LinkSuite™: formally robust ontology-based data and information integration

Werner Ceusters[a], Barry Smith[b], James Matthew Fielding[a,b]

[a] Language & Computing nv (L&C), Hazenakkerstraat 20a, B-9520 Zonnegem, Belgium
[b] Institute for Formal Ontology and Medical Information Science, University of Leipzig, Härtelstrasse 16-18, 04107 Leipzig, Germany

forthcoming in DILS 2004 (Database Integration in the Life Sciences), Berlin: Springer

Abstract. The integration of information resources in the life sciences is one of the most challenging problems facing bioinformatics today. We describe how Language and Computing nv, originally a developer of ontology-based natural language understanding systems for the healthcare domain, is developing a framework for the integration of structured data with unstructured information contained in natural language texts. L&C's LinkSuite™ combines the flexibility of a modular software architecture with an ontology based on rigorous philosophical and logical principles that is designed to comprehend the basic formal relationships that structure both reality and the ways humans perceive and communicate about reality.

Introduction

The integration of information resources in the life sciences is one of the most challenging problems facing bioinformatics today [1]. Researchers are flooded with information from a variety of sources and in a variety of formats, ranging from raw lab instrument data, gene expression profiles, raw sequence traces, chemical screening data, and proteomic data, to metabolic pathway models and full-fledged life science ontologies developed according to a myriad of incompatible and typically only loosely formalized schemas. Ultimately, if the robust integration of the information deriving from all of these sources is to be possible at all, tools for the formal analysis of these different types of data within a single consistent framework will have to be supplied. As a step in this direction, we describe here a methodology for information integration that is able to manipulate information scattered over many data stores while supporting a single view across the whole. The methodology can handle data stores that are owned by different organizations and located physically in different places. It can support the integration of data that are inherently heterogeneous in nature, including structured data stored in relational databases as well as data that is semi-structured via XML or HTML hyperlinking. Most importantly, it can comprehend data that is totally unstructured, including collections of text documents such as clinical discharge summaries as well as scientific journal articles.

In realizing this methodology Language and Computing nv (L&C) is working towards a framework for data-, information- and ontology-integration across all levels of generalisation and including equally information in both structured and unstructured forms. We believe that, given the complexity of the problem, the ultimate solution will be arrived at only if at least the following three tasks are dealt with in an appropriate way:

1. identifying the basic ontological foundations of a framework expressive enough to describe life science data at all levels;
2. carrying out the research in information engineering needed to create technology able to exploit this ontological framework in a way that can support the integration of massively heterogenous structured and semi-structured life science databases;
3. developing the tools for natural language understanding in the domain of the life sciences needed to extract structured data from free text documents.

L&C's LinkSuite™ environment reflects an on-going effort to implement the philosophically sound top-level ontology developed by the Institute for Formal Ontology and Medical Information Science in Leipzig [2, 3, 4]. LinkSuite™ consists of a number of mutually complementary modules, each addressing one or other of the mentioned tasks sufficiently successfully to have been granted from the technology watchers Frost and Sullivan the Healthcare Information Technology and Life Sciences Product of the Year Award during the Global Excellence in Healthcare & Life Sciences Awards Banquet in San Diego in November 2003 [5].

We first elaborate on the three tasks mentioned above. We then provide a description of the LinkSuite™ system, and finally we motivate our design choices and future directions of our research.

Key requirements for ontology-based information integration

Information integration can be realised only adequately if a number of requirementsare satisfied. These fall into three categories: requirements for the sort of ontology used, requirements for the integration of structured data, and requirements for making information in text documents machine readable. We'll discuss each of these in detail.

Basic ontological foundations for life science information integration

Ontology is currently perceived in many circles as providing the necessary starting point for a solution to the problem of information integration both from a domain-independent perspective [6] as also in the specific field of bio-informatics [7]. We believe that ontologies can support the sort of reasoning power that is required both to optimize data-extraction from text corpora and also to optimize data-integration from a variety of disparate sources only if they rest on powerful formal-logical tools [8]. In the longer term such ontologies can also enable reasoning with the data that results from integration in ways that can open the way for large-scale hypothesis checking.

But one problem continues to stand in the way of achieving these ends: terminology-oriented life science databases marked by fundamental logical inadequacies continue to evolve and expand even while incorporating ambiguities and inconsistencies with respect to such basic ontological relationships as *is-a* and *part-of* [9]. These ambiguities and inconsistencies, which result from the lack of a standard unified framework for understanding the basic relationships that structure our reality, are an obstacle to database integration and thus to the sort of computer processing of biomedical data which is the presupposition of advance in the bio-informatics field.

