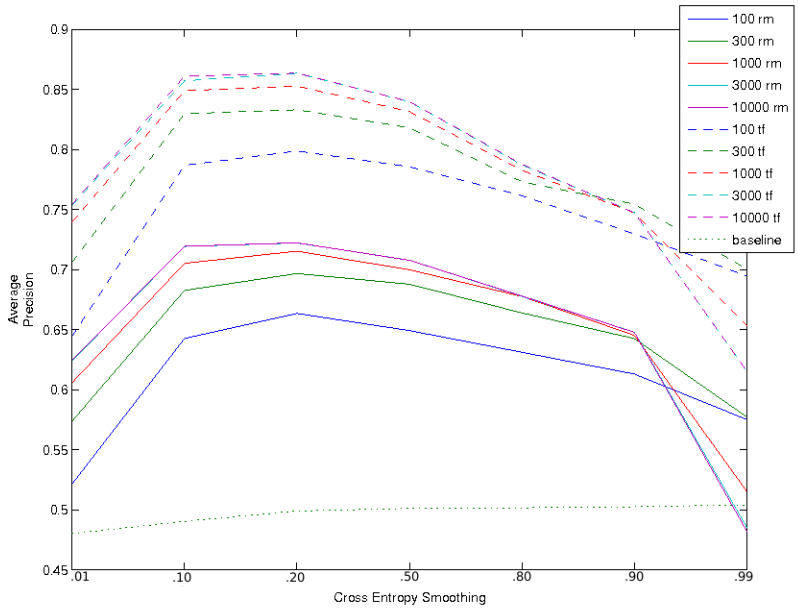

Figure 9.6: Average Precision Scores for Language Model Parameters: Relevance Feedback From Hypertext to Semantic Web

9.4.1.2 Discussion

Of all parameter combinations, the *okapi* relevance feedback works best in combination with a moderate sized word-window ($m = 100$) and with the *inquery* weighting scheme. It should be noted its performance is identical from a statistical standpoint with *ponte*, but as both relevance feedback components are similar and both use *inquery* comparison and *BM*25 weighing, and not surprisingly the algorithms are very similar. Why would *inquery* and *BM*25 be the best performing? The area of optimizing information retrieval is infamously a black art. In fact, *BM*25 and *inquery* combined present the height of heuristic-driven information retrieval algorithms as explored in Robertson and Spärck Jones (1976). While its performance increase over *lca* is well-known and not surprising, it is interesting that *BM*25 and *inquery* perform significantly better than the language model approach.

The answer is rather subtle. Another observation is in order; note that for vector models, *inquery* always outperformed *cosine*, and that for language models tf always outperformed *rm*. Despite the differing frameworks of vector-space models and language models, both *cosine* and *rm* share the common characteristic of normalization. In essence, both *cosine* and *rm* normalize by documents: *cosine* normalizes term fre-

quencies per vector before comparing vectors, while *rm* constructs a relevance model on a per-relevant document basis before creating the average relevance model. In contrast, *inquery* and tf do not normalize: *inquery* compares weighted term frequencies, and tf constructs a relevance model by combining all the relevance documents and then creating the relevance model from the *raw pool* of all relevant document models.

Thus it appears the answer is that any kind of normalization by length of the document hurts performance. The reason for this is likely because the text automatically extracted from hypertext documents is 'messy,' being of low quality and bursty, with highly varying document lengths. As observed in Chapter 7, the amount of triples in Semantic Web documents follow a power-law, so there are wildly varying document lengths of both the relevance model and the document models. Due to these factors, it is unwise to normalize the models, as that will almost certainly dampen the effect of valuable features like crucial keywords (such as 'Paris' and 'tourist' in disambiguating various `eiffel`-related queries).

Then the reason *BM*25-based vector models in particular perform so well is that, due to its heuristics, it is able to effectively keep track of a term's both document frequency and inverse document frequency accurately. Also, unlike most other algorithms, *BM*25 provides a slight amount of rather unprincipled non-linearity in the importance of the various variables (Robertson et al., 2004). This is important, as it provides a way of extenuating the effect of one particular parameter (in our case, likely term frequency and inverse term frequency) and then massively lowering the power of another parameter (in our case, likely the document length). While *BM*25 can be outperformed normally by language models (Lavrenko, 2008) in TREC competitions featuring high-quality samples of English, in the non-normal conditions of comparing natural language and pseudo-natural language terms extracted from structured data in RDF, it is not surprising that *okapi*, whose non-linearity allows certain highly relevant terms to have their frequency 'non-linearly' heightened, provides better results than more principled methods that derive their parameters by regarding the messy RDF and HTML-based corpus as a sample from a general underlying language model.

9.4.2 Semantic Web to Hypertext Feedback

In this section, we assume that the user or agent program has somehow accessed or otherwise examined the associated descriptions from the Semantic Web URIs, and these associated descriptions then form relevance corpus that can then be used as relevance

feedback to expand a query for the hypertext Web. In this way, the feedback cycle has been reversed.

9.4.2.1 Results

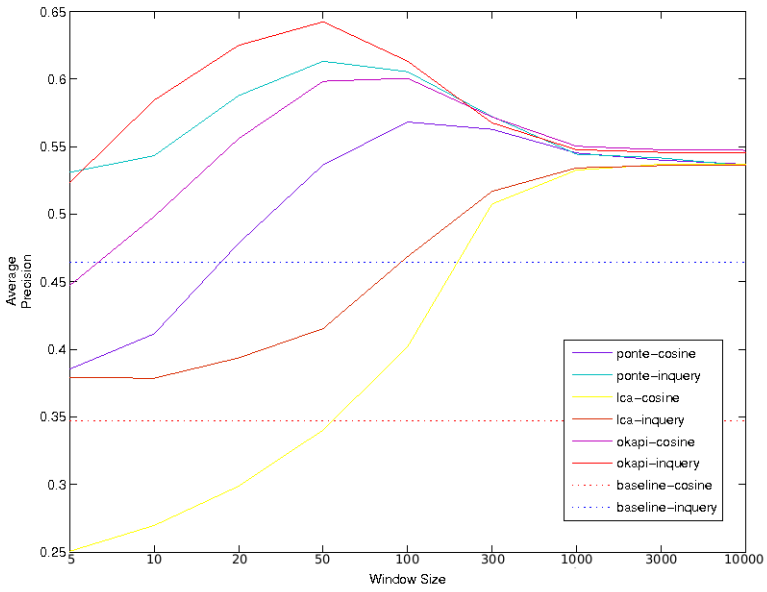

Figure 9.7: Average Precision Scores for Vector-space Model Parameters: Relevance Feedback From Semantic Web to Hypertext

The results for using Semantic Web documents as relevance feedback for hypertext Web search are surprisingly promising. The same parameters as explored in Section 9.4.1.1 were again explored. The average precision results for vector-space models are given in Figure 9.7. The general trends from Section 9.4.1.1 were similar in this new data-set. In particular, *okapi* with a window size of 100 and the *inquery* comparison function again performed best with an average precision of 0.6423 ($p < .05$). Also *ponte* performed almost the same, again an insignificant difference from *okapi*, producing with the same window size of 100 an average precision of 0.6131 ($p > .05$). Utilizing again a large window of 3,000, *lca* had an average precision of 0.5359 ($p < .05$). Similarly, *inquery* consistently outperformed *cosine* in comparison, with *inquery* having a baseline average precision of 0.4643 ($p < .05$) in comparison with the average precision of *cosine* being 0.3470 ($p < .05$).

The results for language modeling were similar to the results in Section 9.4.1.1

and are given in Figure 9.8, although a few differences are worth comment. The best performing language model was *tf* with a *m* of 10,000 and a cross entropy smoothing factor ε to .5, which produced an average precision of .6549 ($p < .05$). In contrast, the best-performing *rm*, with a *m* of 3,000 and ε=.5, only had an average precision of 0.4858 ($p < .05$). The *tf* relevance models consistently performed better than *rm* relevance models ($p < .05$). The baseline for language modeling was also fairly poor with an average performance of 0.4284 ($p < .05$). This was the 'best' baseline using again an *m* of 10,000 for document models and cross entropy smoothing ε of .99. The general trends from the previous experiment then held, except the smoothing factor was more moderate and the difference between *tf* and *rm* was even more pronounced. However, the primary difference worth noting was that best performing *tf* language model outperformed, if barely, the *okapi* (*BM*25 and *inquery*) vector model by a relatively small but still significant margin of .0126. Statistically, the difference was significant ($p < .05$).

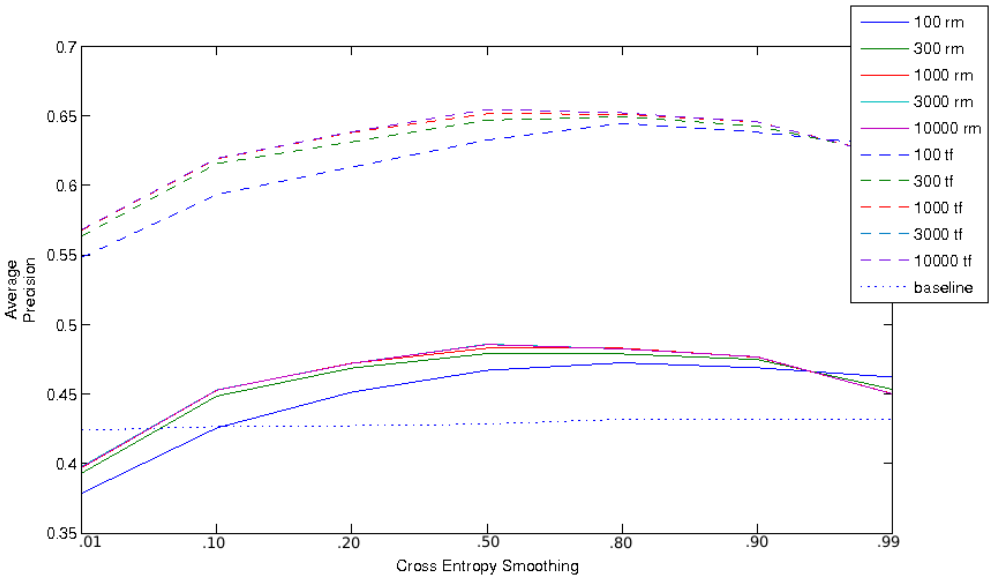

Figure 9.8: Average Precision Scores for Language Model Parameters: Relevance Feedback From Hypertext to Semantic Web

9.4.2.2 Discussion

Why is *tf* relevance modeling better than *BM*25 and *inquery* vector-space models in using relevance feedback from the Semantic Web to hypertext? The high performance

of *BM*25 and *inquery* has already been explained, and that explanation about why document-based normalization leads to worse performance still holds. Yet the rise in performance of tf language models seems odd. However, it makes sense if one considers the nature of the data involved. Recalling Chapter 7, there are two distinct conditions that separated this data-set from the more typical natural language samples as encountered in TREC (Hawking et al., 2000). In the case of using relevant hypertext results as feedback for the Semantic Web, the relevant document model was constructed from a very limited amount of messy hypertext data, which had many text fragments, with a large percentage coming from irrelevant textual data to deal with issues like web-page navigation. This was then compared against Semantic Web data. However, in using the Semantic Web for relevance feedback, these issues are reserved: the relevant document model is constructed out of relatively pristine Semantic Web data and compared against noisy hypertext documents.

Rather shockingly, as the Semantic Web data is mostly manually high-quality curated data from sources like DBpedia, the actual natural language fragments found on the Semantic Web, such as Wikipedia abstracts, are much better samples of natural language than the natural language samples found in hypertext. Furthermore, the distribution of 'natural' language terms extracted from RDF terms (such as 'sub class of' from `rdfs:subClassOf`), while often irregular, will either be repeated very heavily or fall into the sparse long tail. These two conditions can then be dealt with by the generative tf relevance models, since the long tail of automatically generated words from RDF will blend into the long tail of natural language terms, and the probabilistic model can properly 'dampen' without resorting to heuristic-driven non-linearities. Therefore, it is on some level not surprising that even hypertext Web search results can be improved by Semantic Web data, because used in combination with the right relevance feedback parameters, in essence the hypertext search engine is being 'seeded' with high-quality structured and accurate descriptions of the referent of the query to be used for query expansion.

9.4.3 Evaluating Deployed Systems

However, one area we have not explored is how our system performs against state of the art systems. The performance of relevance feedback in Section 9.4.1.1 and Section 9.4.2.1 was only compared to baselines that were versions of our weighting function without a relevance feedback component. While that particular baseline is principled,

the obvious other needed comparison is against actual deployed commercial or academic systems. So we compare the best parameters of the system against actually deployed systems. The obvious baseline to choose to test against is the Semantic Web search engine, FALCON-S, from which we derived our original Semantic Web results used in both the analysis of the Semantic Web in Chapter 7 and in the experiment in Section 9.1. We used the original ranking of the top 10 results given by FALCON-S to calculate its average precision, 0.6985. We then compared both the best baseline, *inquery*, as well as the best (*okapi* with *inquery* and $m = 100$) feedback based system in Figure 9.9. As shown, our feedback based system had significantly ($p < .05$) better average precision (0.8914) than both FALCON-S (0.6985) and the baseline without feedback ($p < .05$).

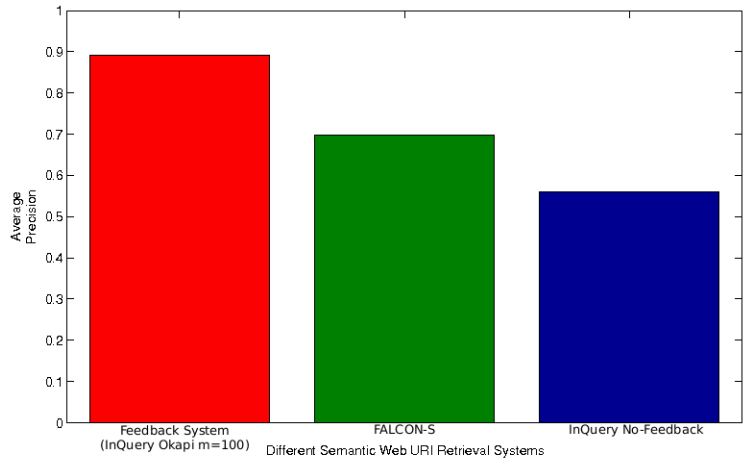

Figure 9.9: Summary of Best Average Precision Scores: Relevance Feedback From Hypertext to Semantic Web

Average precision does not have an intuitive interpretation, besides the simple fact that a system with better average precision will in general deliver more accurate results closer to the top. In particular, one scenario we are interested in is having *only* relevant RDF data accessible from a single URI returned as the top result, so that this result is easily consumed by some program. For example, given the search 'amnesia nightclub ibiza,' a program should be able to consume RDF returned from the Semantic Web to produce with high reliability a single map and opening times for a particular nightclub in Ibiza in the limited screen space of the browser, instead of trying to display structured data for every nightclub called 'amnesia' in the entire world. In Table 9.4, we show that for a significant minority of URIs (42%), FALCON-S returned a

non-relevant Semantic Web URI as the top result ('Non-Relevant Top'). Our feedback system achieves an average precision gain of 20% over FALCON-S in returning a relevant result in the top rank ('Relevant Top'). While a 20% gain in average precision may not seem huge, in reality the effect is quite dramatic, in particular as regards boosting relevant URIs to the top rank. So in Table 9.4, we present results of how our best parameters $okapi - inquery$ with $m = 100$ lead to the most relevant Semantic data in the top result. In particular, notice that now 89% of resolved queries now have relevant data at the top position, as opposed to 58% without feedback. This would result in a noticeable gain in performance for users, which we would argue allows Semantic Web data to be retrieved with high-enough accuracy for actual deployment.

While performance is boosted for both entities and concepts, the main improvement comes from concept queries. Indeed, as concept queries are often one word and often ambiguous, not to mention the case where the name of a concept has been taken over by some company, music band, or product, it should not be surprising that results for concept queries are considerably boosted by relevance feedback. Results for entity queries are also boosted, and are now the most difficult kind of URI for our system to disambiguate. A quick inspection of the results reveals that the entity queries that gave both FALCON-S and our feedback system problems were mainly very difficult queries which have a number of Semantic Web URIs that all share similar natural language content in their associated descriptions. An example would be a query for '`sonny and cher`,' which results in a number of distinct Semantic Web URIs: one for *Cher*, another one for *Sonny and Cher* the band, and another for "The Sonny Side of Cher,' an album by Cher. For concepts, one difficult concept was the query `rock`. Although the system was able to disambiguate the musical sense from the geological sense, there was a large cluster of Semantic Web URIs for rock music, ranging from *Hard Rock* to *Rock Music* to *Alternative Rock*. With a large cluster of URIs with similar content encoded in their associated descriptions, it is not surprising that both our system and FALCON-S had difficulty with certain queries.