To rectify these problems, both formal-ontological and meta-ontological theories are required. Formal ontology is needed to provide life science applications with a set of standardized formal definitions of basic categories and relations, including the resources to deal with dependent and independent entities and with occurrents and continuants. It must have the resources to deal adequately also with the oppositions between functions and realizations (both normal and mutant genes may share the same function, but only the former can participate in those processes which are the realization of this function), and between universals and particulars (*malaria* as disease class described in textbooks versus *malaria* as particular instance of this class in this particular patient). It must also provide meta-ontological theories such as the theory of Granular Partitions [10] designed to allow navigation between ontologies in ways which can be exploited by reasoning engines. By disambiguating the terms used in the often ontologically ill-formed definitions present in most existing terminologies (cf. the misclassifications of constituents, processes and functions in the Gene Ontology [11, 12, 13] or the ontological mistakes in SNOMED-CT [14]), these formalizations may also aid in the passage of information between users and software agents. They will also help to improve consistency, both with and between ontologies, as well as contributing to the reliability of terminology curation.

Integration of (semi-)structured life science databases

Given the basic ontological framework described above, our idea is to allow external databases (EDBs) to connect dynamically to the LinkSuite™ framework in such a way that all the implicit and explicit relationships between the data within each EDB are mapped onto relationships within the base ontology, and that the logical structure of the latter, together with the associated reasoning paper, are thereby propagated through the entire system of EDBs: the databases can be browsed and relationships between them established as if all the data were part of a single base ontology. To this end, the expressive resources of the base ontology must be sufficiently rich that we can map onto it both whole database columns and cell record data in such a way that not only the meta-data but also the incorporated instance data of the EDBs are properly apprehended. We require also that:

1. The mapping should not change the actual state of the data in either the base ontology or the EDBs, meaning that individual databases can be coupled and de-coupled at will without the mapping between the remaining EDBs and the base ontology becoming inconsistent.
2. The flow of EDB data to the ontology should occur in real-time, so that the EDBs do not need to be pre-processed in any way. The only manual intervention should

be in the provision of an initial description of the structure of the database in a form that enables the latter to be mapped onto the base ontology in the appropriate way. Once this is realised, the database is dynamically mapped in such a way that all the data it contains becomes automatically accessible through the base ontology even when updates (at least those updates that do not alter the structure of the database) subsequently appear.

3. The EDBs should continue to interoperate with external applications in the same way as they did before the mapping was effected.

Working with textual information resources

At least 95% of the life science information currently publicly available resides in journals and research reports in natural language form. Even in hospitals that use an electronic medical record system, the majority of the data resides in electronic reports written in free text. Research devoted to making accessible this information has focused thus far on techniques such as document indexing and extraction of simple data elements such as dates, places, names or acronyms. Very few attempts have been made to use such techniques to obtain ontologically relevant information, for example to use automatic analysis of text documents to trigger requests for updating of ontologies such as GO or SNOMED-CT. This requires text data mining mechanisms combining statistical, linguistic and knowledge-based processing working together to overcome the problems intrinsically associated with each type of mechanism taken separately. Statistical approaches have to rely not only on a very large set of domain-related documents, but also on a reliable corpus representing non-domain-specific language usage that is needed in order to filter out what is statistically relevant in the context of the specific domain in question. Linguistic approaches can work only for processing documents in languages for which the necessary *lingware* (lexicons, grammars, machine-readable dictionary resources such as WordNet [15], and so forth) is available [16]. Such tools are certainly available for English, though not for all specialised domains, and they are available only to a limited degree for other languages. Hence a combination of different approaches is necessary, and in such a way that they complement each other mutually [17].

The LinkSuite™ platform

L&C's products are based upon research initiated in the late 1980s with the objective of developing applications for natural language understanding in the healthcare domain. Ontology was identified already early on as a key presupposition of success in this regard, and the need for ontology as a language-independent representation of reality was generally well accepted in both the medical informatics [18] and natural language processing communities [19], as also was the usefulness of s*ituated ontologies,* i.e. ontologies that are developed for solving particular problems in specialised domains [20]. However, our experience and research convinces us that ontologies that have to operate with natural language processing applications are

better suited to assist language understanding when the concepts and relations used are linguistically motivated [21]. This requirement is less important if structured data are only viewed using conventional browsers.

For these reasons, L&C has built its NLU (Natural Language Understanding) technology around a medico-linguistic ontology called LinkBase®, authored using the in-house ontology management system LinkFactory®. The applications that use these resources include TeSSI®, FreePharma® and L&C's Information Extraction Engine. In addition, L&C's OntoCreator is an NLU application that is designed to feed into LinKBase® information drawn from text documents, while MaDBoKS® feeds into LinkBase® instance data drawn from external databases. All components are developed using L&C's workflow architecture, which allows them to be plugged in or out dynamically wherever they are needed. To meet normal industrial standards of good practice for software engineering, all components are developed under a strict quality assurance process which is itself supported by its own quality assurance software and embraces product versioning, and a system for tracking and resolving errors.

LinKFactory®

LinKFactory® [22] is the proprietary L&C environment used for creating and modeling ontologies. L&C uses LinKFactory® for maintaining LinKBase®, the L&C medical ontology. This tool can use various database management systems, including Oracle and Sybase and is developed using a three-tier client-server architecture [23].