Although less impressive than the results for using hypertext web-pages for relevance feedback for the Semantic Web, the feedback cycle from the Semantic Web to hypertext does improves significantly the results of even commercial hypertext web-engines, at least for our set of queries about concepts and entities. The hypertext results for our experiment were given by Yahoo! Web Search (simply called 'Yahoo!'), and we calculated a mean average precision for Yahoo! to be 0.4039. This is slightly less than our baseline *inquery* ranking, which had an average precision of of

Results:	Feedback	FALCON-S
Top Relevant:	118 (89%)	76 (58%)
Top Non-Relevant:	14 (11%)	56 (42%)
Top Non-Relevant Entity:	9 (64%)	23 (41%)
Top Non-Relevant Concept:	5 (36%)	33 (59%)

Table 9.4: Table Comparing Hypertext-based Relevance Feedback and FALCON-S

0.4643. One might wonder why Yahoo! would not use an *inquery* vector-space model to optimize their own system in order to achieve better performance. The reasoning is relatively straightforward: Yahoo! and other commercial search engines must return results within seconds, and doing vector-space comparisons of the results in order to re-rank would take too long. While the exact algorithm behind Yahoo! is unknown, it is likely to be some version of PageRank in combination with a highly-optimized for performance *BM*25. Therefore, the similar precision for Yahoo! and *inquery* make sense. As shown in Figure 9.10, our feedback based system had an average precision of 0.6549 and so performs significantly ($p < .05$) better than Yahoo! and ($p < .05$) the baseline *inquery* system.

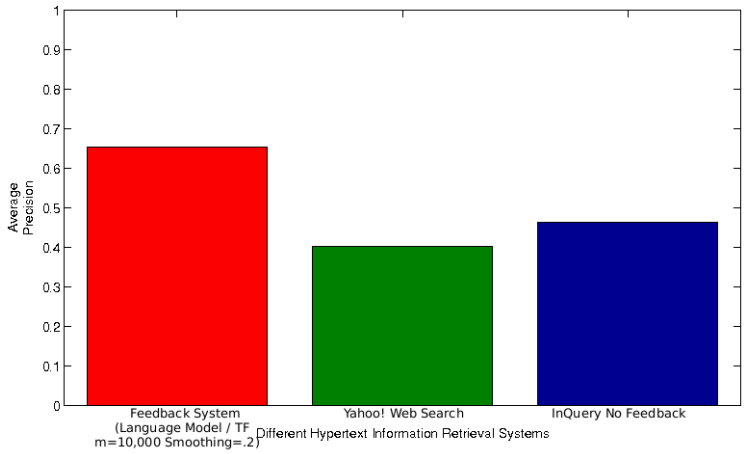

Figure 9.10: Summary of Best Average Precision Scores: Relevance Feedback From Semantic Web to Hypertext

9.5 Discussion

These results are not in need of a large discussion, as they clearly show our relevance feedback method works significantly better than various baselines, both internal baselines and state of the art commercial hypertext search engines and Semantic Web search engines. The parametrization of the precise information retrieval components used in our system is not entirely arbitrary, as argued above in Section 9.4.1.2 and Section 9.4.2.2. The gain of our relevance feedback system, a respectable 19% in average precision over the engine FALCON-S, intuitively makes the ability of our system to place the correct URI in response to a query acceptable for most users. The most difficult step is to select the 'right' Semantic Web URI for the user's need, and in this regard, even small differences can make a huge impact, so an improvement to 89% average precision for a given natural language query makes a large difference.

Second, by incorporating human relevance from the Semantic Web, we make substantial gains over state of the art baseline systems for hypertext Web search. One important factor is the constant assault of hypertext search engines by spammers and others. Given the prevalence of a search engine optimization and spamming industry, it is not surprising that the average precision of even a commercial hypertext engine is not the best, and that it performs less well by a mean average precision of 29% than Semantic Web search engines. Semantic Web search engines have a much smaller and cleaner world of data to deal with than the unruly hypertext Web. Thus, even without relevance feedback from the Semantic Web, an average precision of 69% is impressive, although far from the almost of 89% precision that can be achieved using relevance feedback from the hypertext Web. Improving hypertext Web search is difficult even with relevance feedback. Even with the help of relevance feedback from the Semantic Web, hypertext search is unlikely to achieve near-perfect results anytime soon.

9.6 Conclusion

The final results of our experiment unequivocally demonstrate that our approach of using feedback from hypertext Web search helps users discover relevant Semantic Web URIs and associated descriptions. The gain is significant over both baseline systems without feedback and the state of the art page-rank based mechanism used by FALCON-S and Yahoo! Web search. These results, due to the significant and ran-

domized number of queries used and the fact that relevance judgments involved three judges, point to a high reliability for these results, so we have reason to be believe the results will scale. The operative question is: Why does this work? It is precisely because the same *sense* is encoded in hypertext and the Semantic Web results that these two disparate sets of data be used to aid each other. As distant as it seems, the philosophical work in Chapter 3 on sense and reference laid the ground for improved search performance.

The key reason why we have improved search performance to the point where it should be able to find the 'best' relevant URI for an entity or concept is because we have used relevance feedback for disambiguating concepts and entities. There has been considerable previous research in disambiguating entities on the Web. Some of the work consists of finding common patterns to disambiguate proper names in general from other natural language words, such as the technology we employ to determine the presence of entities in the query log (Mikheev et al., 1998), while further research attempts to link these named entities to their correct sense as given in a list of senses in some knowledge representation (Vu et al., 2007). Current research in entity disambiguation, as exemplified by the approach of Nguyen and Cao, use the previous identification of names for entities as a basis to disambiguate named entities whose sense is unknown (2008). However, this stream of entity disambiguation research has a number of limitations, being dependent on a pre-existing knowledge representation of some sort that literally lists the senses, be it a formal ontology or a more informal thesaurus or even just some textual corpus. These techniques are usually evaluated over a corpus such as news stories where the number of entities is bound and so can be correlated with the pre-existing knowledge representation (Nguyen and Cao, 2008). This general methodology ignores the point made by Masterman that the senses of English words are fundamentally open-ended, such that polysemy can infect even the most mundane of entity names over time (Wilks, 2005a). As noticed by Wilks (Wilks, 2005b), this applies to knowledge representations as well, for after decades of development even the formal terms used in Cyc are experiencing a 'drift' in terms of their sense. While in natural language and in formal languages before the Web, this ability for names to change meaning and for new names to appear happened relatively slowly over the lifetime of an individual, on the Web new entity names appear all the time, and previously stable names are 'cannibalized' by new entities on a regular basis. This was observed in our analysis of the query logs in Chapter 7. So Masterman's thesis about the open-ended number of senses is even more important on the Web than it is

in natural language.

Our technique succeeds insofar as it incorporates just the necessary amount of disambiguation needed, without fully disambiguating a name for an entity or concept to some pre-existing bounded knowledge representation. The entire point of the Semantic Web is that knowledge representation is open and unbounded, and thus new senses with their attendant descriptive knowledge representations in RDF may be added to the Semantic Web at any point. Given any new name with a sense, it is likely that information that connects that name to its sense it is likely to be found somewhere on the Web by an hypertext search engine. Thus, our technique, by applying an relevance feedback between hypertext web-pages on the open-ended Web to the equally open-ended Semantic Web knowledge representation presents a *general* technique for sense disambiguation of names on the Web, although our experiments show that the senses in the Semantic Web trail far behind the senses in the hypertext Web, as only 2% queries could find no relevant sense information on the hypertext Web, while 34% of the queries could not find a sense as the Semantic Web. However, for those queries where at least one sense could be found on the Semantic Web, how can we determine what is the best URI for that sense? Crucially, the queries by themselves are usually ambiguous as regards sense. The URIs also may have many different shades of senses. For example, is a WordNet URI for the Eiffel Tower a sense for the term 'Eiffel Tower' while somehow a DBpedia URI for the Eiffel Tower is a sense for Eiffel Tower itself? Do these two URIs share the same sense, or only the same sense to some degree? What if the URI is connected to some information that is incorrect, but some information that is correct, about some particular sense? These questions make the problem of sense disambiguation much less of a simple matching problem between named entities and senses, but more of a ranking of senses.

We employ our Wittgensteinian intuition that the context provided by the clicking of the user on web-pages can provide not complete named-entity disambiguation – which would require some closed list of senses – but the *minimum disambiguation necessary to get the task at hand complete*. Our technique of relevance feedback is in fact a form of sense disambiguation. Furthermore, we take into account the open-ended nature of senses by providing the lists of Semantic Web URIs for senses as a ranking of URIs, with the degree of relevance of the sense of the query being – if our algorithm performs well – approximated by its place in the ranking of search results. We clinch the sense disambiguation necessary since we crucially provide for the judges making the relevance judgments in Section 9.1.3 a 'snapshot' of the rele-

vant web-page clicked on by the original user who entered the query in addition to the query keywords. Our experiment has a very strict definition of relevance that confines the judges to only clicking on web-pages that definitely share a sense, and our experiments showed judges had agreement on this task, and so agreement on the senses of web-pages. Then, the entire list of clicked web-pages then are used as the necessary context needed to counter the sparsity of context given by the query keywords themselves. While this list of clicked web-pages may not provide enough context to make the sense of the desired entity or concept completely unambiguous, it provides enough context to make it *unambiguous enough*. This position is in line with our Wittgensteinian public language position that does not seek to eliminate ambiguity, but only to alleviate it as much as needed. The re-ranking of the returned Semantic Web URIs by relevance feedback then takes this new disambiguation context on board. The results given in Section 9.4.3 demonstrate that this method clinches the necessary disambiguation information as human judges believe the results are better. The 'best' Semantic Web URI is then not one that simply 'stands-in' for the sense. Instead, the 'best' Semantic Web URI for a sense is one whose knowledge representation matches the aggregated relevant information in the hypertext web-pages, information that crucially disambiguated among an open-ended continuum of senses.

One can imagine a new and improved day in the life of the Semantic Web if our system was deployed on a large scale. Prior to our system, on the Semantic Web there was little if any attempt to share and re-use URIs, primarily due to an inability to find them. Suppose Ralph was to visit the Eiffel Tower and wanted to reference it in some RDF triples produced by his Semantic Web-aware calendar planning software and then graph merge these triples with other triples, so he could serendipitously discover his friend Dan Brickley had just moved from Bristol to Paris. However, he would have to find the best URI for the Eiffel Tower, disambiguating the URI for the Eiffel Tower itself from that of the film *A View from the Eiffel Tower*. Also, Ralph would need to find a URI for Dan Brickley the Web developer, making sure it is disambiguated from the URI for Dan Brickley the fashion model. He could use a Semantic Web search engine like FALCON-S, but he would have to manually dig through rather unfriendly RDF triples, and Ralph is not a Semantic Web expert. However, with our system he can seamlessly use natural language queries in a normal hypertext search engine to find Semantic Web URIs and relevant information about the Eiffel Tower. Does he want the latitude and longitude of the Eiffel Tower? All Ralph has to do is type in `eiffel tower` and begin clicking on results as he normally does when he searches

the Web. By simply clicking on a result for the Eiffel Tower in Paris as opposed to the movie, he resolves a Semantic Web URI for the Eiffel Tower from DBpedia and gets valuable information about it, such as its latitude and longitude `48.8583, 2.2945` and location in Paris. When he wishes to correlate this data with his friends, when he types in `dan brickley` into a search engine, Ralph clicks on Dan's homepage. Dan's information, such as latitude and longitude and his being in Paris on the dates Ralph is in Paris, emerges. Also, Ralph notices that in an almost eerie fashion, as the Semantic Web information is consumed by his calendar program, his search results in the search engine improve. Ralph has stepped into the 'virtuous cycle' of the Semantic Web and Web search (Baeza-Yates, 2008).

Chapter 10

Conclusion and Future Directions

Language is the body of the mind. **Anton Pannekoek** (1912)

10.1 Conclusion

As described in the introduction to the thesis, we have given both a thorough analysis of the central problem of the Semantic Web and a practical solution. Here we will describe how we arrived at an analysis of the theoretical problem via the contributions of each chapter. We will also discuss whether or not our engineering solution to the problem is sufficient, namely by discussing some of the drawbacks of our system. Lastly, we will briefly demarcate some space opened for future theoretical research by the thesis.

The main theoretical problem confronting the Semantic Web in particular and the Web is 'what does a URI refer to?' In order to analyze and answer this question, we employed previous work in the philosophy of language. After asking the initial question in Chapter 1, in order for the question to be taken seriously, in Chapter 2 we gave a brief overview of the development of the Web. The history of the Web was traced from from Licklider's 'Man-Machine Symbiosis' hypothesis, through to Douglas Engelbart's 'Human Augmentation Project' and finally to the familiar hypertext Web and the Berners-Lee's vision for URIs to be universal identifiers. Far from a detour, this chapter sets up the crucial notion that architecture of the Web *itself* should be a first-class citizen of investigation. In Chapter 3, we step back and present a sketch of a unified terminological account of the philosophy of information and the philosophy of language. The main contribution of this chapter was our re-affirmation of Dummett's neo-Fregean doctrine of sense and reference, which we expanded by showing that nat-

ural language is *just* one possible language in which the distinction between sense and reference appears, and so the distinction between sense and reference is also present in any exchange of information, including those of computers communicating via formal languages. In Chapter 4 we exemplify the analysis put forward in Chapter 3 by laying out the architecture of the Web, which consists of a set of terms such as 'resource' and principles such as the 'Principle of Linking' that define the ideal interactions of those terms, using terminology from the philosophy of language and information. Most importantly, we claim that issues of meaning can be broken into the two separate issues of sense and reference. We claim a URI is an identifier for some kind of content independent of a particular encoding, and so a URI identifies a sense.

In Chapter 5, we show how the project of the Semantic Web naturally follows from the hypertext Web, by demonstrating how the primary Semantic Web language, RDF, is an application of Web architecture to the much older knowledge representation of semantic networks. Our analysis of the Semantic Web in terms of philosophy of language and information leads to a new insight, that the problem of determining the sense and reference of URIs is fundamentally the *unsolved* problem put forward by the Semantic Web. Also, in Chapter 5, we acknowledge Karen Spärck Jones's critique of the Semantic Web as a mere repetition of logic-based classical artificial intelligence. However, we escape unscathed from her criticism, since the use of URIs as names for things is the real *new* claim of the Semantic Web, not any particular knowledge representation scheme. So it is precisely within the realm of URIs that *technical* advance must be made.

In Chapter 6, we analyze the two most prominent positions on reference and URIs. The first position, the logicist position advanced primarily by Hayes, states that for the Semantic Web, the meaning of a URI is given by whatever model(s) satisfies the formal semantics of the Semantic Web. This position is shown to be a direct descendant of the philosophical descriptivist theory of reference, namely that the referent of a name is given by whatever satisfies the descriptions associated with the name, as put forward by Carnap, Russell, and Tarski. However, the practical failure in deployment of the early Semantic Web seems to vindicate the predictions of Spärck Jones that any purely logicist approach was doomed to failure. Another position is the direct reference position of Berners-Lee, which states that the referent of a URI is whatever was intended by the owner of the URI, which is a direct philosophical descendant of the causal theory of reference, that any name refers via some causal chain directly to a referent, as championed by Kripke and Putnam. However, due to the observation of

the principles of Web architecture to the Semantic Web and Berners-Lee's direct reference position, it appears that a new second generation of the Semantic Web, known as Linked Data, is experiencing large growth. In Chapter 7, we do the first empirical large-scale study of this new Linked Data Web, using queries from a large hypertext search engine to sample the Semantic Web. While we find that many of the principles of Web architecture are actually being followed, we also observe that with the tremendous release of data on the Web in the form of Linked Data there is still very little reuse or sharing of URIs, so that the same referent will tend to have multiple URIs. Instead of solving the problem, with the direct reference position everyone simply mints their own URIs, and little communication or merge data happens. Thus, we have shown so far in the thesis our analysis of the problem, namely that **the Semantic Web is a kind of language that can be defined by its conformance to the principles of Web architecture, but nonetheless inherits the problems regarding reference and meaning from the philosophy of natural language**.

In Chapter 8, we lay out a solution to the question of 'what does a URI refer to?' in the form of a new philosophical position based on Wittgenstein and a practical application based on applying relevance feedback from hypertext search engines to discover Semantic Web URIs. This public language position holds that the Semantic Web is a form of language, and as a language exists as a mechanism for co-ordination among multiple agents, then the meaning – and so the sense – of a URI is the use of the URI by a community of agents. We argue for this by noting that both the causal and descriptivist theories of names attempt to banish the notion of sense in favor of building an entire theory of meaning on top of only reference, and that their lack of success on the Semantic Web points to a return to the notion of a Fregean public and objective notion of sense. Then we argue that if the Semantic Web wants to be used as a *new* language of URIs, then it has no alternative but to build off of already-existing natural languages and activities such as hypertext Web search. In this vein, the Semantic Web needs a way to query for a natural language name for some concept or entity and get precisely the 'best' URI for the concept or entity. As shown in Chapter 7, currently state of the art Semantic Web search engines only return a relevant Semantic Web URI in return to a query 58% of the time. Therefore, we propose a *novel* solution to the problem; since both hypertext web-pages and Semantic Web data about the same referent share the same *sense* as defined in Chapter 3, regardless of their encoding, we can use relevance feedback from the hypertext Web search engines to bootstrap the Semantic Web. Finally, in Chapter 9 we test a deployment of the system on a subset

of the queries for concepts and entities used in Chapter 7. Using human subjects to manually judge the relevance of both Semantic Web data and hypertext web-pages in response to a query, we show that our system successfully uses the relevance feedback from hypertext web to boost the discovery of relevant Semantic Web URIs for concepts and entities. After exploring relevant parameters, our system performs better in terms of average precision than a baseline without feedback as well as FALCON-S, producing a relevant Semantic Web URI 89% of the time. Lastly, we show that using the relevant Semantic Web URIs as relevance feedback to a hypertext Web search engine also improves performance, resulting in better performance in the top 10 results than both a baseline without feedback and Yahoo! Web search. Therefore, our thesis conclusively demonstrates that **a theory of sense and reference suitable enough to encourage identifier re-usage on the Web can be implemented by employing relevance feedback from search engine results.**

10.2 Future Directions

There are two kinds of future directions the work in this thesis should take. The first is various technical improvements that should be implemented by our relevance-feedback systems, and the second is a more theoretical extension of the philosophical territory of the thesis.