The first tier of the program runs on the user's workstation and is called the LinKFactory® Client. This tier contains a layout manager with which the user can define different frames into which he can load modules that are referred to as *beans*. Beans are Java user interfaces that facilitate communication between the user and the application and provide a visual representation of some selected portion of the underlying ontology. Examples of available beans are *concepttree* and *fulldeftree*. Beans can be assembled in a layout and linked to one another and share information by event spawning, as when the selection of a concept in the *concepttree* tells the *fulldeftree* to show the information associated with that concept in the *fulldeftree* format. The use of beans allows L&C easily to expand the functionality of LinKFactory® without disturbing the functions already supplied. Sixteen beans have been defined thus far.

The second tier of LinKFactory® runs on a server and contains the actual business logic: this tier knows what actions to perform when a user clicks a button or types in a term in a specific data capture field on his user-interface. The first and second tiers communicate through the LinKFactory® Server Interface (LSI), which can be seen as a high level API using Java RMI (remote method invocation). The LSI allows the LinKFactory® Client to pass high level requests such as 'add concept' and 'get concept' to the LinKFactory® Server's business logic layer. This tier translates a high level request (such as 'add concept') into a series of actions. In addition, it also performs authorization checks to see if the user is allowed to perform the requested actions.

The series of actions initiated by a request such as 'add concept' does not depend on any specific relational database platform because of the third tier: the data access

layer, which translates Data Access Objects into relational tables containing the LinKBase® medical ontology and SQL query templates. The LinKFactory® program has been written in Java. Thus it too operates in a system-independent way and is ready to perform in a distributed network environment.

LinKBase®

LinKBase® contains over 2 million language-independent medical and general-purpose *domain-entities*, representing universals and particulars in the sense of Aristotelian ontology [24]. As such, domain-entities abstract away from the specific features of natural language representations, fulfilling to that end the same function as *concepts* in other terminologies or ontologies. They are however not equivalent to concepts, since they represent not abstractions from how humans think about real-world entities, but rather the entities themselves to which such thoughts are directed. The concepts in people's minds are clearly separated from the LinkBase® ontology proper by being represented as what are called *meta-entities*, a category which is included also in order to allow mappings to third party terminologies and ontologies. Domain-entities are associated with more than 4 million terms derived from a number of different natural language sources [25]. A *term* in this connection is a sequence of one or more *words*, which may be associated with other concepts in their turn. Domain-entities are linked together into a semantic network in which some 480 link types are used to express different sorts of relationships. The latter are derived from formal-ontological theories of mereology and topology [26, 27], time and causality [28], and also from the specific requirements of semantics-driven natural language understanding [19, 29]. Link types form a multi-parented hierarchy in their own right. At the heart of this network is the formal subsumption (*is-a*) relationship, which in LinKBase® covers only some 15% of the total number of relationships involved. Currently, the system is being re-engineered in conformity with the IFOMIS theories of Granular Partitions [10] and Basic Formal Ontology [30, 31].

MaDBoKS

The MaDBoKS (Mapping Databases onto Knowledge Systems) tool is an extension of Linkfactory® constructed to enable external relational databases to be connected to LinkBase® in the manner described above [32]. Database schemas from existing databases, for example from hospital patient databases or electronic patient records, can be retrieved and mapped to the ontology in such a way as to establish a two-way communication between database and ontology. The latter thereby comes to serve as a central switchboard for data integration, so that the database schemas themselves function as semantic representations of the underlying data (analogous to the semantic representations of natural language utterances that are yielded through processing by natural language understanding software). In an NLU system, a semantic parser bridges the gap between the ontology and the documents from which information is to be extracted. Here an analogous piece of software, called a *mediator*, bridges the gap between the ontology and the databases to be integrated [33]. L&C has thus far been able to successfully integrate the Gene Ontology [11], Swiss-Prot [34] and the Taxonomy database of the National Center for Biotechnology Information [35] using this approach.

TeSSI®: Terminology Supported Semantic Indexing

TeSSI® is a software application performing semantic indexing. TeSSI® first segments a document into its individual words and phrases. It then matches words and phrases in the document to corresponding LinKBase® domain-entities [36]. This step introduces ambiguity, since some entities share homonymous terms: e.g. the word *ventricle* can be used in a text to denote a *cardiac ventricle* or a *cerebral ventricle*; the phrase *short arm* may equally well denote a reduction anomaly of the upper limb or a part of a chromosome. To resolve cases of ambiguity, TeSSI® uses domain knowledge from LinkBase® to identify which domain-entity out of the set of domain-entities that are linked to a homonymous word or phrase best fits with the meaning of the surrounding words or phrases in the document.

TeSSI® then uses the matches between words and phrases identified in the document and the domain knowledge in LinkBase® to infer additional domain-entities which are only implicitly part of the subject matter of the document. The end result of this process is a graph structure whose nodes correspond to the LinkBase® domain-entities explicitly or implicitly present in the document and whose arcs correspond respectively to 1) semantic relationships derived from the LinkBase® domain ontology and 2) co-occurrence relationships derived from the position of terms in the document. The inclusion of the latter is motivated by many studies showing that co-occurring terms are likely to be also semantically related [37]. Nodes are weighted according to the number of occurrences of the corresponding terms in the document. Arcs are weighted according to the semantic distance between the corresponding entities in LinKBase® and according to the proximity of the corresponding terms in the document.