10.2.1 Technical Improvements

There are a number of areas where our project needs to be more thoroughly integrated with other approaches and improved. In particular, we could use better language modeling and better explicit query expansion, the incorporation of multimedia and machine-translation, the creation of new Semantic Web URIs when none exists for a query, and increased scale.

10.2.1.1 Adapting Language Models and Query Expansion to the Web

While language models, particularly generative models as given by (Lavrenko, 2008), should in general have theoretically higher performance than vector-space models, our experiment in Chapter 9 showed a slight but significantly better performance for vector-space than language models in relevance feedback from hypertext web-pages

to the Semantic Web, likely due to the parameters of the language model being generated by the infamously messy and non-parametric natural language data of the Web. Furthermore, the reason why large-scale search engines do not in general implement language models for information retrieval is that the computational complexity of calculating distributions over billions of documents does not scale. However, there is reason to believe that relevance models could be scaled to work with Web search in general and Semantic Web search in particular if they built their language sample from a 'clean' and suitably large sample of natural language (as was done in our relevance-feedback experiment using relevant Semantic Web results) then these relevance models would be more effective. The computational complexity could be reduced via caching and the use of Bloom filters for the language model. This, combined with some sort of statistical query expansion that would help a user resolve ambiguous queries like `rock` into `rock music` or `geological rock`, would likely get our performance to about 89%. Further natural language processing, including better stemming and lemmatization, would also likely improve performance.

10.2.1.2 Integration of Multimedia and Machine Translation

Despite the fact that we maintained that the traditional problems of sense and reference should hold in *any* information in *any* language, including formal languages, we did not investigate any way to incorporate multimedia and other non-natural languages into our system. Instead, we reduced knowledge representation languages to a pseudo-natural language for processing. The incorporation of multimedia semantics would make the entire approach stronger. Also, this approach would fail for queries given in foreign languages. A query for `tour de eiffel` should return the same Semantic Web URI for the Eiffel Tower. Yet as our system relies on natural language term overlap with RDF in the associated descriptions, only integration with machine translation would allow the system to be able to resolve associated descriptions across different natural languages.

10.2.1.3 Automatic Creation of New URIs

One of the looming deficits of our system is that for a substantial amount of our queries there are *no* Semantic Web URIs. This amount is estimated in Chapter 7 as 34% of all queries, almost as many as there were queries where a non-relevant Semantic URI was the first result. However, as shown also in Chapter 7, these queries with no

Semantic Web URIs in general *do* have relevant information on the hypertext Web, if not the Semantic Web. In this manner, the automatic generation of Semantic Web triples from natural language text as explored by Brewster et al. (2007) and Cimiano et al. (2005) could be used in combination with our system to create new URIs, with accessible and automatically generated associated descriptions, in response to user queries. Furthermore, one could even imagine the reverse of new information being created, that is information that is not shared by hypertext web-pages being removed from associated descriptions.

10.2.1.4 Scale

Lastly, our system and experiment was only a *proof of concept* system, and it was tested only over a relatively small (although statistically significant) number of users and queries automatically harvested from a query engine. Far better would be to deploy this system with a global-scale hypertext search engine. The benefit to users would be instant: they would have access to structured data that could be taken advantage of by programs like SearchMonkey that could automatically format it in response to certain types of queries (Mika, 2008). The statistics over the whole Semantic Web and user queries would be interesting, allowing the identification of communities and a more data-driven approach to the creation of Semantic Web vocabularies. Given the growing interest in 'Semantic Search' in some version or another from large hypertext search companies like Google and Microsoft, the adoption of our feedback system in the wild is not impossible.

10.2.1.5 The Statistical Semantic Web

What should be apparent here is this project is but the first step in a new direction for the Semantic Web, one away from both the logicist Semantic Web and the Linked Data Web of databases to the *Statistical Semantic Web*, a Semantic Web constructed statistically from the behavior and language use of users of the Web. One could argue that the large hypertext Web is in fact precisely this statistical Semantic Web, but we would argue that without the use of URIs and the de-linking of the content of the data from particular encodings through the principles of Web architecture, these 'lower-case' statistical semantic webs created by hypertext are actually not part of the Web but closed data, whereas the Statistical Semantic Web would be an open Web of URIs and information created through statistical methods. However, we do reiterate the

central position of hypertext search engines should not be underestimated on the Web, and we find it astounding that the Semantic Web has ignored hypertext Web search engines, such that this thesis is the first to show how they can realistically be put in a mutually beneficial feedback cycle.

10.2.2 Theoretical Extensions

One fascinating possibility for future theoretical work is the impact of the Web to investigate the questions of intelligence and embodiment. These questions deserve more than the cursory treatment we give them here, but this treatment here shows the potential productivity of considering the Web a first class object of philosophical investigation.

10.2.2.1 The Extended Mind Hypothesis on the Web

The Extended Mind thesis sets the framework for our understanding of the utility of these digital representations on the Web (Clark and Chalmers, 1998). To explain the Extended Mind thesis, Clark introduces us to Ralph, a man with an impaired memory who navigates about his life via the use of his notebook, in particular to the Museum of Modern Art (1998). We will rephrase this example in the more familiar terms of Ralph's visit to the Eiffel Tower from Chapter 3. Let us assume Ralph has a serious memory impairment. Ralph is trying to navigate to the Eiffel Tower from the airport, and uses his notebook as a surrogate memory in order to discover the location. Ralph has a map in his notebook to the Eiffel Tower made for the precise purpose of navigating individuals to the monument. Ralph can get to the museum with the map, but without the map he would be lost. In this regard, the map qualifies as an 'external' representation that can drive the cognitive processes of an agent in a similar fashion to the way that classical artificial intelligence assumed internal representations did. Interestingly enough, Clark point out that if external factors are driving the process, then they deserve some of the credit: "If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process" (1998). The map and other external representations have been dubbed "cognitive technology" by Clark (2000).

The Web then presents an interesting twist on the Extended Mind Hypothesis extension that we presented earlier. Again, Ralph is using a web-page on his mobile

phone to find his way to the Museum of Modern Art. While our previous example had Ralph using the Web as ordinary Web users did years ago, simply downloading some directions and following them, we now add a twist. Imagine that Inga and Ralph are using a map-producing Web site that allows users to add annotations and corrections, a sort of wiki of maps. Inga, noticing that the main entrance to the Museum of Modern Art is closed temporarily due to construction and so the entrance has moved over a block, adds this annotation to the map, correcting an error as regards where the entrance to the Museum of Modern Art should be. This correction is propagated at speeds very close to real-time back to the central database behind the Web site. Ralph is running a few minutes behind Inga, and because this correction to the map is being propagated to his map on his personal digital assistant, Ralph can successfully navigate to the new entrance a block away. This (near) real-time updating of the representation was crucial for Ralph's success. Given his memory issues, Ralph would have otherwise walked right into the closed construction area around the old entrance to the Museum and been rather confused. This active manipulation with updating of an external representation lets Inga and Ralph possess some form of dynamically-changing collective cognitive state. Furthermore, they can use their ability to update this shared external representation to influence each other for their greater collective success. In this manner, the external representation is clearly social, and the cognitive credit must be spread across not only multiple people, but the representation they use in common to successfully accomplish their behavior. Clark and Chalmers agree that cognition can be socially extended, "What about socially extended cognition? Could my mental states be partly constituted by the states of other thinkers? We see no reason why not, in principle" (1998). How we extend their story is that socially extended cognition is now mediated by external representations, in particular by digital representations and other information accessible on the Web via URIs.

One of the obvious requirements for any process to be part of an extended mind is that it is accessible when needed to solve some problem. The obvious requirement is that the representation needed by the subject be within its effective reach, not separated from the subject in space or time. So if Ralph's notebook with the map to the Eiffel Tower has been left at home in Boston when he is in Paris, the notebook cannot count as part of his extended mind. Furthermore, if his notebook exists only in the past, such that it was destroyed in a fire before Ralph could use it, then the notebook also could not count as part of Ralph's extended mind at the current moment. The point here is that at least a minimal condition for anything to be cognitive technology is that it be

accessible over the bounds of space and time when needed with a reasonable latency. In other words, the external representation must have "reliable coupling," (Clark and Chalmers, 1998). The technical trajectory of Licklider's "Man-Machine Symbiosis" project, which could be considered the engineering twin of the philosophical Extended Mind thesis, is precisely to overcome the barriers of time and space that separate representations and their users. The Semantic Web is just the latest incarnation of this trend.

10.2.2.2 Embodiment Reconsidered

One of the strange repercussions that follows straightforwardly from a Wittgensteinian and neo-Fregean approach to sense as inherently objective and external is that as more and more of language, and thus our shared sense that guides our behavior, gets encoded in external representations with the possibility of low-latency Web access, it becomes unclear where the precise boundary point is in these feedback cycles between the individual and their external representation. If the cycle of connection and disconnections happens constantly, over many individuals, as it would if a major hypertext search engine pursued the Semantic Search approach given here, the very boundaries of agents become difficult to detect. If we become dependent on the Web, defining intelligence in terms of a fully autonomous agent then becomes not even an accurate portrayal of human intelligence, but "a certain conception of the human individual that may have applied, at best, to that faction of humanity who had the wealth, power, and leisure to conceptualize themselves as autonomous beings exercising their will through individual agency and choice" (Hayles, 1999). By jettisoning this conception, yet reconstructing the commitment to a certain kind or degree of embodiment, a new kind of philosophy that takes the Web seriously can do justice to complex phenomenon such as the advent of the Web and the increasing recognition of what Engelbart termed "collective intelligence" (Engelbart and Ruilifson, 1999). Pierre Levy notes that cognitive science "has been limited to human intelligence in general, independent of time, place, or culture, while intelligence has always been artificial, outfitted with signs and technologies, in the process of becoming, collective"(1994). The vast technological changes humanity has engendered across the world are now reshaping the boundaries of human bodies, and so the domain of cognitive science. This has been a process that has been ongoing since the dawn of humanity, and whose most momentous event was the evolution of natural language. Only now due to the incredible rate of technological progress, as exemplified by the growth of collective intelligence and new languages

like the Semantic Web on the Web, do changes in language become self-evident within the scope of a single lifetime.

10.2.2.3 The Science of the Web

While firmly based on Wittgenstein, the position that the Semantic Web is an attempt to create a new kind of public language goes against a certain quietism that Wittgenstein exhibits when he states that "philosophy may in no way interfere with the actual use of language; it can in the end only describe it" (Wittgenstein, 1953). Berners-Lee responds to such notions with a radical riposte, that on the Web "we are not analysing a world, we are building it" (Berners-Lee, 2003a). This radical outlook that engineering systems *are* philosophy given a digital embodiment is best summarized by Berners-Lee himself in the statement that "we are not experimental philosophers, we are philosophical engineers" (2003a). In contrast to any purely descriptive science, the primary difference of what has been termed the "science of the Web" is that not only can engineered systems be constructed to test theories, as done in traditional modeling in almost all scientific fields, but these models can be released upon the world at large through the Web (Berners-Lee et al., 2006b).

We hope that by integrating the Semantic Web with work on information retrieval as pioneered by Karen Spärck Jones, the Semantic Web itself can have a new lease on life and be tested on a large scale. Spärck Jones's objection to the Semantic Web was that it needed a single agreed upon ontology. As our exegesis of Berners-Lee and the Semantic Web has shown, this single agreed upon ontology is not a requirement for the Semantic Web. In contrast, Berners-Lee has long maintained that instead decentralized agreement on the use of URIs is enough, as put by Hendler, "a little semantics goes a long way" (Hendler, 2007). Yet how do we boot-strap this decentralized agreement, and let users find and re-use the best URIs? After analyzing the Semantic Web as a new kind of public language, we hypothesized that the Semantic Web should be grounded in the everyday behavior of the cybernetic form of life, the widespread use of search engines. While hypertext web search is already done via some form of adapted information retrieval, we show how time-tested techniques like relevance feedback can be used on a new kind of Web search: the search for Semantic Web URIs for concepts and entities. Our innovation is that we use well-known techniques for relevance feedback between the hypertext Web and the Semantic Web to increase the performance on this kind of semantic search. This demonstrates how users of the Semantic Web can take advantage of the use of search engines over vast amounts of text to give a statistical

semantics based in natural language to URIs and their attendant Semantic Web knowledge representations, and so find and re-use the best URIs for concepts and entities. The results of our experimental attempt to prove this are promising. If there is anything to be learned from Wittgenstein and the Web, it is that although one can never escape philosophical problems, one can make progress by interpreting them anew.

Appendix A

An Ontology for Web Architecture

The task of classifying all the words of language, or what's the same thing, all the ideas that seek expression, is the most stupendous of logical tasks. Anybody but the most accomplished logician must break down in it utterly; and even for the strongest man, it is the severest possible tax on the logical equipment and faculty. **Charles Sanders Peirce, letter to editor B. E. Smith of the Century Dictionary**

In order to better understand the nature of the 'Web' in the Semantic Web, a formal ontology called the 'Identity of Resources on the Web' (IRW) ontology was created. This ontology formally shows how terms in earlier chapters, particularly Chapters 4 and Chapter 5, can be related to the terminology given in Chapter 3. Formal ontologies have a long history of use in clarifying potentially confusing domains. Traditionally, the domains that have been most amendable to formal ontologies have been domains that are already highly structured, such as scientific domains like biology. However, one of the most exciting developments in modern knowledge representation is the advent of the Semantic Web, which hopes to combine the principles of the Web with the principles of knowledge representation in order to see many small, linked formal ontologies develop in a vast number of heterogeneous domains. The hope is that by combining the principles of the Web with formal ontologies, both the Web and formal ontologies will co-evolve together.

Although researchers have paid much attention to what kinds of logic best underlies knowledge representation on the Web, very little work has been done from the side of the knowledge representation community on understanding what exactly are the core principles and components of the Web itself. This is not surprising, as Web architecture is mostly an informal body of knowledge phrased in a combination of Internet and Web

standards, tutorials and notes, running code, and even an 'oral' tradition passed down in IRC chats and e-mail discussions. However, while each document itself is usually clear and self-contained, over the years many of the documents have been replaced with newer versions and extended in various manners, using the same vocabulary differently. Some parts of this myriad number of documents have been deprecated, and only some components are best practices. Furthermore, other informally written notes and even the products of long e-mail list-serv discussions have had an influence on the core architecture of the Web. Thus, the knowledge of Web architecture itself can be to outsiders, especially those coming from a background in knowledge representation, a rather obscure and even vague field despite its unreasonable effectiveness, since many of its principles are embodied primarily in the minds of its primary architects who do not in general attend academic conferences, specifications that are not mentioned in academic literature, and the running code that has been built off these specifications.

We model these terms and the debates around them using a lightweight formal ontology in OWL-DL, which we call *IRW*, for 'Identity of Resources on the Web.' IRW is meant to be an helpful formal tool for resolving conflicting arguments about identity and URIs, and as a consequence, it provides a supporting vocabulary for implementing practical solutions in a variety different scenarios (Halpin and Presutti, 2009). Further details of the ontology are available in Halpin and Presutti (2009). While there are limits to any formal ontology in describing such a multi-faceted field, a single ontology of how the terms in the various specifications fit together into a coherent body of knowledge is necessary. First, we will informally describe some of the components of the Web itself in order to then formally elucidate these components a formal ontology that allows us to model Web architecture. We will the end by presenting a number of surprisingly utilitarian uses that this formal ontology provides.

A.1 Related Work

The foundations of Web architecture have primarily been laid out in various specifications from the World Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF). The W3C and IETF have different structures and terminology. The W3C is a more recent and formal body technically concerned with the World Wide Web, and not the Internet as a whole. The W3C is a membership organization that features a strict formal process that moves, ideally on a limited time-scale, to create a normative W3C Recommendation that defines a Web standard. In contrast, the

IETF has existed almost since the the dawn of the Internet as an open organization that runs off ‘rough consensus and running code,’ and in lieu with its philosophy calls its documents ‘Request for Comments’ (RFCs), although it does distinguish a level of confidence, with kinds of documents progressing from the informal ‘Informational’ to the more mature ‘Internet Drafts’ and finally to ‘Standards Track’ documents.

The main source for our terminology is a document entitled *The Architecture of the World Wide Web* (AWWW) as described in in Chapter 4. The other group of standards that we will investigate is the various IETF RFCs around Uniform Resource Identifiers (URIs, such as *http://www.example.org*) and the HyperText Transfer Protocol (HTTP), which were both developed within the IETF, also described in Chapter 4. In particular the specifications around URIs, originally called ‘Universal Resource Identifiers,’ were first put forward by Berners-Lee in the ‘informational’ IETF RFC 1630 (Berners-Lee, 1994a). However, the IETF could not agree on this name and as such, the later RFC for Uniform Resource Locations (URLs), in the form of IETF RFC 1738 came out (Berners-Lee et al., 1994). URLs became URIs again with the publication of IETF RFC 2396 (Berners-Lee et al., 1998), which after a number of minor amendments, was later itself superseded by the full Internet Standard IETF RFC 3986 (Berners-Lee et al., 2005). Likewise, HTTP was first defined in RFC 2068 (Fielding et al., 1997), which was then shortly superseded by IETF the ‘Standards track’ IETF RFC 2616 (Fielding et al., 1999). When possible, we will use primarily the definitions of the later IETF RFCs when it obsoletes a previous RFC. W3C Recommendations, unlike IETF RFCs, are generally not made obsolete.