Having identified all the medical (and non-medical) terms in a document, TeSSI® then ranks the corresponding domain-entities in the order of their relevance to the document as a whole, thus identifying the topics (main subjects) of the document. Relevance scores are on a scale of 0 to 100, with 100 representing the most relevant doman-entity. To determine these scores, TeSSI® uses a constraint spreading activation algorithm on the constructed graph [38]. In this way, semantically related domain-entities reinforce each other's relevance rankings. The rationale for this algorithm stems from the observation that the domain-entities referred to by terms in any particular document will vary in their degree of semantic independence from each other. For example, a document might contain one mention each of the terms "heart failure," "aortic stenosis," and "headache", the first two being clearly more closely related to each other than to the third. An indexing system based entirely on term frequency will treat these three terms independently, thus assigning them all the same relevance. Intuitively, however, the document has twice as many mentions of heart disease as of headache. TeSSI® takes advantage of its underlying medical ontology in order to represent more accurately this type of phenomenon.

L&C's Information Extraction System

The L&C Information Extraction System consists of a number of components that successively add structured information to an unstructured text. The system takes a text document in natural language as input and creates an XML document as output. The latter then serves as the basis for further user-defined operations including querying and template-filling. As such, the information extraction system itself is an

essential component in a query answering system. The XML output is created via a natural language processing procedure involving the use of Full Syntactic Parsing for syntactic analysis and LinKBase® for analysing semantics. In addition Text Grammar Analysis (TGA) is applied, which means that the system looks for relations (such as *summarization*, *elaboration*, *argumentation*, *exemplification*, and so forth) between parts of text. We believe that it is only through TGA carried out on top of syntactic and semantic analysis of individual sentences that a full understanding of the meaning of text in natural language will be possible in the future.

The basic components of the L&C information extraction process are:

Segmentation
Section Labeling
Clause/Phrase Segmentation
Fragment Labeling
Information Extraction,

We deal with each of these in turn.

The input text is first segmented into paragraphs and sentences. Each sentence is then decomposed into its basic constituents, which are tagged with markers for syntactic and semantic information. Segmentation uses rules easily adaptable to the client's particular document requirements. An important step in the process is *section segmentation* carried out at whole document level rather than sentence level. A text is not an unordered succession of separate chunks of data. Rather it is a structured whole, in which each piece of information enters in at a certain functionally appropriate stage. Recognizing the different sections in a text is thus important for getting at its meaning. As an example, the first sections of this paper are: title, authors, affiliation, and abstract. In medical discharge summaries, typical initial sections are: patient-related administrative data, anamnesis, clinical findings, and so forth.

Each section is automatically assigned a label that reflects the context of the information that the section contains. Labeled sections are used to limit the scope of search when looking for particular information to be extracted. For example, discharge medication will only be looked for in sections in which discharge medication is known to appear. Labeling is based on labeling rules gathered from a training corpus that take into account a number of weighted features. Users of our system can choose whether they want to adopt an existing training corpus or create their own. If they choose the latter they are supplied with a fully customizable basic set of possible section labels with their descriptions. A graphical user interface for labeling texts is included with the system. It can be used to first label manually a training corpus that then serves as input for a supervised learning algorithm which generates in turn the rules to label similar texts automatically. Labeled sections are used to limit the scope of search when looking for particular information to be extracted. The accuracy of the labeler for medical discharge summaries amounts currently to 97.23% (tested on 4421 sections in 100 medical reports); this is an increase from the level of 96.4% achieved in 2001 [39].

Sections consist of sentences, and each sentence can be divided in its turn into clauses and phrases. To effect this division we use our Full Syntactic Parser, a hybrid system combining both symbolic and statistical approaches. We use a dependency grammar-based formalism to capture the syntactic relations between the words in a

sentence, which enables us for instance to capture immediately the scope of negated elements in a sentence.

The different fragments that are recognized by the Clause/Phrase Segmenter with embedded Full Syntactic Parser receive a functional label – such as "clinical finding" or "diagnosis" – according to their content. The Fragment Labeler uses the same techniques as the Section Labeler and thus also needs to be built up by means of a training corpus, which again can be provided either by L&C or by the client. Fragment label information is used to further narrow down the amount of text in which information will be searched for.

The Information Extraction component uses information from the Section Labeler and the Fragment Labeler, as well as conceptual information from LinKBase®, syntactic information from our Full Syntactic Parser, and novel machine-learning techniques, which in combination go much further than standard text analysis algorithms relying on string matching and similar techniques.