Informal notes are another major source of information. The W3C AWWW is an exegesis of Tim Berners-Lee’s notes on ‘Design Issues: Architectural and philosophical points’ that exist a collection of unordered personal notes available at: `http://www.w3.org/DesignIssues/`. Another major source of information is Roy Fielding’s dissertation “Architectural Styles and the Design of Network-based Software Architectures,” as Fielding was one of the principal architects of HTTP (2000). Lastly, much of the interest in Linked Data comes from the ‘How to Publish Linked Data on the Web’ note, itself a practical tutorial built from Berners-Lee’s informal ‘Linked Data’ note (Bizer et al., 2007).

A.2 The Use of a Formal Ontology

The primary use of a formal ontology in the context of Web architecture is to allow us to model formally the various distinctions employed by Web architecture. Although some other formal logic that deals with actions and events may be more suitable for modeling the temporal transactions of a client and server interactions on the Web, an ontology is necessary in order to capture the various distinctions given in Web specifications first. As even Web architects find themselves confused about the distinctions between 'entities' in HTTP and 'representations' in Web architecture (Mogul, 2002), this ontology could be of use as a reference to anyone interested in understanding or even extending existing Web specifications, as well as those interested in correctly implementing best practices that are dependent on rather obscure corners of Web architecture, such as Linked Data.

One of the most interesting uses of the ontology should be to phrase the arguments around the Identity Crisis in a way that allow those involved in debates to model formally their positions using extensions to a common ontology as a starting point. To this aim, IRW can be discussed, reviewed, and comment on the Ontology Design Patterns wiki[1]. To serve the aim of elucidating arguments, additional modules of IRW have been developed, in particular to deal with the debate between Berners-Lee and Hayes, and are briefly introduced in Section A.3.

There have been previous attempts to model at least a subset of Web architecture as given in Chapter 4 in a formal ontology, but all lack coverage of some crucial concepts. For example, while the ontology given by RDF Schema touches upon the vocabulary of resources via its term `rdfs:Resource`, it does not cover the distinction between information and non-information resources. The IRE (Identifiers, Resources, and Entities), based on Dolce Ultra Lite (DUL),[2] a light version of the widely-known DOLCE foundational ontology and its extension for describing information objects[3] (IOL, described in (Gangemi, 2008)), attempted to model some of these concepts earlier (Presutti and Gangemi, 2008). However, many aspects were not included in IRE, such as the distinctions between resources and their Web representations, or the concept of accessing a web-page via a web server, that are crucial to the efforts within the W3C, while many of the distinctions drawn by DUL+IOL were found to be too 'heavy-weight' for these communities (Gangemi et al., 2002). In response to these concerns, the IRE ontology

[1] `http://ontologydesignpatterns.org/wiki/Submissions:IRW`
[2] `http://www.loa-cnr.it/ontologies/DUL.owl`
[3] `http://www.loa-cnr.it/ontologies/IOLite.owl`

has been evolved into the IRW ontology.

We show graphically how this ontology can model the 303 redirection needed for the Semantic Web via an example. In the example an agent trying to access a URI for the Eiffel Tower itself, `http://dbpedia.org/resource/Eiffel_Tower`. Upon attempting to access that resource with a HTTP GET request on a URI, since the Eiffel Tower itself is not an information resource, no Web representations are directly available. Instead, the agent gets a `303 See Other` that in turn redirects them to an information resource that hosts Web representations about the Eiffel Tower, such as `http://dbpedia.org/page/Eiffel_Tower`. When this URI returns the 200 status code in response to an HTTP GET request, the agent can infer that the URI `http://dbpedia.org/page/Eiffel_Tower/` is actually an information resource.

The Semantic Web URI which is used to refer to the Eiffel Tower itself (not the web-page), `http://dbpedia.org/resource/Eiffel_Tower`, could be any kind of resource and so *could* be a non-information resource (Connolly, 2006). This example is illustrated in Figure A.1, using terms from the IRW ontology introduced in Section A.3.

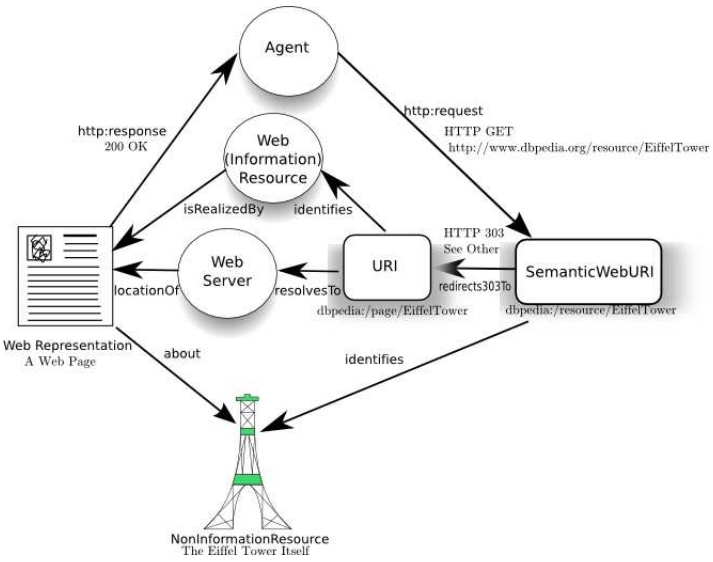

Figure A.1: 303 Redirection for Semantic Web URIs

In order to introduce the IRW ontology, we will first introduce its core concepts one by one, and distinguish when we are communicating about a module or part of the core ontology. The components of the ontology will then be used to model successfully the two primary use-cases, the modeling of the retrieval of hypertext web-pages and then the retrieval of Semantic Web data using the Linked Data principles.

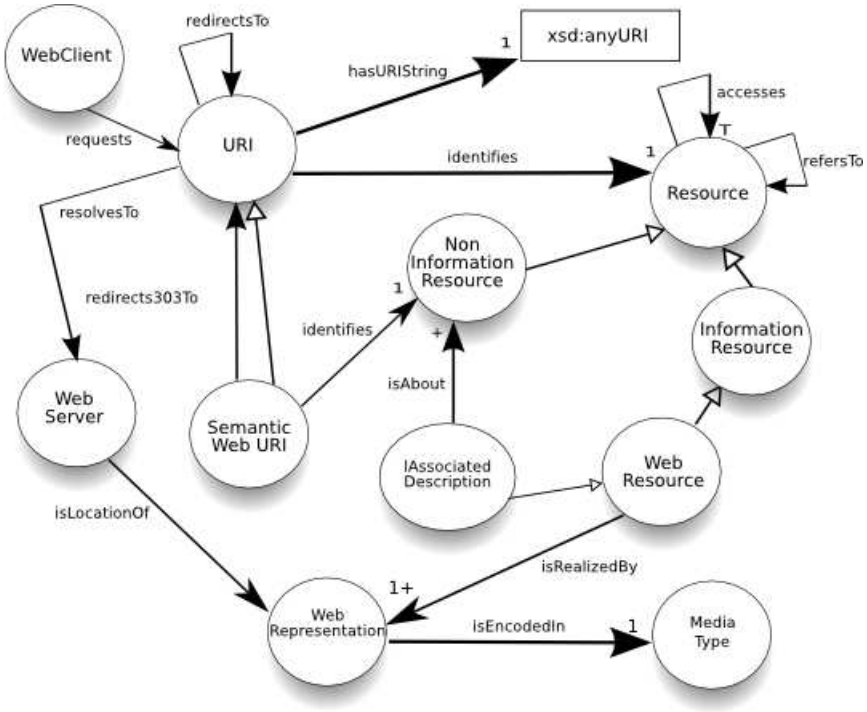

Figure A.2: The IRW ontology illustrated as a graph. Rounded nodes are classes, while rectangular ones are datatypes. Arcs ending with an empty triangle are `rdfs:subClassOf` relationships. Arcs ending with a filled triangle are either object properties or datatype properties depending of the range node. Arcs' direction indicates the domain and range of the property. A '1' associated to a property means it is functional, a 'T' means it is transitive, '1+' means 'at least one'. Prefixes are indicated only if different from `irw:`.

A.3 The IRW Ontology

The prefix `irw:` is for the namespace `http://purl.org/NET/irw/` of the IRW ontology. Terms in the ontology will be given in `teletype` font, and if no namespace is given, we will assume the `irw:` namespace. The stable version of the ontology can also be accessed via its PURL. The latest version of the IRW ontology may be accessed online.[4] The prefix `rdfs:` is used for the RDF(S) namespace `http://www.w3.org/2000/01/rdf-schema#` Note `ir:` is the 'information realization' ontology, also online.[5] While the IRW ontology in full can not graphically explicated due to lack of space, the primary classes and properties are given in Figure A.2. The IRW-related elements needed for the example

[4] `http://ontologydesignpatterns.org/ont/web/irw.owl`
[5] `http://www.ontologydesignpatterns.org/cp/owl/informationrealization.owl`

of 303 redirection are given in Figure 5.5. The IRW ontology starts with `Resource`. While this class expresses the same intuition as `rdfs:Resource`, we have defined it again because this version of IRW is within OWL-DL expressivity. In OWL Full, this class is equivalent to `rdfs:Resource`. Now, we move to modeling the debates around the Identity Crisis.

A.3.1 Resources and URIs

The notion of a URI is modeled as a class, `URI` that has exactly one value for the datatype property `hasURI` allowing to specify its value. Modeling URIs as a class allows us to talk about different kinds of URIs, such as IRIs (Internationalized Resource Identifiers) and Semantic Web URIs. A property `identifies` can then connect a URI to a resource. Since we want to associate a URI with a particular character string like '`http://www.example.org`' for the URI, we also have a property for called `hasURIString`. This property then have various sub-properties for future modules such as the conversion of IRIs to URIs, so that a IRI given in the Japanese character set can be converted to a URI. The sub-properties of `hasURIStrong` may be included like `hasRelativeURIString` and has `hasAbsoluteURIString` for the conversion of relative URIs to absolute URIs.

- ***Resource***: *An OWL Class*. "Anything that might be identified by a URI" (Jacobs and Walsh, 2004). This class is meant to express the same intuition of rdfs:Resource but it is defined here in order to have OWL-DL compaibility. In an OWL Full version of this ontology this class would be `owl:equivalentClass rdfs:Resource`.

- ***URI***: *An OWL Class*. An abbreviation for Uniform Resource Identifier. "A global identifier in the context of the World Wide Web" (Jacobs and Walsh, 2004). Any identifier that follows either fulfills the role given in IETF RFC 3986 can fulfill this class, even if it an identifier such as IRI that has a conversion to a URI or uses a scheme such as URN (Moats, 1997) or URL (Berners-Lee et al., 1994) that has been subsumed by the concept of URIs.

  subClassOf: `Resource`

- ***identifies***: *An OWL Object Property*. The relationship between a URI and a resource. It is functional as the W3C states one should "assign distinct URIs to distinct resources" (Jacobs and Walsh, 2004).

Inverse Property: `isIdentifiedBy`

Domain: `URI`

Range: `Resource`

subPropertyOf: `refersTo`

functional

- ***hasURIString***: *An OWL Object Property*. The relationship between a URI and the character encoding of a URI.

  Domain: `URI`

  Range: `xsd:anyURI`

A.3.2 Access and Reference

One of the largest re-occurring debates in Web is about whether the notion of 'identifies' between URI and resources is actually coherent. According to Berners-Lee, URIs identify exactly one resource (i.e. `identifies` is a functional property) via some causal and historical chain given by the owner or creator of the URI, a similar position towards names and reference as given by Kripke (1972). Hayes would disagree with `identifies` being functional, and would prefer the term be dropped all together from Web architecture (Hayes and Halpin, 2008). Instead, Hayes would use the more precise terms *access* (`accesses`) and *reference* (`refersTo`) (Hayes and Halpin, 2008). In the tradition of formal model theory and the Russellian descriptivist theory of reference (Russell, 1905), Hayes argues that a URI can refer to a referents in any interpretation that satisfies the model given by the formal semantics of RDF (Hayes and Halpin, 2008). In this way, a URI can refer to more than one resource, and so this can be modeled by the object property `refersTo`, which is non-functional unlike `identifies`, and so `identifies` can be sub-property of `refersTo`. One aspect of reference is that the object of reference can be "immediately causally disconnected" from its subject (Hayes and Halpin, 2008). This is important, as reference is used as a property between URIs and resources, including not only web-pages but also resources like the Eiffel Tower or integers that are necessary for the Semantic Web

However, it seems there should be another relationship besides reference: the relationship of 'access' for when "the name provides a causal pathway to the thing, perhaps mediated by the Web" (Hayes and Halpin, 2008). We call this relationship

the `accesses` property, which is a causal connection to the thing identified. Since this is an exceptionally common use of Web architecture, it is used within the core `irw` module. This is modeled again as a property between URIs and resources, although it is transitive, unlike `refersTo`. If one can access *a* and *a* accesses *b*, and *b* accesses *c*, then *a* accesses *c* (via *b*). Note that access and reference are not disjoint, for as "the architecture of the Web determines access, but has no direct influence on reference" and that one can use a URI that accesses a web-page to also refer to that web-page, or even something completely different.

- ***accesses***: *An OWL Object Property*. The relationship between a resource and another resource where the former provides a causal pathway to the latter.

  Inverse Property: `isAccessedBy`

  domain:`Resource`

  range: `Resource`

  transitive

- ***refersTo***: *An OWL Object Property*. The relationship between a resource and another resource where the former may be immediately causally disconnected from the latter.

  Inverse Property: `isAbout`

  domain: `Resource`

  range: `Resource`

A.3.3 Information Resources

There is a controversial sub-classes of `Resource` outlined in AWWW known as 'information resources.' As the AWWW defines the notion of *information resource* as "a resource which has the property that all of its essential characteristics can be conveyed in a message" (Jacobs and Walsh, 2004), which we model as `InformationResource`. This definition has widely been thought of as unclear, and defining what set of individuals belong in this class and what do not, has been a source of perpetual debate on various list-servs, and our formal modeling in combination with a few classes from a subset of DOLCE, DUL+IOL (Gangemi, 2008), hopefully will clarify the notion. An `InformationResource` is viewed to be equivalent to the notion of *information object* from DUL+IOL (Gangemi, 2008), such as a musical composition, a text, a word, or

a picture. An information object is an object defined at a level of abstraction, independently from how it is concretely realized. This means an information resource has, via the `ir:realizes` property (with inverse `ir:isRealizedBy`), at least one `ir:InformationRealization`, a concrete *realization*. This term is again imported from DUL+IOL (Gangemi, 2008). So an information resource's "essential characteristics can be conveyed in a single message" implies that everything from a bound book to an HTTP message can be a realization of an information resource (Jacobs and Walsh, 2004).

Examples of this are descriptions of a resource using natural language or depictions of a resource using images. Information resources also can, but not necessarily, be identified (either accessed or referred to) with a URI. In this manner, the text of Moby Dick can be an information resource since it could be conveyed as a single message in English, and can be realized by both a particular book or a webpage containing that text. The definition of information object and information realization can be thought of as the classic division in philosophy of mind between an object given on a level of abstraction and some concrete thing that realizes that abstraction, where a single abstraction may have multiple realizations. This is similar, but more broad, that the type-token distinction in philosophy and the *TBox* and *ABox* distinction from description logic used in OWL.

- ***InformationResource***: *An OWL Class*. "A resource which has the property that all of its essential characteristics can be conveyed in a message" (Jacobs and Walsh, 2004).

  subClassOf: `Resource`

  equivalentClass: `iol:InformationObject`, which is defined by IOL as "a concrete realization of an expression, e.g. the written document containing the text of a law" (Gangemi, 2008).

- ***ir:isRealizedBy***: *An OWL Object Property.* Imported from IOL. "A relation between an information realization and an information object, e.g. the paper copy of the Italian Constitution realizes the text of the Constitution" (Gangemi, 2008).

  Inverse Property: `ir:realizes`

  Domain: `ir:InformationRealization`

  Range: `ir:InformationObject`

- ***ir:InformationRealization***: *An OWL Class*. Imported from IOL. "A piece of information, such as a musical composition, a text, a word, a picture, independently from how it is concretely realized" (Gangemi, 2008). This is equivalent to the broadest notion of *representation* as defined in AWWW as "data that encodes information about resource state" (Jacobs and Walsh, 2004).

A.3.4 Web Resources and Web Representations

Up until now, all the work done by the ontology has not had much to do with the Web per se, but more with the more general ideas of information and resources that apply equally as well to books as to web-pages. However, we can now specialize this ontology to the Web. In particular, representations can be transferred over a protocol such as HTTP. However, in doing so they become something we call *Web representations* (`WebRepresentation`) with entity body and entity headers. Therefore, this use of the term 'representation' is more narrow that the AWWW's use, which is equivalent to the notion of any information realization in the large, and instead focused on representations sent over the Web. This is due to the AWWW specifying that "new protocols created for the Web should transmit representations as octet streams typed by Internet media types" (Jacobs and Walsh, 2004). Note also that therefore, as given in IETF RFC 2616, a Web representation may be defined as "an entity included with a response that is subject to content negotiation" such that "there may exist multiple representations associated with a particular response status" (Fielding et al., 1999). Furthermore, one can distinguish *Web resources* (`WebResource`), a subset of information resources that *are* usually Web-accessible, such as web-pages, from things that simply carry information, like the text of Moby Dick, regardless of whether it is on the Web or not.