FreePharma®

L&C developed a novel approach to formally representing and managing the information present in medication prescriptions: FreePharma® [40], The input to which is constituted by free text medication prescriptions. The latter are first parsed syntactically using full syntactic parsing aided by semantic disambiguation and statistical reinforcement: if a pure syntactic parse leads to many possible solutions, semantics and statistics are used to prune the parse tree. The syntactic parser uses semantico-syntactic labels to represent the relations between the terms in the medication prescriptions and uses various statistics to decide which analysis is the most probable. The XML-structure generated by the parser is then mapped onto a standard, pre-defined XML-template by means of semantic knowledge from a medical ontology for disambiguation and semantic slot analysis. The output of the system is thus an XML message providing a structured representation of the extracted (and initially unstructured) prescription information.

Using this formalism, we are able to gather the required information from natural language text. Because the system is not limited to structured text input, this greatly improves its flexibility and applicability. Its hybrid syntactic, semantic and statistical approach allows the system to deal with highly complex medication prescriptions.

A randomly selected corpus of 300 prescriptions, with a large coverage of possible prescription formats – including decreasing and increasing doses, as well as tapering doses and conjunctions of doses – yields a syntactic recall of 98.5%, with a syntactic precision of 96.2%. The precision of the final semantic representation amounts to 92.6%. (*Precision* here means the percentage of syntactic or semantic labels correctly assigned to terms and/or phrases in the prescriptions within the population of all the labels assigned. *Recall* refers to the percentage of correctly assigned labels with respect to the number of labels that should have been assigned.)

OntoCreator

Since so much life science information resides in journals and research reports in natural language form, it is worthwhile to develop a methodology, algorithms and software implementations that enable us to derive life-science data from free text

documents and to use data extracted from these sources also in developing situated ontologies along the lines described above. OntoCreator is L&C's first and still modest attempt automatically to produce raw ontologies that can subsequently be validated and edited by users using the facilities of LinkFactory®. The module exploits the machinery described above to combine both statistical and linguistic text analysis techniques to produce raw ontologies out of text repositories covering the life sciences.

OntoCreator consists of a set of components that enable the user to analyze documents in various languages, to access and modify ontologies that are already mapped to LinKBase® and to construct graphical interfaces for accessing and editing the extracted data.

Its functionalities include the ability to:

1. extract domain-relevant terms that can be added to terminology lists, including terms not known in advance to stand in any relationship to the terms already processed;
2. propose such terms as representing domain-entities either already recognized or needing to be added to those already existing in the ontology;
3. extract semantic relationships between terms in the documents analyzed and add corresponding relations to the ontology;
4. submit the extracted terms with a relevance weight and possible semantic relationships in the form of XML documents;
5. enable the user to edit the results of the automatic ontology extraction.

Related work

In [41], 53 ontology authoring systems were reviewed. Only three systems were reported to combine the functionalities of multi-user authoring, information extraction, merging of distinct ontologies and lexical support (the latter being defined by the reviewers as "*capabilities for lexical referencing of ontology elements* (*e.g., synonyms*) *and processing lexical content, e.g., searching/filtering ontology terms*)", all features that are necessary to develop and maintain the very large ontologies that will be vital to future biomedical research. Of the three, LinKFactory® is the only commercial system, the two others being OntoBuilder from the University of Savoy [42], and WebOde from the Technical University of Madrid [43]. The latter however resorts to synchronisation methods to allow multi-user access [44], and such methods are insufficient to prevent inconsistencies when two or more users are working with the same data at the same time. In addition, almost all ontology systems reviewed lack the resources to deal not only with *classes* but also with individual *instances*, i.e. entities bound to specific locations in space and time [31, 45,]. If, however, we are to incorporate instance-based data in a framework for biomedical ontology integration, then an ontology management system must go beyond what Brachman called the T-Box (of classes, or general concepts) [46] (which has served hitherto as the main focus of almost all researchers in our field) and take account also of the A-Box (containing data pertaining to the individual instances of such classes in spatio-temporal reality).

In [47] LinKBase® was reported to be the largest (in terms of number of domain-entites) medical terminology system available worldwide.

[48] describes a system that comes very close to MaDBoKS® and that is specifically designed to mediate between structured life-science databases using a much smaller ontology than LinKBase®. As is also the case for Tambis [49], however, this system is not intended to work with free-text document-based resources. Data integration in the system described in [48] is also limited to the external database schemas and does not take into account cell data.

Relevant on-going research in the combined use of structured and unstructured information sources is being conducted under the auspices of the European-funded ORIEL-project, but the currently available literature does not make it possible to assess the results obtained thus far [50].

Integrating biomedical ontologies

It is for us no surprise that Kalfoglou and Schorlemmer, after having reviewed 35 systems for their ontology mapping capacities, conclude that ontology mapping still "*faces some of the challenges we were facing ten years ago when the ontology field was at its infancy. We still do not understand completely the issues involved*." [51]. Many researchers seem to forget that ontology is a discipline that was in its infancy not 10 but rather some 2400 years ago, when the seminal ideas of Aristotle on categories, definitions, and taxonomies were first presented – ideas which can now be seen to have enjoyed an astonishing prescience. In our view, applying philosophical and logical rigour in a way which builds on the type of realism-based analysis initiated by Aristotle is the only way to provide a coherent and unified understanding of the basic ontological distinctions required to successfully integrate the diverse domain-specific terminologies that have grown up in uncontrolled fashion in the separate parts of the biomedical informatics community [2]. Integration of heterogenous biological resources (instance-level) and integration and re-usability of bio-ontologies (class-level) are indeed the most important challenges facing the life sciences today [52]. Hence the importance of dynamic techniques such as those described above to integrate external databases with a domain-ontology.