However, one problem is that it appears a client may only access a Web representation of a resource as a response, and so we need a term for describing the request for a representation itself. To do this, we turn to the notion of *entity* (`Entity`) as defined by HTTP (Fielding et al., 1999). Entities may be used either for a request or response, but a representation is only for a response. Something can be an entity without necessarily being a representation or being transferred as bits over the wire from any particular Web resource. For example, the entity headers and entity body of a POST request, or even a 404 response, is an entity, but does not necessarily represent the state of a particular Web resource. The same entity may be transferred as the request or response of many particular actions by a client. Also, different URIs may return the same entity,

such as when one URI hosts a copy of a resource given by another URI. In order to model entities, we use the popular `hasComponent` ontology design pattern.

- ***WebResource***: *An OWL Class*. "A network data object or service" (Fielding et al., 1999). As such, this is a resource that is accessible via the Web (Hayes and Halpin, 2008). Therefore, a Web Resource must have at least one URI and be realized by at least one Web Representation.

  subClassOf: `InformationResource`

  ir:isRealizedBy: `WebRepresentation` where *minCardinality(1)*

  isIdentifiedBy: `URI` where *minCardinality(1)*

- ***Entity***: *An OWL Class*. "The information transferred as the payload of a request or response" (Fielding et al., 1999). "An entity consists of metainformation in the form of entity-header fields and content in the form of an entity-body" (Fielding et al., 1999).

  subClassOf: `iol:InformationRealization`

  hasComponent: `EntityHeader` where *minCardinality(1)*

  hasComponent: `EntityBody`

- ***EntityBody***: *An OWL Class* Whatever information is sent "in the request or response is in "a format and encoding defined by the entity-header fields" (Fielding et al., 1999). Also called in HTTP the 'content' of a message (Fielding et al., 1999).

- ***EntityHeader***: "Entity-header fields define metainformation about the entity-body or, if no body is present, about the resource identified by the request" (Fielding et al., 1999). Sometimes called in HTTP "meta-information" (Fielding et al., 1999). Various sub-classes of this class can define HTTP status codes (`StatusCode`), content encoding (`MediaType`), content language (`ContentLanguage`), date of creation (`DateCreation`, date of modification(`DateModification` and the like.

- ***WebRepresentation***: *An OWL Class*. "A sequence of octets, along with representation metadata describing those octets, that constitutes a record of the state of the resource at the time when the representation is generated" (Berners-Lee et al., 2005). Note that the term 'representation" is used for this class in IETF

RFC 3968, but has been changed to ‘Web Representation’ to separate it from the more general notion of ‘representation’ used in the W3C AWWW (Jacobs and Walsh, 2004)

subClassOf: `Entity`

A.3.5 Media Types, Generic, and Fixed Resources

One intriguing problem, central to the notion of Web representations and resources, is the connection between media types and resources. Very little work has been done in this area, likely due to the lack of use of content negotiation in general on the hypertext Web. For example, instead of using content negotiation to return versions of the same resource in multiple languages, many sites use explicit links. The only substantial work on this has been Berner-Lee’s note *Generic Resources* where he outlines an ontology of types of resources, conditioned by how the resource varies over HTTP requests (Berners-Lee, 1996b). Berners-Lee has informally said that a generic resource is equivalent to information resources, since the main important part of a generic resource is the information itself, not any particular realization of the information. So, for example, a resource like ‘the weather report of Oaxaca’ is a generic resource, as is the text of Moby Dick in any language. However, the ‘weather report of Oaxaca today’ is not a generic resource, nor is Moby Dick in English. Resources may also vary over time. For example, the text of Moby Dick will be the same over time and so be “time-invariant,” but the resource for the ‘weather report of Oaxaca’ will change over time and be “time-specific” (Berners-Lee, 1996b). Furthermore, resources may vary over media-type. For example, the same information may be given in some custom XML dialect or RDF, or the same depiction may be given in different formats like JPG and SVG. These resources are all imported from Berners-Lee’s *ont* ontology, and all quotes in the following definitions are from the ontology.[6] There are also ‘fixed resources’ that, regardless of time and natural language, always delver the same representation. For example, a resource for Moby Dick that gave always the same edition in the same language as plain text would be a fixed resource. The idea of a fixed resource is surprisingly common, as it equates a single web-page with a resource.

- ***ont:GenericResource***. *An OWL Class*. “This resource is a resource that can vary by media-type over any number of dimensions” (Berners-Lee, 1996a).

[6] Available at `http://www.w3.org/2006/gen/ont`.

subClassOf: `Resource`

equivalentClassWith: `InformationResource`

- ***ont:TimeInvariantResource***. *An OWL Class*. "A resource of which all representations are in the same version. Representations of the resource will not change as a result of the resource being updated to a version with time" (Berners-Lee, 1996a).

  subClassOf: `Resource`

  disjointClassWith: `ont:TimeSpecificResource`

  ir:realizedBy `WebRepresentation` where `hasComponent DateCreation` and `DateLastModifiation` and where *maxCardinality (1) and minCardinality(1)*

- ***ont:LanguageSpecificResource***. *An OWL Class*. "A resource of which all representations are in the same language" (Berners-Lee, 1996a).

  subClassOf: `Resource`

  disjointClassWith: `ont:LanguageInvariantResource`

  ir: realizedBy: `WebRepresentation` where `hasComponent ContentLanguage` *maxCardinality(1) and minCardinality(1)*

- ***ont:MediaTypeSpecificResource***. *An OWL Class*. "A resource of which all representations are in the same media-type" (Berners-Lee, 1996a).

  subClassOf: `Resource`

  disjointClassWith: `ont:LanguageInvariantResource`

  realizedBy: `WebRepresentation` where `hasComponent MediaType` where *maxCardinality(1) and minCardinality(1).*

- ***ont:FixedResource***. *An OWL Class*. "A resource from which only one entity will ever come" (Berners-Lee, 1996a).

  subClassOf: `Resource`

  disjointClassWith: `ont:LanguageInvariantResource`

  realizedBy: `WebRepresentation` where *maxCardinality(1) and minCardinality(1).*

A.3.6 Hypertext Web Transactions

The typical Web transaction is started by an agent, given by a class `Agent`, which is some client in the context of the Web (Jacobs and Walsh, 2004). This agent can have a *request* (`request`) from a URI an representation. The Entity then contains a URI, which is the URI where that identifies the URI of the resource the request is acting upon, and this is modeled via the *request* mechanism. Both of these properties are a sub-property of `access`. An `Agent` can then request a representation from a URI. We also introduce the class `Web Server` for the generic notion of a *web server*, which has a `resolves` property. The property *resolves* is the resolution of a URI to a concrete Web server, which currently is done by mapping a URI to an IP address or addresses. So each `WebServer` has at least one `URI`. In order for the resolution to be successful, the Web It also has a `locationOf` property with at least one `Web Representation`, indicating the Web Server concretely can respond to an HTTP request with a particular Web Representation. Since `requests`, `resolves`, and `responses` are all sub-properties of the transitive property `accesses`, this part of the ontology models the physical and causal pathway between a given request for a URI and a responded to Web Representation. Then there is a *response* property that is the inverse of the `request` property that concretely returns the representation.

The entity given in the request may have a preferred media-type, and the response should have a media-type as well. The media-type, such as `application/xml` or `application/rdf+xml`, tells the agent how to interpret the entity body of the response (the returned Web representation of the resource). The media-types are found in the list given online by IANA.[7] Each of the media-types can be given a sub-class of our `MediaType` class. The relationship between a `MediaType` and a `Entity` is given by the `encodes` relationship. Note that each Web Representation should have a single media-type.

A URI may also have a `redirectsTo` property, a sub-property of `accesses`, that we can use to model HTTP redirection. This can be done via a number of different techniques, ranging from a ‘Content-Location’ HTTP entity header to a 300-hundred level HTTP status code. Note that, even in the light of the W3C TAG’s *httpRange-14* decision, since redirection can be used between just information resources that have nothing to do with the Semantic Web, their domain and range say nothing about the type of resource.

[7] http://www.iana.org/assignments/media-types/

- ***Agent***: *An OWL Class*. A program that establishes connections for the purpose of sending requests (Fielding et al., 1999). Also known as an 'Agent' in the W3C AWWW, which is "A person or a piece of software acting on the information space on behalf of a person, entity, or process" (Jacobs and Walsh, 2004).

  subClassOf: `Resource`

- ***request***: *An OWL Object Property*. "A request message from a client to a server includes, within the first line of that message, the method to be applied to the resource, the identifier of the resource, and the protocol version in use" (Fielding et al., 1999).

  Inverse Property: `response`

  subPropertyOf: `accesses`

  domain: `Agent`

  range: `URI`

- ***WebServer***: *An OWL Class*. "An application program that accepts connections in order to service requests by sending back responses"(Fielding et al., 1999). Note that "any server may act as an origin server, proxy, gateway, or tunnel, switching behavior based on the nature of each request" (Fielding et al., 1999).

  subClassOf: `Resource`

- ***resolvedBy***: *An OWL Object Property*. The relationship between a Web Server and a Web URI that hosts a representation of the resource identified by the URI.

  Inverse Property: `resolves`

  subPropertyOf: `accesses`

  domain: `WebServer`

  range: `URI`

  minCardinality(1)

- ***response***: *An OWL Object Property*. "After receiving and interpreting a request message, a server responds with an HTTP response message" (Fielding et al., 1999).

  Inverse Property: `isResponseBy`

  subPropertyOf: `access`

domain: `Entity`

- ***locatedOn***: *An OWL Object Property*. A relation between a Web Representation and a Web Server, indicating that the entity can be obtained by e.g. an HTTP request to the Web server.

  InverseProperty: `isLocationOf`

  subPropertyOf: `access`

  domain: `WebRepresentation`

  range: `WebServer`

- ***MediaType***: *An OWL Class*. "the media type of the underlying data" of a response (Fielding et al., 1999). The various registered media-types and their associated IETF RFC and can each be given its own sub-class.

  subClassOf: `Resource`

- ***isEncodedIn***: *An OWL Object Property*. The relationship between a entity and its media type.

  InverseProperty: `encodedIn`

  domain: `Entity`

  range: `MediaType`

  minCardinality(1) and *maxCardinality(1)* if applied to a `WebRepresentation`.

- ***redirectsTo***: *An OWL Object Property*. The relationship between one URI and another where any requested Entity is sent to the URI given as the object of this property.

  Inverse Property: `redirectedFrom`

  subPropertyOf: `access`

  domain: `URI`

  range: `URI`

A.3.7 Modeling the Semantic Web and Linked Data

In order to model explicitly the redirection solution to the "Identity Crisis" by the W3C TAG, a few new properties have been minted. These new properties are the

redirects303To and redirectsHashTo. Obviously, redirects303To models the TAG's 'solution' to *httpRange-14* while redirectsHashTo represents the hash convention.

With these kinds of redirections in hand, we can now model the typical Semantic Web transaction. A new sub-class of URI, SemanticWebURI is given, where the *Semantic Web URI* has a constraint that it must have at least one redirects property. The Semantic Web is supposed to use URIs not for Web resources ('documents') but for abstract concepts and real-world things themselves. The redirection allows the URI to refer to or identity a resource that is not accessible on the Web. In the 'Linked Data Tutorial' note, these are called *non-information resources* (Bizer et al., 2007). Although this term is controversial and hard to define abstractly, operationally it simply means a resource that is not Web-accessible that therefore should, to comply with the Linked Data initiative, using redirection to resolve to another resource. Although the space of non-information resources is relatively large and hard to draw precise boundaries around, we list a few exemplars in order to serve as what Dennett would call "intuition-pumps" in order to help us understand this concept (1981). In particular, a new class called NonInformationResource that represents things that can not themselves – for whatever reason – be realized as a single digitally encoded message, is introduced and is disjoint with InformationResource. A number of different kinds of things may be NonInformationResources. Since this concept is the cause of much confusion and debate, it is detailed with two disjoint sub-classes. A *physical entity resource* (PhysicalEntityResource), is a resource that is 'touchable' like physical people, artifacts, places, bodies, chemical substances, biological entities, etc. mapping to a subset of "entities" within OKKAM (Bouquet et al., 2007a). A *conceptual resource* (ConceptualResource) refers to resources that are created in a social process that can't be completely realized digitally, such as legal entities, political entities, social relations, as well as the concept of horse and imaginary objects like unicorns.

This kind of resource is an *associated descriptions* (AssociatedDescription), which is just an Web resource that can be accessed via redirection from a Semantic Web URI (Bizer et al., 2007). For example, in DBpedia[8] the resource dbpedia:/resource/Eiffel_Tower redirects to some RDF/XML at dbpedia:/data/Eiffel_Tower, and to an HTML page at dbpedia:/page/Eiffel_Tower depending on the requested media type (Auer et al., 2007). This Linked Data typical scenario can be generalized: a WebClient

[8]Prefix dbpedia: is used for the namespace http://dpedia.org

`requests` a `SemanticWebURI` *x* and the request is redirected (e.g. via hash or 303 redirection) to another URI, where this second URI identifies an `AssociatedDescription` that has one `isAbout` property to a non-information resource. We model `AssociatedDescription` as a subclass of `WebResource`.

- ***SemanticWebURI***: *An OWL Class*. A URI used to identify any resource that is not accessible on the Web.

  subClassOf: `URI`

  redirectsTo: `AssociatedDescription`

  identifies: `NonInformationResource`

- ***NonInformationResource***: *An OWL Class*. All resources that are not information resources

  subClassOf: `WebResource`

  complementOf: `InformationResource`

  redirectedFrom: `SemanticWebURI`

- ***PhysicalEntityResource***: *An OWL Class*. Some thing that occupies its own space and has its own mass in the real world but is not Web-accessible.

  subClassOf: `NonInformationResource`

- ***ConceptualResource***: *An OWL Class*. Resources that are created in the social communication process. A conceptual resource does not exist if it's not in a social communication. For example: legal entities, political entities, social relations, concepts, and the like.

  subClassOf: `NonInformationResource`

- ***AssociatedDescription***: *An OWL Class*. A resource that exists primarily to describe a non-Web accessible resource.

  subClassOf: `WebResource`

  redirectedFrom: `SemanticWebURI`

- ***redirects303To***: *An OWL Object Property*. A redirection that uses the HTTP 303 status code.

  Inverse Property: `redirected303From`

domain: `URI`

range: `URI`

functional

- ***redirectsHashTo***: *An OWL Object Property*. A redirection that works via the fragment identifier being removed from the URI.

  Inverse Property: `redirectedHashFrom`

  domain: `URI`

  range: `URI`

A.4 Uses of the IRW Ontology

The IRW ontology has many uses in the real world of the Web. These uses can operate on a number of different levels, both theoretical and practical. On the level of theory, it can help clarify the various arguments over Web architecture, such as the relationship between resources and representations. On a practical level, the TAG's decision of *httpRange-14* has been considered ambiguous, and the IRW ontology can resolve this difficulty by making resources more self-describing. Lastly, it can be used to determine if some URI is enabled to host Linked Data.

A.4.1 Resolving the Identity Crisis

One purpose of this ontology is to describe, in formal detail, the exact nature of the conflicts between the various sides of the Identity Crisis debate. The main conflict between Hayes and Berners-Lee can then be cast as an argument over three IRW properties. Berners-Lee's slogan that 'URIs identify one thing' is modeled by having the `identifies` property be *functional*, i.e. a URI can only identify one resource. Furthermore, he would also hold that a `SemanticWebURI` `refersTo` exactly one `NonInformationResource`.

Hayes's response would be that `identifies` should be eliminated and there can be no constraints whatsoever on `refersTo` and thus no constraints on the usage of URIs for referring to things on the Semantic Web, while typical hypertext Web transactions can be modeled functionally with `accesses`. Although IRW models Berners-Lee's

more general notion of identification via `identifies`, IRW also captures Hayes's perspective with the properties `refersTo` and `accesses`. Lastly, the criticisms of redirection modeled with `redirectsTo` has mainly to do with the fact that the domain can only be a `URI` rather than a `SemanticWebURI`, which we also explicitly model. Thus, there is no way to ever definitely be sure that a URI is a Semantic Web URI and so one can never be sure that a URI identifies a non-information resource. We show how IRW can solve this problem in Section A.4.2.

A.4.2 The Self-Describing Semantic Web

The IRW ontology can help explicitly model and make available to the rest of the Semantic Web the often subterranean details of Web architecture. The IRW ontology can also solve the problem noted earlier that currently it is impossible to describe whether or not some resource describes some non-Web accessible thing, such that there is no "definition, description, some other kind of indication of what the identifier is intended to identify" (Pepper, 2006). Solving this can be done on via adding IRW statements to associated descriptions accessible via Semantic Web URIs. There would be a number of advantages if web-pages that have RDF content could distinguish themselves as such, in the same way that HTML 'valid' documents are currently validated by W3C Validators. This can be done by embedding a IRW statement in RDF/XML documents, RDF returned from SPARQL endpoints, and RDFa or GRDDL statement in XHTML or XML documents (Adida et al., 2008). Ideally, this would be in conjunction with some sort of graphical logo to distinguish the page as 'Semantic Web Enabled,' much as current web-pages can be marked up with a logo for 'XHTML 1.0 Valid.' This is useful because detecting RDF 'in the wild' on the Web, such as embedded RDFa, can be difficult for humans. The main problem is that an `NonInformationResource` has no Web representation to embed such a statement in. Take for example the Semantic Web URI created by Pat Hayes for himself: `www.ihmc.us/users/phayes/PatHayes.html`. While originally a stand-alone web-page, currently Hayes has the URI use 303 redirection to `http://www.ihmc.us/users/phayes/PatHayesAbout.html`.[9] This latter web-page could easily use a combination of RDFa as IRW to mark itself up as a representation of a non-information resource by including the statement that `phayes:PatHayes.html rdf:type NonInformationResource` and adding

[9] Let `phayes:` stand for `http://www.ihmc.us/users/phayes`.