Philosophical rigour must be applied in two equally essential dimensions. The first is in setting up the base ontology to be used as framework for integrating the separate external databases. The second is in calibrating the ontologies used in these external databases in terms of the categories and relations supplied by the base ontology.

Ontology-like structures, such as the Gene Ontology [11] and SNOMED-CT [14], are 'controlled vocabularies', i.e. they have a clean syntactic structure, which is often mistaken for a semantic structure. They consist of systems of concepts joined together via binary relations such as *is-a* and *part-of*. For the most part however, these concepts and relations are formulated only in natural language and are used in a variety of inconsistent ways even within a single ontology, and this sets obstacles in the way of ontology alignment [53]. To define a robust common structure in which ontological elements from such information resources may be mapped [2, 8, 54] thus

constitutes the first dimension of philosophical rigour in the enterprise of life-science ontology integration.

The second dimension of rigour turns on the fact that mapped elements of external ontologies inherit the logical structure in which the entities and relations of the base ontology are defined and axiomatized. In this way the rigour of the base ontology is imported into external ontologies from the outside. This importation is meta-ontological, in the sense that changes designed to bring about consistency are made not directly within the external database itself, but rather via corresponding adjustments in its representation within the base ontology. This method makes it possible for us to navigate between ontologies derived from distinct external sources in consistent fashion even when the latter are not themselves consistent.

We do not thereby resolve the inconsistencies and other problems within the external ontologies themselves. While many of these problems are eliminated, or at least ameliorated, through the adoption of an approach like the one presented here, i.e. via the imposition of clear formal-ontological distinctions, it is not our intention to remodel existing databases to reflect such distinctions. Each terminology has its own purposes and advantages, and from the LinkSuite™ perspective the task of integrating the corresponding ontologies involves focusing precisely on *integration*, and not on that of *assimilation* – drawing hereby on the fact that we can attain the desired degree of consistency necessary to map these databases onto each other (by going always through the base-ontology) and adding structural information at the meta-ontological level without actual changes in the external databases themselves.

Although we have already come far, much remains to be done, especially in the area of automatic extraction of situated-ontologies from free text document collections in domains for which no formal ontology with adequate coverage thus far exists. The essential steps to be performed are: first, that of identifying terms and phrases in text documents that represent entities instantiating ontological categories such as functions and roles or dependent and non-dependent entities; and second, assessing whether or not the entities thereby found are already part of the ontology as thus far developed. Clearly these two steps must be performed in parallel, and thus some approach like that of agent-based parallel processing must be adopted, in which each agent can operate independently from the others yet is constantly generating information useful to the latter on the basis of processing that has been effected thus far while at the same time also constantly looking out for information generated by these other agents that might improve its own processing. There need to be agents that extract terms from documents that can be added to terminology lists, relate terms to already existing entities in the ontology, extract semantic relationships between entities, and request assistance from a user to assess the validity of results when no automatic mechanism can be called upon. Moreover, research must be focused around a global strategy for managing the sorts of ambiguity and uncertainty which are inevitably introduced at each decision step when an agent is deriving structured information from unstructured texts.

Conclusion

We have described a series of problems which arise in the study of life science ontologies, terminologies and databases, and we have sketched the design of a platform that is able to deal with them appropriately. Most problems encountered illustrate a general pattern, present in some form or another in all existing biomedical ontologies. The latter are, when assessed from the perspective of what can be achieved when an appropriate degree of formal-ontological rigour is imposed from the start, conspicuously *ad hoc* (this is the main cause of the Tower of Babel problem in current biomedical research). This *ad hoc* character is not without its history: those engaged in terminology research were forced, during the initial stages of moving from printed dictionaries and nomenclatures to digitalized information resources, to make a series of decisions about complex ontological issues – indeed about the very same issues that philosophers have pondered for millennia – at a time when the ontological nature of these issues was still not clear. To date, the importance of philosophical scrutiny in software application ontologies has been obscured by the temptation to seek immediate solutions to apparently localized problems. In this way, the forest has been lost for the trees, and the larger problems of integration have been rendered unsolvable. *Ad hoc* solutions have fostered further *ad hoc* problems.

Our research thus far, and its embodiment in LinkSuite™, constitutes a convincing demonstration of the increased adaptability that can be gained through the application of philosophical knowledge and techniques. If this success is any indicator, we have great reason to expect that further research will greatly enhance our ability to effect direct integration of further, larger and even more complex terminologies.

References

1 Kirsten T, Do H, Rahm E. Data Integration for Analyzing Gene Expression Data. *Proc. 2nd Biotech Day*, University of Leipzig, May 2003. 88-89.

2 Smith B., Ceusters W.: Towards Industrial-Strength Philosophy; How Analytical Ontology Can Help Medical Informatics. *Interdisciplinary Science Reviews*, 2003, 28: 2, 106-111.

3 Ceusters W, Smith B, Van Mol M. *Using ontology in query answering systems: scenarios, requirements and challenges*. In Bernardi R, Moortgat M (eds) Proceedings of the 2nd CoLogNET-ElsNET Symposium, 18 December 2003, Amsterdam, 5-15.

4 Fielding JM, Simon J, Ceusters W, Smith B. *Ontological Theory for Ontology Engineering*. In Proceedings of The Ninth International Conference on the Principles of Knowledge Representation and Reasoning, 2004 (in press)

5 Frizzell, J.: Frost & Sullivan Recognizes Language & Computing with Product of the Year Award, Nov 5, 2003. (http://awards.frost.com/prod/servlet/press-release.pag?docid= 7868843&ctxixpLink=FcmCtx1&ctxixpLabel=FcmCtx2)

6 Partridge, C.: The Role of Ontology in Integrating Semantically Heterogeneous Databases. Technical Report 05/02, LADSEB-CNR Padova, Italy, June 2002

7 Lambrix, P., Habbouche, M. and Pérez, M.: Evaluation of ontology engineering tools for bioinformatics. *Bioinformatics* 19: 12, 1564-1571, 2003.

8 Ceusters W, and Smith B.: Ontology and Medical Terminology: why Descriptions Logics are not enough. *Proceedings of the Conference Towards an Electronic Patient Record* (TEPR 2003), San Antonio, 10-14 May 2003 (electronic publication).

9 Smith B. and Rosse C.: The role of foundational relations in the alignment of biomedical ontologies, *Proceedings of Medinfo*, San Francisco, 7-11 September 2004, in press.
10 Bittner, T. and Smith, B.: A Theory of Granular Partitions. In: *Foundations of Geographic Information Science*, Matthew Duckham, Michael F. Goodchild and Michael F. Worboys (eds.), London: Taylor & Francis Books, 2003, 117-151.
11 Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* 25:25-29, 2000.
12 Smith B, Williams J and Schulze-Kremer S, 2003, The Ontology of the Gene Ontology, in *Biomedical and Health Informatics: From Foundations to Applications*, Proceedings of the Annual Symposium of the American Medical Informatics Association, Washington DC, November 2003, 609-613.
13 Smith B, Köhler J and Kumar A, On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology, in this volume.
14 Ceusters W., Smith B., Kumar A. and Dhaen C.: Mistakes in Medical Ontologies: Where Do They Come From and How Can They Be Detected? in Pisanelli DM (ed). *Ontologies in Medicine. Proceedings of the Workshop on Medical Ontologies, Rome October 2003*, Amsterdam: IOS Press, 2004 (in press).
15 Fellbaum, C. (ed.): *WordNet: An Electronic Lexical Database*. Cambridge MA: MIT Press, (1998).
16 Bod, R. and Scha, R.: Data-oriented language processing in: Young, S, and Bloothooft, G. (eds.), *Corpus-Based Methods in Language and Speech Processing*, Kluwer Academic Publishers, 137-173, 1997
17 Widdows, D.: Unsupervised methods for developing taxonomies by combining syntactic and statistical information. *Proceedings of HLT-NAACL 2003*, 197-204
18 Rector, A.L., Rogers, J.E. and Pole, P.: The GALEN High Level Ontology. In Brender J, Christensen JP, Scherrer J-R, McNair P (eds.) *MIE 96 Proceedings*. Amsterdam: IOS Press 1996, 174-178.
19 Bateman, J. A.: Ontology construction and natural language. *Proc Int Workshop on Formal Ontology*. Padua, Italy, 1993: 83-93.
20 Mahesh K. and Nirenburg S.: A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*. Montreal, Canada, 1995.
21 Deville, G, and Ceusters, W.: A multi-dimensional view on natural language modelling in medicine: identifying key-features for successful applications. Supplementary paper in *Proceedings of the Third International Working Conference of IMIA WG6*, Geneva, 1994.
22 Ceusters, W., Martens, P., Dhaen, C., Terzic, B.: LinKBase: an Advanced Formal Ontology Management System. Interactive Tools for Knowledge Capture Workshop, *KCAP-2001, October 2001, Victoria B.C., Canada* (http://sern.ucalgary.ca/ksi/K-CAP/K-CAP2001/).
23 Sadoski, D.: Client/Server Software Architectures--An Overview. http://www.sei.cmu.-edu/str/descriptions/clientserver_body.html, 2003.
24 Burkhardt, H. and Smith, B. (eds.): *Handbook of Metaphysics and Ontology*. Philosophia, Munich, Germany, 2 vols. (1991).
25 Montyne, F.: The importance of formal ontologies: a case study in occupational health. *OES-SEO2001* International Workshop on *Open Enterprise Solutions: Systems, Experiences, and Organizations, Rome, September 2001* (http://cersi.luiss.it/oes-seo2001/papers/28.pdf).
26 Smith, B. and Varzi A.C.: Fiat and bona fide boundaries, *Proc COSIT-97*, Berlin: Springer. 1997: 103-119.
27 Smith, B.: Mereotopology: a theory of parts and boundaries, *Data and Knowledge Engineering* 1996; 20: 287-301.
28 Buekens, F., Ceusters, W. and De Moor, G.: The explanatory role of events in causal and temporal reasoning in medicine. *Met Inform Med* 1993; 32: 274-278.