`phayes:PatHayes.html irw:redirects303To phayes:PatHayesAbout.html`.

A.4.3 Linked Data Validation

One subset of this second application is the use of IRW to systematize the process of Linked Data validation. Currently, the only Linked Data validator is *Vapour*, which is coded procedurally and whose results can not themselves be presented as RDF (Berrueta et al., 2008). The IRW and the HTTP in RDF vocabulary can be used to record whether or not each Linked Data resource is properly redirected using 303 redirection, and the IRW vocabulary can be used to make sure that the `303` redirection can lead access *both* an associated description in HTML and in RDF (Koch et al., 2008). An example of Linked Data validation is given below. Assuming that the URI `http://dbpedia.org/resource/Eiffel_Tower` is claiming to be hosting data in accordance with the Linked Data principles, we can check it in the following ways:[10]

Input
`http://dbpedia.org/resource/Eiffel_Tower`

Check

- If HTTP 303 Request with content request type `application/rdf+xml` returns a RDF file
- If HTTP 303 Request with content request type `text/html` returns an HTML file

Output (if succeeded)

```
dbpedia:resource/Eiffel_Tower redirects303To
dbpedia:page/Eiffel_Tower
dbpedia:resource/Eiffel_Tower redirects303To
dbpedia:data/Eiffel_Tower
```

Inferences

```
dbpedia:data/Eiffel_Tower isAbout
dbpedia:resource/Eiffel_Tower
dbpedia.org:page/Eiffel_Tower isAbout
```

[10] With the namespace `dbpedia:` being for `http://dbpedia.org`.

```
dbpedia.org:resource/Eiffel_Tower
dbpedia:resource/Eiffel_Tower rdf:type NonInformationResource
dbpedia.org:data/Eiffel_Tower rdf:type AssociatedDescription
dbpedia:page/Eiffel_Tower rdf:type AssociatedDescription
```

A.5 Conclusion

Overall, the IRW ontology can serve as a foundational ontology of Web architecture, the "dark side of Semantic Web" that Hendler believes may give the Semantic Web a crucial advantage over previous efforts in knowledge representation (2007). What is surprising is that it has taken so long for an ontology to be created for Web architecture. However, the debates between advocates of Web architecture can themselves be highly contentious and the terminology often misunderstood. Furthermore, the various documents that describe this problem are spread throughout many informal and semi-formal notes and standards (and arguments over e-mail lists), so systematizing the terminology and modeling it formally was perhaps more difficult than would be expected. Future work needs to be done to standardize IRW and further evolve the ontology through the W3C and the wider communities around the Semantic Web and Web architecture, which will doubtless result in refinements to IRW. It is far too easy to take the Web for granted. It is always those things that are closest to us that are the most difficult to speak about. Yet by developing a coherent language for describing Web architecture, a concrete step in establishing a new kind of philosophy of the Web has been taken.

Bibliography

Adida, B., Birbeck, M., McCarron, S., and Pemberton, S. (2008). RDFa in XHTML: Syntax and Processing. W3C Recommendation, W3C. http://www.w3.org/TR/rdfa-syntax/.

Allan, J., Connell, M., Croft, W. B., Feng, F. F., Fisher, D., and Li, X. (2000). INQUERY and TREC-9. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pages 551–562, Gaithersburg, Maryland, USA.

Althusser, L. (1912). *Marxism And Darwinism*. Verso, London, United Kingdom. http://www.marxists.org/archive/pannekoe/1912/marxism-darwinism.htm (Last accessed Jan. 9th 2008). Translated by Nathan Weiser.

Althusser, L. (1963). Marxism and Humanism. In *For Marx*. Verso, London, United Kingdom. Republished in 2005 by Verso, translated by Ben Brewster.

Andrews, K., Kappe, F., and Maurer, H. (1995). The Hyper-G network information system. *Journal of Universal Computer Science*, 1(4):206–220.

Anklesaria, F., McCahill, M., Linder, P., Johnson, D., Torrey, D., and Alberti, B. (1993). IETF RFC 1436 the Internet Gopher protocol. Category: Informational. http://www.ietf.org/rfc/rfc1436.txt (Last accessed on Oct. 5th 2008).

Anscombe, G. (1958). On brute facts. *Analysis*, 18:68–72.

Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *Proceedings of the International and Asian Semantic Web Conference (ISWC/ASWC2007)*, pages 718–728, Busan, Korea.

Baeza-Yates, R. (2008). From capturing semantics to semantic search: A virtuous cycle. In *Proceedings of the 5th European Semantic Web Conference*, pages 1–2, Tenerife, Spain.

Baeza-Yates, R., Calderon-Benavides, L., and Gonzalez, C. (2006). Understanding user goals in web search. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, pages 98–109, Glasgow, United Kingdom.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley-Longman, New York City, New York, USA.

Baeza-Yates, R. A., Ciaramita, M., Mika, P., and Zaragoza, H. (2008). Towards semantic search. In *Proceedings of Conference on Applications of Natural Language to Information Systems (NLDB)*, pages 4–11, London, United Kingdom.

Barabasi, A.-L., Albert, R., Jeong, H., and Bianconi, G. (2000). Power-law distribution of the World Wide Web. *Science*, 287:2115.

Bardini, T. (2000). *Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing*. Stanford University Press, Stanford, California, USA.

Batelle, J. (2003). The database of intentions. http://battellemedia.com/archives/000063.php (Last accessed Dec. 11th 2008).

Bateson, G. (2001). *Steps to an Ecology of Mind*. University of Chicago Press, Chicago, Illinois, USA.

Battelle, J. (2005). *The Search*. Portfolio, New York City, New York, USA.

Beckett, D. and Berners-Lee, T. (2008). Turtle - Terse RDF Triple Language. Member submission, W3C.

Beged-Dov, G., Brickley, D., Dornfest, R., Davis, I., Dodds, L., Eisenzopf, J., Galbraith, D., Guha, R., MacLeod, K., Miller, E., Swartz, A., and van der Vlist, E. (2001). RDF Site Summary (RSS) 1.0. Technical report, http://web.resource.org/rss/1.0/spec.

Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2):61–71.

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716.

Bergson, H. (1911). *Creative Evolution*. Dover Press, New York City, New York, USA. Republished in 1998, translated by Arthur Mitchell.

Berners-Lee, T. (1989). Information management: A proposal. Technical report, CERN. http://www.w3.org/History/1989/proposal.html (Last accessed on July 12th 2008).

Berners-Lee, T. (1991). Document naming. Informal Draft. http://www.w3.org/DesignIssues/Naming (Last accessed on July 28th 2008).

Berners-Lee, T. (1994a). IETF RFC 1630 Universal Resource Identifier (URI). http://www.ietf.org/rfc/rfc1630.txt (Last accessed on May 3rd 2008).

Berners-Lee, T. (1994b). World Wide Web Future Directions. Plenary Talk. http://www.w3.org/Talks/WWW94Tim/ (Last accessed on Oct. 5th 2008).

Berners-Lee, T. (1996a). Generic resource ontology. http://www.w3.org/2006/gen/ont (Last accessed on January 8th 2009).

Berners-Lee, T. (1996b). Generic resources. Informal Draft. http://www.w3.org/DesignIssues/Generic.html (Last accessed on Dec. 4th 2008).

Berners-Lee, T. (1996c). Universal Resource Identifiers: Axioms of Web Architecture. Informal Draft. http://www.w3.org/DesignIssues/Axioms.html (Last accessed Sept. 5th 2008.

Berners-Lee, T. (1998a). Cool URIs don't Change. http://www.w3.org/Provider/Style/URI (Last accessed on Nov 19th 2008).

Berners-Lee, T. (1998b). Semantic Web Road Map. Informal Draft. http://www.w3.org/DesignIssues/Semantic.html (Last accessed on April 12th 2008).

Berners-Lee, T. (1998c). What the Semantic Web can represent. Informal Draft. http://www.w3.org/DesignIssues/rdfnot.html (Last accessed on Sept. 12th 2008).

Berners-Lee, T. (2000). *Weaving the Web*. Texere Publishing, London.

Berners-Lee, T. (2003a). Message on www-tag@w3.org list. http://lists.w3.org/Archives/Public/www-tag/2003Jul/0158.html (Last accessed on May 20th 2008).

Berners-Lee, T. (2003b). Message to www-tag@w3.org. http://lists.w3.org/Archives/Public/www-tag/2003Jul/0127.html (Last accessed on May 20th 2008).

Berners-Lee, T. (2003c). Message to www-tag@w3.org. http://lists.w3.org/Archives/Public/www-tag/2003Jul/0022.html (Last accessed on May 20th 2008).

Berners-Lee, T., Cailliau, R., Groff, J.-F., and Pollermann, B. (1992). World-Wide Web: The Information Universe. In *Electronic Networking: Research, Applications and Policy*, pages 74–82. Meckler, Westport, Connecticut, USA.

Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanara, R., Hollenbach, J., Lerer, A., and Sheets, D. (2006a). Tabulator: Exploring and analyzing Linked Data on the Web. In *Proceedings of the Third International Semantic Web User Interaction Workshop*, Athens, Georgia, USA.

Berners-Lee, T. and Connolly, D. (1993). IETF Working Draft HyperText Markup Language (HTML): A Representation of Textual Information and MetaInformation for Retrieval and Interchange. http://www.w3.org/MarkUp/draft-ietf-iiir-html-01.txt.

Berners-Lee, T., Fielding, R., and Frystyk, H. (1996). IETF RFC 1945 Hypertext Transfer Protocol (HTTP/1.0). http://www.ietf.org/rfc/rfc1945.txt (Last accessed on Oct. 5th 2008).

Berners-Lee, T., Fielding, R., and Masinter, L. (1998). IETF RFC 2396 Uniform Resource Identifier (URI): Generic Syntax. http://www.ietf.org/rfc/rfc2396.txt (Last accessed on Sept. 15th 2008).

Berners-Lee, T., Fielding, R., and Masinter, L. (January 2005). IETF RFC 3986 Uniform Resource Identifier (URI): Generic Syntax. http://www.ietf.org/rfc/rfc3986.txt (Last accessed on April 2th 2008).

Berners-Lee, T., Fielding, R., and McCahill, M. (1994). IETF RFC 1738 Uniform Resource Locators (URL). http://www.ietf.org/rfc/rfc1738.txt (Last accessed on Sept. 3th 2008).

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., and Weitzner, D. J. (2006b). Creating a science of the Web. *Science*, 313(5788):769–771.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):35–43.

Berners-Lee, T. and Kagal, L. (2004). The fractal nature of the Semantic Web. *AI Magazine*, 29(3).

Berrueta, D., Fernandez, S., and Frade, I. (2008). Cooking HTTP content negotiation with Vapour. In *Proceedings of Scripting for the Semantic Web Workshop at the European Semantic Web Conference*, Tenerife, Spain,.

Biron, P. and Malhotra, A. (2004). XML Schema Part 2: Datatypes. Recommendation, W3C. http://www.w3.org/TR/xmlschema-2/ (Last accessed March 13th 2008).

Bizer, C., Cygniak, R., and Heath, T. (2007). How to publish Linked Data on the Web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (Last accessed on May 28th 2008).

Bizer, C., Heath, T., Idehen, K., and Berners-Lee, T. (2008). Linked Data on the Web. In *Proceedings of the WWW2008 Workshop on Linked Data on the Web*, Beijing, China.

Bizer, C. and Seaborne, A. (2004). D2RQ: Treating non-RDF databases as virtual RDF graphs. In *Proceedings of International Semantic Web Conference*, Hiroshima, Japan.

Bobrow, D. and Winograd, T. (1977). Experience with KRL-0: One cycle of a knowledge representation language. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 213–222.

Boley, H. and Kifer, M. (2008). RIF Basic Logic Dialect. Recommendation, W3C. http://www.w3.org/TR/rif-bld/ (Last accessed August 8th 2008).

Booth, D. (2008). URIs Declaration versus Use. In *Proceedings of Identity, Reference, and the Semantic Web Workshop at the European Semantic Web Conference*, Tenerife, Spain,.

Borden, J. and Bray, T. (2002). Resource Directory Description Language (RDDL). http://www.rddl.org/(Last accessed August 8th 2008).

Bouquet, P., Stoermer, H., and Giacomuzzi, D. (2007a). OKKAM: Enabling a Web of Entities. In *I3: Identity, Identifiers, Identification. Proceedings of the WWW2007*

Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007., CEUR Workshop Proceedings, ISSN 1613-0073. online http://CEUR-WS.org/Vol-249/submission_150.pdf.

Bouquet, P., Stoermer, H., Tummarello, G., and Halpin, H., editors (2007b). *Proceedings of the WWW2007 Workshop I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007*, CEUR Workshop Proceedings. CEUR-WS.org.

Bouquet, P., Stoermer, H., Tummarello, G., and Halpin, H., editors (2008). *Proceedings of the ESWC2008 Workshop on Identity, Reference, and the Web, Tenerife, Spain, June 1st, 2008*, CEUR Workshop Proceedings.

Box, D., Ehnebuske, D., Kakivaya, G., Layman, A., Mendelsohn, N., Nielsen, H., Thatte, S., and Winer, D. (2000). Simple Object Access Protocol (SOAP) 1.1. http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

Brachman, R. (1983). What IS-A is and isn't: An analysis of taxonomic links in semantic networks. *IEEE Computer*, 16(10):30–36.

Brachman, R. and Schmolze, J. (1985). An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):151–160.

Brachman, R. and Smith, B. (1980). Special issue on knowledge representation. *SIGART Newsletter*, 70:1–38.

Bray, T., Paoli, J., and Sperberg-McQueen, C. (1998). Extensible Markup Language (XML). Recommendation, W3C. http://www.w3.org/TR/1998/REC-xml-19980210 (Last accessed on March 10th 2008).

Brewster, C., Iria, J., Zhang, Z., Ciravegna, F., Guthrie, L., and Wilks, Y. (2007). Dynamic iterative ontology learning. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP)*, Borovets, Bulgaria.

Brickley, D. and Guha, R. V. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. Recommendation, W3C. http://www.w3.org/TR/rdf-schema/ (Last accessed on Nov. 15th 2008).

Brickley, D. and Miller, L. (2000). FOAF Vocabulary Specification. http://xmlns.com/foaf/spec/ (Last accessed on Nov 20th 2008).

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 107–117, Brisbane, Australia.

Broder, A. (2002). A taxonomy of Web search. *SIGIR Forum*, 36(2):3–10.

Bundy, A., McNeill, F., and Walton, C. (2006). On repairing reasoning reversals via representational refinements. In *Proceedings of FLAIRS (Florida AI Research Society) Conference*, Melbourne Beach, Florida USA.

Bush, V. (1945). As we may think. *Atlantic Monthly*, 1(176):101–108.

Cancho, R. F. and Sole, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100:788–791.

Caracciolo, C., Euzenat, J., Hollin, L., Ichise, R., Isaac, A., Malaise, V., Meilicke, C., Pane, J., Shvaiko, P., Stuckenschmidt, H., Svab-Zamazal, O., and Svatek, V. (2008). Results of the Ontology Alignment Evaluation Initiative 2008. In *Proceedings of The International Workshop on Ontology Matching*, Karlsruhe, Germany.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Carnap, R. (1928). *The Logical Structure of the World*. University of California Press, Berkeley, California, USA. Republished in 1967.

Carnap, R. (1947). *Meaning and Necessity: a Study in Semantics and Modal Logic*. University of Chicago Press, Chicago, Illinois, USA.

Carnap, R. (1950). Empiricism, semantics, and ontology. *Revue Internationale de Philosophie*, 4:20–40.

Carnap, R. and Bar-Hillel, Y. (1952). An outline of a theory of semantic information. Technical Report RLE-TR-247-03150899, Research Laboratory of Electronics, Massachusetts Institute of Technology.

Carpenter, B. (1996). IETF RFC 1958 Architectural Principles of the Internet. http://www.ietf.org/rfc/rfc1958.txt (Last accessed on March 12th 2008).

Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):261–272.

Cerf, V. and Kahn, R. (1974). A protocol for packet network intercommunication. *IEEE Transactions on Communications*, 22(4):637–648.

Cheng, G., Ge, W., and Qu, Y. (2008). FALCON-S: Searching and browsing entities on the semantic web. In *Proceedings of the the World Wide Web Conference*, Beijing, China.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, Paris, France.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge, United Kingdom.

Cimiano, P. and Volker, J. (2005). Text2Onto - A framework for ontology learning and data-driven change discovery. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513, pages 227–238, Alicante, Spain.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, Massachusetts, USA.

Clark, A. (2000). *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford University Press, Oxford, United Kingdom.

Clark, A. (2002). Minds, brains, tools. In Clapin, H., editor, *Philosophy of Mental Representation*, pages 66–90. Clarendon Press, Oxford, United Kingdom.

Clark, A. and Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1):7–19.

Clark, K. (1978). Negation as failure. In Gallaire, H., Minker, J., and Nicolas, J., editors, *Logic and Databases*. Plenum, New York City, New York, United States.

Clauset, A., Shalizi, C., and Newman, M. (2007). Power-law distributions in empirical data. http://arxiv.org/abs/0706.1062v1 (Last accessed October 13th 2008).

Connolly, D. (1998). The XML revolution. *Nature*. http://www.nature.com/nature/webmatters/xml/xml.html (Last accessed on April 3rd 2008).

Connolly, D. (2002). An evaluation of the World Wide Web with respect to Engelbart's requirements. Informal Draft. http://www.w3.org/Architecture/NOTE-ioh-arch (Last accessed on Dec. 4th 2008).

Connolly, D. (2006). A pragmatic theory of reference for the Web. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, Edinburgh, Scotland. http://www.ibiblio.org/hhalpin/irw2006/dconnolly2006.pdf (Last accessed November 22nd 2008).

Connolly, D. (2007). Gleaning Resource Descriptions from Dialects of Languages (GRDDL). Technical report, W3C. Recommendation.

Craswell, N., Zaragoza, H., and Robertson, S. (2005). Microsoft Cambridge at TREC-14: Enterprise Track. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, page http://research.microsoft.com/apps/pubs/default.aspx?id=65241 (Last accessed January 10th 2009, Gaithersburg, Maryland, USA.

Cummins, R. (1996). *Representations, Targets, and Attitudes*. MIT Press, Cambridge, Massachusetts, USA.

Davis, G. and Olson, M. (1985). *Management information systems: Conceptual foundations, structure, and development*. McGraw-Hill, New York City, New York, USA.

Deacon, T. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. W.W. Norton, New York City, New York, USA.

Deleuze, G. and Guattari, F. (1991). *What is Philosophy?* Columbia University Press, New York City, New York, USA. Translated by Janis Tomlinson and Graham Burchell (1996).

Delugach, H. (2007). ISO Common Logic. Standard, ISO. http://cl.tamu.edu/ (Last accessed on March 8th 2008).

Dennett, D. (1981). *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, Cambridge, Massachusetts, USA.

DeRose, S., Maler, E., and Orchard, D. (2001). XML Linking Language (Xlink) Version 1.0. Recommendation, W3C. http://www.w3.org/TR/xlink/ (Last accessed on Nov. 12th 2008).

Ding, L. and Finin, T. (2006). Characterizing the Semantic Web on the Web. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 242–257, Athens, Georgia, USA.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V. C., and Sachs, J. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, pages 652–659, Washington, D.C., USA.

Dowty, D. (2007). Compositionality as an Empirical Problem. In Barker, C. and Jacobson, P., editors, *Direct Compositionality*, pages 23–101. Oxford University Press, Oxford, United Kingdom.

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press, Cambridge, Massachusetts, USA.

Dreyfus, H. (1979). *What Computers Still Can't Do: A critique of artificial reason*. MIT Press, Cambridge, Massachusetts, USA.

Dummett, M. (1973). *Frege: Philosophy of Language*. Duckworth, London, United Kingdom.

Dummett, M. (1993). What is a Theory of Meaning. In *The Seas of Language*, pages 1–33. Oxford University Press, Oxford, United Kingdom. Originally published in *Truth and Meaning: Essays in Semantics* in 1976.

Engelbart, D. (1962). Augmenting Human Intellect: A Conceptual Framework. Technical report, Stanford Research Institute. AFOSR-3233 Summary Report.

Engelbart, D. (1990). Knowledge-domain interoperability and an open hyperdocument system. In *Proceedings of the Conference on Computer-Supported Collaborative Work*, pages 143–156, Los Angeles, California.

Engelbart, D. and Ruilifson, J. (1999). Bootstrapping our collective intelligence. *ACM Computer Survey*, 31(4):38.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D., and Yates, A. (2004). Web-scale information extraction in KnowItAll. In *Proceedings of the International World Wide Web conference (WWW)*, pages 100–110, New York City, New York, USA.

Euzenat, J. and Shvaiko, P. (2007). *Ontology Matching*. Springer-Verlag, Berlin, Germany.

Evans, G. (1982). *The Varieties of Reference*. Oxford University Press, Oxford, United Kingdom. Edited by John McDowell.

Ferraiolo, J. (2002). Scalable Vector Graphics (SVG) 1.0 Specification. Recommendation, W3C. http://www.w3.org/TR/2001/REC-SVG-20010904/ (Last accessed April 22nd 2008.

Fielding, R. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., and Berners-Lee, T. (1999). IETF RFC 2616 Hypertext Transfer Protocol - HTTP 1.1. http://www.ietf.org/rfc/rfc2616.txt (Last accessed on April 2nd 2008).

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. (1997). IETF RFC 2068 hypertext transfer protocol - HTTP 1.1. http://www.ietf.org/rfc/rfc2068.txt (Last accessed on March 12th 2008).

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Floridi, L. (2004). Open problems in the philosophy of information. *Metaphilosophy*, 35(4):554–582.

Fodor, J. (1975). *The Language of Thought*. MIT Press, Cambridge, Massachusetts, USA.

Fountain, A., Hall, W., Heath, I., and Davis, H. (1990). Microcosm: An open model for hypermedia with dynamic linking. In *Proceedings of Hypertext: Concepts, Systems and Applications (ECHT)*, pages 298–311, Paris, France.

Fredkin, E. (2003). An introduction to digital philosophy. *International Journal of Theoretical Physics*, 42(1):189–247.

Frege, G. (1892). Uber Sinn und Bedeutung. *Zeitshrift fur Philosophie and philosophie Kritic*, 100:25–50. Reprinted in The Philosophical Writings of Gottlieb Frege (1956), Blackwell, Oxford, United Kingdom (1956), translated by Max Black.

Galloway, A. (2004). *Protocol: How Control Exists After Decentralization*. MIT Press, Boston, Massachusetts, USA.

Gangemi, A. (2008). Norms and plans as unification criteria for social collectives. *Journal of Autonomous Agents and Multi-Agent Systems*, 16(3):70–112.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, R., and Schneider, L. (2002). Sweetening ontologies with DOLCE. In *Proceedings of International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 166–181, Siguenza, Spain. Springer.

Gerber, A., van der Merwe, A., and Barnard, A. (2008). A Functional Semantic Web Architecture. In *Proceedings of the 5th European Semantic Web Conference*, pages 273–287, Tenerife, Spain.

Ginsberg, A. (2006). The big schema of things. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, Edinburgh, Scotland. http://www.ibiblio.org/hhalpin/irw2006/aginsberg2006.pdf.

Glaser, H., Millard, I., and Jaffri, A. (2008). RKBExplorer.com: A knowledge driven infrastructure for Linked Data providers. In *Proceedings of European Semantic Web Conference (ESWC)*, pages 797–801, Tenerife, Spain.

Goodman, N. (1968). *Languages of Art: An Approach to a Theory of Symbols*. Bobbs-Merrill, Indianapolis, Indiana, USA.

Granka, L., Joachims, T., and Gay, G. (2004). Eye-tracking analysis of user behavior in www search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479, Sheffield, United Kingdom.

Grice, P. (1957). Meaning. *The Philosophical Review*, 66:377–388.

Guha, R., McCool, R., and Miller, E. (2003). Semantic search. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 700–709, Budapest, Hungary.

Guha, R. V. (1996). Meta Content Framework: A White paper. http://www.guha.com/mcf/wp.html (Last accessed Aug. 11th 2008).

Haarslev, V. and Mueller, R. (2003). Racer: An OWL reasoning agent for the Semantic Web. In *Proceedings of the International Workshop on Applications, Products and Services of Web-based Support Systems at the IEEE International Conference on Web Intelligence*, pages 91–95, Halifax, Canada.

Hafner, K. and Lyons, M. (1996). *Where Wizards Stay Up Late: The Origins of the Internet*. Simon and Schuster, New York City, New York, USA.

Halasz, F. and Schwartz, M. (1994). The Dexter hypertext reference model. *Communications of the ACM*, 37(2):30–39.

Halpin, H. (2004). The Semantic Web: The Origins of Artificial Intelligence Redux. In *Proceedings of Third International Workshop on the History and Philosophy of Logic, Mathematics, and Computation (HPLMC-04 2005)*, Donostia San Sebastian, Spain. Republished in 2007 by Icfai University Press in The Semantic Web. http://www.ibiblio.org/hhalpin/homepage/publications/airedux.pdf (Last accessed April 2nd 2008).

Halpin, H. (2006). Representationalism: The hard problem for artificial life. In *Proceedings of Artificial Life X*, pages 527–534, Bloomington, Indiana.

Halpin, H. (2008a). Foundations of a philosophy of collective intelligence. In *Proceedings of Convention for the Society for the Study of Artificial Intelligence and Simulation of Behavior*, Aberdeen, Scotland.

Halpin, H. (2008b). Philosophical Engineering: Towards a Philosophy of the Web. *APA Newsletter on Philosophy and Computers*, 7(2):5–11.

Halpin, H., Hayes, P., and Thompson, H. S., editors (2006). *Proceedings of the WWW2006 Workshop on Identity, Reference, and the Web, Edinburgh, United Kingdom, May 23, 2008*. http://www.ibiblio.org/hhalpin/irw2006 (Last accessed May 1st 2008).

Halpin, H. and Presutti, V. (2009). An Ontology of Resources: Solving the Identity Crisis. In *Proceedings of European Semantic Web Conference (ESWC)*, pages 521–534, Heraklion, Crete.

Halpin, H., Robu, V., and Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th International World Wide Web Conference (WWW'07)*, pages 211–220, Banff, Canada.

Halpin, H. and Thompson, H. (2005). Web Proper Names: Naming Referents on the Web. In *Proceedings of The Semantic Computing Initiative Workshop at the World Wide Web Conference*, Chiba, Japan.

Halpin, H. and Thompson, H. S. (2006). One document to bind them: combining XML, Web Services, and the Semantic Web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 679–686, Edinburgh, Scotland.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D*, 42:335–346.

Haugeland, J. (1981). Analog and analog. In *Mind, Brain, and Function*, pages 213–226. Harvester Press, New York City, New York, USA.

Haugeland, J. (1991). Representational genera. In *Philosphy and Connectionist Theory*, pages 61–89. Erlbaum, Mahwah, New Jersey, USA.

Hausenblas, M., Halb, W., Raimond, Y., and Heath, T. (2008). What is the size of the Semantic Web? In *Proceedings of Conference on Semantic Systems (iSemantics)*, Graz, Austria. http://tomheath.com/papers/hausenblas-isemantics2008-size-of-semantic-web.pdf (Last accessed May 1st 2008).

Hawking, D., Voorhees, E., Craswell, N., and Bailey, P. (2000). Overview of the TREC-8 web track. In *Proceedings of the Text REtrieval Conference (TREC)*, pages 131–150, Vienna, Virginia.

Hayes, P. (1977). In defense of logic. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 559–565, Cambridge, Massachusetts, USA.

Hayes, P. (1979). The Naive Physics Manifesto. In *Expert Systems in the Micro-Electronic Age*, pages 242–270. Edinburgh University Press, Edinburgh, Scotland.

Hayes, P. (2002). Catching the dream. http://www.aifb.uni-karlsruhe.de/ sst/is/WebOntologyLanguage/hayes.htm (Last accessed Oct. 17th 2008).

Hayes, P. (2003a). Message to www-rdf-comments@w3.org. http://lists.w3.org/Archives/Public/www-tag/2003Jul/0147.html (Last accessed on May 20th 2008).

Hayes, P. (2003b). Message to www-rdf-comments@w3.org. http://lists.w3.org/Archives/Public/www-tag/2003Jul/0198.html (Last accessed on May 20th 2008).

Hayes, P. (2004). RDF Semantics. Recommendation, W3C. http://www.w3.org/TR/rdf-mt/ (Last accessed Sept. 21st 2008).

Hayes, P. (2006). In defense of ambiguity. In *Proceedings of the Identity, Reference, and the Web Workshop at the WWW Conference*, Edinburgh, Scotland. http://www.ibiblio.org/hhalpin/irw2006/hayes.pdf (Last accessed on Oct. 5th 2008).

Hayes, P. and Halpin, H. (2008). In defense of ambiguity. *International Journal of Semantic Web and Information Systems*, 4(3).

Hayes, P. and Menzel, C. (2001). A semantics for the knowledge interchange format. In *Proceedings of 2001 Workshop on the IEEE Standard Upper Ontology*, Seattle, Washington, USA. Available at http://reliant.teknowledge.com/IJCAI01/HayesMenzel-SKIF-IJCAI2001.pdf (last accessed Nov 19th 2008.

Hayles, N. K. (1999). *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics*. University of Chicago Press, Chicago, Illinois.

Hayles, N. K. (2005). *My Mother was a Computer: Digital Subjects and Literary Texts*. University of Chicago Press, Chicago, Illinois.

Heath, T. and Motta, E. (2007). Revyu.com: A reviewing and rating site for the web of data. In *Proceedings of the International Semantic Conference and Asian Semantic Web Conference (ISWC/ASWC2007)*, pages 718–728, Busan, Korea.

Hegel, G. (1959). *Säammtliche Werke*. Fromann, Stuttgart, Germany.

Hendler, J. (2007). The Dark Side of the Semantic Web. *IEEE Intelligent Systems*, 22(1):2–4.

Hirst, G. (2000). Context as a spurious concept. In *Proceedings of Context in Knowledge Representation and Natural Language, AAAI Fall Symposium*, pages 273–287, North Falmouth, Massachusetts.

Horrocks, I. (1998). Using an expressive description logic: FaCT or fiction? In *Proceedings of Sixth Principles of Knowledge Representation and Reasoning Conference*, pages 636–647, San Francisco, California.

Huynh, D. F., Karger, D. R., and Miller, R. C. (2007). Exhibit: Lightweight structured data publishing. In *Proceedings of the World Wide Web Conference (WWW)*, pages 737–746, Banff, Canada.

Israel, D. and Perry, J. (1990). What is information? In Hanson, P., editor, *Information, Language, and Cognition*, pages 1–19. University of British Columbia Press, Vancouver, Canada.

Jacobs, I. (1999). W3C Mission Statement. Technical report, W3C. http://www.w3.org/Consortium/.

Jacobs, I. and Walsh, N. (2004). Architecture of the World Wide Web. Technical report, W3C. http://www.w3.org/TR/webarch/ (Last accessed Oct 12th 2008).

Jaffri, A., Glaser, H., and Millard, I. (2008). Managing URI synonymity to enable consistent reference on the Semantic Web. In *Proceedings of the Workshop on Identity, Reference, and the Web (IRSW) at ESWC2008*.

Jameson, F. (1981). *The Political Unconscious*. Cornell University Press, Ithaca, New York, USA.

Janes, J. (1993). On the distribution of relevance judgments. In *Proceedings of the American Society for Information Science*, pages 104–114, Medford, NJ.

Jansen, B. J., Booth, D. L., and Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. *Information Process and Management*, 44(3):1251–1266.

Jones, K. S. (1964). Synonymy and semantic classification. Thesis, Cambridge University. Republished in 1984 by Edinburgh University Press.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Jones, K. S. (1999). Information Retrieval and Artificial Intelligence. *Artificial Intelligence Journal*, 114:257–281.

Jones, K. S. (2004). What's new about the Semantic Web?: Some questions. *SIGIR Forum*, 38(2):18–23.

Kahn, R. and Wilensky, R. (2006). A framework for distributed digital object services. *International Journal on Digital Libraries*, 6(2).

Kilgarriff, A. (1993). Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26:356–387.

Klyne, G. and Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. Recommendation, W3C. http://www.w3.org/TR/rdf-concepts/.

Koch, J., Velasco, C. A., and Abou-Zahra, S. (2008). HTTP Vocabulary in RDF. W3C Working Draft, W3C. http://www.w3.org/TR/EARL10-Schema/.

Koller, D. and Pfeffer, A. (1998). Probabilistic frame-based systems. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 580–587, Madison, Wisconsin.

Kripke, S. (1972). *Naming and Necessity*. Harvard University Press, Cambridge, Massachusetts, USA.

Kwok, C., Etzioni, O., and Weld, D. (2001). Scaling question answering to the Web. In *Proceedings of the International World Wide Web conference (WWW)*, pages 150–161, Hong Kong, Hong Kong.

Lassila, O. and Swick, R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. Deprecated W3C Recommendation, W3C. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Lavrenko, V. (2008). *A Generative Theory of Relevance*. Springer-Verlag, Berlin, Germany.

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., and Thomas, S. (2002). Relevance models for topic detection and tracking. In *Proceedings of Human Language Technologies Conference, HLT 2002*, pages 104–110.

Lavrenko, V. and Croft, W. B. (2001). Relevance-based language models. In *Proceedings of the Twenty-Fourth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, New Orleans, Louisiana, USA. ACM Press.

Leiner, B., Cerf, V., Clark, D., Kahn, R., Kleinrock, L., Lynch, D., Postel, J., Roberts, L., and Wolff, S. (2003). A brief history of the internet. http://www.isoc.org/internet/history/brief.shtml (Last accessed March 20th 2008).

Lenat, D. (1990). Cyc: Towards programs with common sense. *Communications of the ACM*, 8(33):30–49.

Levensque, H. and Brachman, R. (1987). Expressiveness and tractability in knowledge representation and reasoni ng. *Computational Intelligence*, 3(1):78–103.

Levy, P. (1994). *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Plenum Press, New York City, New York, USA.

Lewis, D. (1971). Analog and digital. *Nous*, 1(5):321–327.

Licklider, J. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, 1:4–11.

Luntley, M. (1999). *Contemporary Philosophy of Thought*. Blackwell, London, United Kingdom.

Lyotard, J.-F. (1988). *The Inhuman: Reflections on Time*. Editions Galilee, Paris, France. Republished 1998 by Blackwell. Translated by Geoffrey Bennington and Rachel Bowlby.

Masterman, M. (1961). Semantic message detection for machine translation, using an interlingua. In *Proceedings of International Conference on Machine Translation of Languages and Applied Language Analysis*, London, United Kingdom.

May, R., Levin, S., and Sugihara, G. (2008). Ecology for bankers. *Nature*, 451:893–895.

McCarthy, J. (1959). Programs with common-sense. *Nature*, 188:77–91. http://www-formal.stanford.edu/jmc/mcc59.html (Last accessed May 1st 2008).

McCarthy, J. (1980). Circumspection – a form of nonmonotonic reasoning. *Artificial Intelligence*, 1(13):27–39.

McCarthy, J. (1992). 1959 memorandum. *IEEE Annals of the History of Computing*, 14(1):20–23. Reprint of original memo made in 1952.

McCarthy, J. and Hayes, P. (1969). Some philosophical problems from the standpoint of Artificial Intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence*, volume 4, pages 463–502. Edinburgh University Press.

McCarthy, J., Minksy, M., Rochester, N., and Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial intelligence. Technical report, Dartmouth College. http://www-

formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (Last accessed March 12th 2008).

McDermott, D. (1987). A critique of pure reason. *Computational Intelligence*, 33(3):151–160.

McKay, D. (1955). The place of meaning in the theory of information. In Cherry, E., editor, *Information Theory*, pages 215–225. Basic Books, New York City, New York, USA.

Mealling, M. and Daniel, R. (1999). IETFRFC 2483 URI resolution services necessary for URN resolution. Experimental. http://www.ietf.org/rfc/rfc2483.txt (Last accessed April 13th 2008).

Mendelsohn, N. (2006). The Self-Describing Web. Draft TAG finding, W3C. http://www.w3.org/2001/tag/doc/namespaceState-2006-01-09.html (Last accessed March 7th 2008).

Mika, P. (2008). Microsearch: An Interface for Semantic Search. In *Proceedings of Semantic Search Workshop at the European Semantic Web Conference*, Tenerife, Spain,.

Mikheev, A., Grover, C., and Moens, M. (1998). Description of the LTG system used for MUC. In *Seventh Message Understanding Conference: Proceedings of a Conference*.

Miles, A. and Bechhofer, S. (2008). SKOS Simple Knowledge Organization System reference. W3c recommendation, W3C. http://www.w3.org/TR/skos-reference/.

Miller, G. (1995). Wordnet: A lexical database for English. *Communications of the ACM*, 11(38):39–41.

Millikan, R. (1984). *Language, Thought and Other Biological Categories: New Foundations for Realism*. MIT Press, Cambridge, Massachusetts, United States.

Millikan, R. (2000). Naturalizing Intentionality. In Elevitch, B., editor, *Proceedings of Twentieth World Congress on Philosophy, Philosophy of Mind*, pages 83–90. Philosophy Documentation Center, Charlottesville, Virginia, USA.

Millikan, R. (2004). *Varieties of Meaning*. MIT Press, Cambridge, Massachusetts, United States.

Minsky, M. (1975). A framework for representing knowledge. In Winston, P., editor, *The Psychology of Computer Vision*, pages 211–277. McGraw Hill, Columbus, Ohio, USA.

Mizarro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832.

Moats, R. (1997). IETF RFC 2141 URN Syntax. http://www.ietf.org/rfc/rfc2141.txt (Last accessed April 20th 2008).

Mockapetris, P. (1983). IETF RFC 882 Domain Names - Concpets and Facilities. http://www.ietf.org/rfc/rfc882.txt (Last accessed on March 12th 2008).

Mogul, J. (2002). Clarifying the fundamentals of HTTP. In *Proceedings of the 11th International World Wide Web Conference*, pages 444–457, Honolulu, Hawaii, USA. ACM.

Monk, R. (1991). *Ludwig Wittgenstein: The Duty of Genius*. Penguin, New York City, New York, USA.

Montague, R. (1970). English as a formal language. In Visentini, B., editor, *Linguaggi nella Societa e nella Tecnica*, pages 189–224. Edizioni di comunita, Milan, Italy.

Mueller, V. (2007). Representation in digital systems. In *Proceedings of Adapation and Representation*, Paris, France. http://www.interdisciplines.org/adaptation/papers/7 (Last accessed March 8th 2008).

Needham, R. (1962). A method for using computers in information classification. In *Proceedings of the IFIP Congress*, pages 284–287, Vienna, Austria.

Nelson, T. (1965). Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of 20th National Conference of the Association for Computing Machinery*, pages 84–100, Cleveland, Ohio, USA.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 1(4):135–183.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351.

Nguyen, H. T. and Cao, T. H. (2008). Named entity disambiguation: A hybrid statistical and rule-based incremental approach. In *Proceedings of the Asian Semantic Web Conference (ASWC2008)*, pages 420–433, Bangkok, Thailand.

Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., and Tummarello, G. (2008). Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics, and Ontologies*, 3(1):37–52.

Paşca, M. (2007). Weakly-supervised discovery of named entities using web search queries. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM)*, pages 683–690, Lisbon, Portugal.

Parsia, B. (2003). Message to www-rdf-comments@w3.org. http://lists.w3.org/Archives/Public/www-rdf-comments/2003JanMar/0366.html (Last accessed on May 20th 2008).

Parsia, B. and Patel-Schneider, P. F. (2006). Meaning and the semantic web. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, Edinburgh, Scotland. http://www.ibiblio.org/hhalpin/irw2006/bparsia2006.pdf.

Pennebaker, W. and Mitchell, J. (1992). Joint photographic still image data compression standard. Standard, ISO.

Pepper, S. (2006). The case for published subjects. In *Proceedigs Identity, Reference, and the Web Workshop at the WWW Conference*, Edinburgh, Scotland. http://www.ibiblio.org/hhalpin/irw2006/spepper2.pdf (Last accessed on Oct. 5th 2008).

Ponte, J. M. (1998). *A language modeling approach to information retrieval*. Phd dissertation, University of Massachusets.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the Twenty-First Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, Melbourne, Australia.

Postel, J. (1982). IETF RFC 821 Simple Mail Transfer Protocol. http://www.ietf.org/rfc/rfc821.txt (Last accessed on March 12th 2008).

Postel, J. (1994). IETF RFC 1590 Media Type Registration Procedure. Category: Informational. http://www.ietf.org/rfc/rfc1590.txt (Last accessed on March 12th 2008).

Postel, J. and Reynolds, J. (1985). IETF RFC 959 File Transfer Protocol: FTP. http://www.ietf.org/rfc/rfc959txt (Last accessed on March 12th 2008).

Presutti, V. and Gangemi, A. (2008). Identity of resources and entities on the web. *International Journal of Semantic Web and Information Systems*, 4(2):49–72.

Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF. Recommendation, W3C. http://www.w3.org/TR/rdf-sparql-query/ (Last accessed March 13th 2008).

Putnam, H. (1975). The meaning of meaning. In Gunderson, K., editor, *Language, Mind, and Knowledge*. University of Minnesota Press, Minneapolis, Minnesota, USA.

Quillian, M. R. (1968). Semantic memory. In Minsky, M., editor, *Semantic Information Processing*, pages 216–270. MIT Press, Cambridge, Massachusetts, USA.

Quine, W. (1960). *Word and Object*. MIT Press, Boston, Massachusetts.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60:20–43.

Raggett, D., LeHors, A., and Jacobs, I. (1999). HTML 4.01 Specification. Recommendation, W3C. http://www.w3.org/TR/REC-html40/ (Last accessed March 7th 2008).

Rees, J. (2008). URI Documentation Protocol. Technical report, Science Commons. http://neurocommons.org/page/URI_documentation_protocol (Last accessed on Nov 20th 2008).

Reiter, R. (1978). On closed world data bases. In *Logic and Databases*. Plenum Publishing, New York City, New York.

Richens, R. H. (1956). Preprogramming for mechanical translation. *Mechanical Translation*, 3(1):20–25.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.

Robertson, S., Zaragoza, H., and Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 42–49, Washington, D.C., USA.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33:294–304.

Robertson, S. E. and Spärck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

Robertson, S. E., Walker, S., and Beaulieu, M. M. (1998). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 253–264, Gaithersburg, Maryland, USA.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126.

Rocchio, J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–32. Prentice-Hall, Inc., Uppder Saddle River, New Jersey, USA.

Russell, B. (1905). On denoting. *Mind*, 14:479–493.

R.V.Guha and D.Lenat (1993). Language, representation and contexts. *Journal of Information Processing*, 15(3).

Ryle, G. (1949). Meaning and necessity. *Philosophy*, 24:68–76.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sauermann, L. and Cygniak, R. (2008). Cool URIs for the Semantic Web. Technical report, W3C Semantic Web Interest Group Note. http://www.w3.org/TR/cooluris/ (Last accessed on Nov. 12th 2008).

Schank, R. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):532–631.

Schmidt-Schauss, M. (1989). Subsumption in KL-ONE is undecidable. In *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning*, pages 421–431, Toronto, Canada.

Scott, D. and Strachey, C. (1971). Toward a mathematical semantics for computer languages. Oxford programming research group technical monograph, Oxford University. PRG-6.

Searle, J. (1969). *Speech Acts*. Cambridge University Press, Cambridge, United Kingdom.

Searle, J. (1995). *The Construction of Social Reality*. The Free Press, New York City, New York, USA.

Shannon, C. and Weaver, W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press. Republished 1963.

Shapiro, S. (1979). The SNePS semantic network processing system. In Findler, N., editor, *Associative Networks: Representation and Use of Knowledge by Computers*. Academic Press, New York City, New York, USA.

Shvaiko, P. and Euzenat, J. (2005). A survey of schema-based matching approaches. *Journal of Data Semantics*, 4:146–171.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12.

Simondon, G. (1958). *Du mode d'existence des objets techniques*. Aubier, Paris, France. English Translation accessed on the Web at http://accursedshare.blogspot.com/2007/11/gilbert-simondon-on-mode-of-existence.html (Last accessed September 7th 2008).

Smith, B. C. (1984). Reflection and semantics in LISP. *Proceedings of 11th ACM SIGACT-SIGPLAN symposium on Principles of programming languages*, pages 23–35.

Smith, B. C. (1987). The correspondence continuum. Report, Center for the Study of Language and Information. no. CSLI-87-71.

Smith, B. C. (1991). The Owl and the Electric Encyclopedia. *Artificial Intelligence*, 47:251–288.

Smith, B. C. (1995). *The Origin of Objects*. MIT Press, Cambridge, Masschusetts, USA.

Smith, B. C. (1997). One hundred billion lines of C++. *LeHigh Cog Sci News*, 1(10). http://www.ageofsignificance.org/people/bcsmith/papers/smith-billion.html (Last accessed on March 1st 2008).

Smith, B. C. (2002a). The Foundations of Computing. In Scheutze, M., editor, *Computationalism: New Directions*. MIT Press, Cambridge, Massachusetts, USA.

Smith, B. C. (2002b). Reply to Dennett. In Clapin, H., editor, *Philosophy of Mental Representation*, pages 237–265. Clarendon Press, Oxford, United Kingdom.

Sollins, K. and Masinter, L. (1994). IETF RFC 1737 Functional Requirements for Uniform Resource Names. http://www.ietf.org/rfc/rfc1737.txt (Last accessed April 20th 2008).

Sowa, J. (1976). Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, 20(4):336–357.

Sowa, J. (1987). Semantic Networks. In Shapiro, S., editor, *Encyclopedia of Artificial Intelligence*, pages 1011–1024. Wiley and Sons, New York City, New York, USA.

Sowa, J. (2006). Review of language, cohesion, and form by margaret masterman. *Computational Linguistics*, 4(32):551–553.

Stevenson, M. and Wilks, Y. (1999). Large vocabulary word sense disambiguation. In Ravin, Y. and Leacock, C., editors, *Polysemy: Theoretical and Computational Contributions*, pages 161–177. Oxford University Press, Oxford, United Kingdom.

Stickler, P. (2005). CBD - Concise Bounded Description. Member Submission. http://www.w3.org/Submission/CBD/ (Last accessed March 13th 2008).

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: a core of semantic knowledge. In *In Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, Banff, Canada.

Suda, B. (2006). *Using Microformats*. O'Reilly. http://safari.oreilly.com/0596528213 (Last accessed Oct 12th 2008).

Tarski, A. (1935). The concept of truth in formalized languages. *Studia Philosophia*, 1:261–405. Reprinted in Logic, Semantics and Metamathematics (1956), Oxford University Press, Oxford United Kingdom, (1956), translated by J.H. Woodger.

Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, 4:341–375.

Thompson, H., Beech, D., Maloney, M., and Mendelsohn, N. (2004). XML Schema Part 1: Structures. Recommendation, W3C. http://www.w3.org/TR/xmlschema-1/ (Last accessed March 13th 2008).

Tsarkov, D. and Horrocks, I. (2003). DL reasoner vs. first-order prover. In *In Proceedings of the 2003 Description Logic Workshop (DL 2003)*, pages 85–94, Rome, Italy.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59:433–460.

van Assem, M., Gangemi, A., and Brickley, D. (2006). RDF/OWL Representation of WordNet. Editor's draft, W3C. http://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion (Last accessed Nov. 20th 2008).

van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworth & Co (Publishers) Ltd, London, UK, second edition.

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., and Studer, R. (2006). Semantic Wikipedia. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 585–594, Edinburgh, Scotland.

Vu, Q. M., Masada, T., Takasu, A., and Adachi, J. (2007). Using a knowledge base to disambiguate personal name in Web search results. In *Proceedings of the 2007 ACM symposium on Applied Computing*, pages 839–843, Seoul, Korea.

Wadler, P. (2003). The Girard-Reynolds Isomorphism. *Information and Computation*, 186(2):260–284.

Waldrop, M. M. (2001). *The Dream Machine: J.C.R. Licklider and the Revolution That Made Computing Personal*. Penguin, New York City, New York, USA.

Walsh, N. and Thompson, H. (2007). Associating resources with namespaces. TAG Finding. http://www.w3.org/2001/tag/doc/nsDocuments/ (Last accessed March 7th 2008).

Watts, D. and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, 6684(393):409–410.

Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press, Cambridge, Massachusetts, USA.

Wheeler, M. (2008). The Fourth Way: A comment on Halpin's 'Philosophical Engineering'. *APA Newsletter on Philosophy and Computers*, 8(1):9–12.

Whitelaw, C., Kehlenbeck, A., Petrovic, N., and Ungar, L. H. (2008). Web-scale named entity recognition. In *Proceedings of Conference on Information and Knowledge Management*, pages 123–132, Napa Valley, California.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, Massachusetts, United States.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74.

Wilks, Y. (2005a). A personal memoir: Margaret Masterman (1910-1986). In Masterman, M., editor, *Language Cohesion and Form*. Cambridge University Press, Cambridge, United Kingdom.

Wilks, Y. (2005b). Unhappy Bedfellows: the relationship of AI and IR. In Tait, J., editor, *Charting a new course: Natural Language Processing and Information Retrieval. Essays in honour of Karen Spärck Jones*. Kluwer, Amsterdam, Netherlands.

Wilks, Y. (2007). Karen Spärck Jones (1935-2007). *IEEE Intelligent Systems*, 22(3):8–9.

Wilks, Y. (2008a). The Semantic Web: Apotheosis of annotation, but what are its semantics? *IEEE Intelligent Systems*, 23(3):41–49.

Wilks, Y. (2008b). What would a Wittgensteinian computational linguistics be like? In *Proceedings of Convention for the Society for the Study of Artificial Intelligence and Simulation of Behavior*, Aberdeen, Scotland.

Winograd, T. (1972). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. *Cognitive Psychology*, 3(1).

Winograd, T. (1976). Towards a procedural understanding of semantics. Stanford Artificial Intelligence Laboratory Memo AIM-292.

Wittgenstein, L. (1921). *Tractatus Logico-Philosophicus*. Routledge, New York City, New York, USA. Republished 2001.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell Publishers, London, United Kingdom. Republished 2001, translated by G.E.M. Anscombe.

Woods, W. (1975). What's in a link: Foundations for semantic networks. In *Representation and Understanding: Studies in Cognitive Science*, pages 35–82. Academic Press, Inc., Orlando, Florida, USA.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the Nineteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, pages 189–196, Cambridge, Massachusetts, USA.

Zhai, C. and Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the Twenty-Fourth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, New Orleans, Louisiana, USA.

Zimmerman, H. (1980). The ISO model of architecture for Open Systems Interconnection. *IEEE Transactions on Communications*, 28(4):425–432.

Zipf, G. (1949). *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, New York City, New York, USA.

Zurawski, M., Smaill, A., and Robertson, D. (2008). Bounded ontological consistency for scalable dynamic knowledge infrastructures. In *Proceedings of the Asian Semantic Web Conference (ASWC2008)*, pages 212–226, Bangkok, Thailand.