29 Ceusters, W., Buekens, F., De Moor, G. and Waagmeester, A.: The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition. *Met Inform Med* 1998; 37(4/5): 327-33.
30 Fielding, J. M., Simon, J. and Smith. B.: Formal Ontology for Biomedical Knowledge Systems Integration. Manuscript (http://ontology.buffalo.edu/ medo/FOBKSI.pdf).
31 Grenon, P. and Smith, B.: SNAP and SPAN: Towards dynamic spatial ontology, forthcoming in *Spatial Cognition and Computation*.
32 Verschelde, J. L., Casella Dos Santos, M., Deray, T., Smith, B. and Ceusters, W.: Ontology-assisted database integration to support natural language processing and biomedical data-mining, under review.
33 Wiederhold, G. and Genesereth, M.: "The Conceptual Basis for Mediation Services (ps of Source)"; *IEEE Expert*, 12: 5, Sep.-Oct. 1997.
34 Boeckmann B., Bairoch A., Apweiler R., Blatter M., Estreicher A., Gasteiger E., Martin M. J., Michoud K., O'Donovan C., Phan I., Pilbout S., and Schneider M.: The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31: 365-370 (2003).
35 Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A. and Rapp, B. A.: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000 Jan 1; 28(1):10-4 (2000)
36 Jackson, B. and Ceusters, W.: A novel approach to semantic indexing combining ontology-based semantic weights and in-document concept co-occurrences. In Baud R, Ruch P. (eds) *EFMI Workshop on Natural Language Processing in Biomedical Applications, 8-9 March, 2002, Cyprus*, 75-80.
37 Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407 (1990).
38 Hendler, J. A.: Marker-Passing over Microfeatures: Towards a Hybrid Symbolic-Connectionist Model. *Cognitive Science* 1989 (1) 79-106.
39 Van Mol, M. and O'Donnell, M.: Automatic Recognition of Generic Structure: Medical Discharge Notices. In: *Text and Texture, Systemic Functional viewpoints on the nature and structure of text*. L'Harmattan, Paris, 2004 (in press).
40 Ceusters, W., Lorré, J., Harnie, A. and Van Den Bossche, B.: Developing natural language understanding applications for healthcare: a case study on interpreting drug therapy information from discharge summaries. *Proceedings of IMIA-WG6, Medical Concept and Language Representation*, Phoenix, 16-19/12/1999, 124-130.
41 Denny, M.: Ontology Building: A Survey of Editing Tools. http://www.xml.com/pub/a/2002/11/06/ontologies.html
42 Roche, C.: Corporate Ontologies and Concurrent Engineering. *Journal of Materials Processing Technology*, 107, 187-193, 2000.
43 Arpfrez, J. C., Corcho, O., Fernandez-Lopez, M. and Gomez-Perez, A.: WebODE: a scalable workbench for ontological engineering. *Proceedings of the First International Conference on Knowledge Capture (K-CAP) Oct. 21-23, 2001, Victoria, B.C., Canada*, 2001.
44 Anonymous. WebODE Ontology Engineering Platform. http://delicias.dia.fi.upm.es/webODE/
45 Bittner, T. and Smith, B.: Directly Depicting Granular Ontologies; presented at the 1st International Workshop on Adaptive Multimedia Retrieval, Hamburg, September 2003 (http://wings.buffalo.edu/philosophy/faculty/smith/articles/DDGO.pdf).
46 Brachman, R.: On the Epistemological Status of Semantic Networks, In Findler,N. (ed.). *Associative Networks: Representation and Use of Knowledge by Computers*, Academic Press, New York, 1979; 3-50.

47 Zanstra, P. E., van der Haring, E. J. and Cornet, R.: Introduction of a Clinical Terminology in The Netherlands, Needs, Constraints, Opportunities. National ICT Instituut in de Zorg, 2003.
48 Ben Miled, Z., Webster, Y., Li, N. and Liu, Y.: An Ontology for the Semantic Integration of Life Science Web Databases, *International Journal of Cooperative Information Systems*, Vol. 12, No. 2, June 2003.
49 Baker, P.G., Goble, C.A., Bechhofer, S., Paton, N.W., Stevens, R. and Brass, A.: An Ontology for Bioinformatics Applications, *Bioinformatics*, 15: 6, 510-520, 1999.
50 Anonymous: Online Research Information Environment for the Life Sciences. http://www.oriel.org/description.html
51 Kalfoglou, Y. and Schorlemmer, M.: Ontology mapping: the state of the art, *The Knowledge Engineering Review* 18(1), 2003.
52 Sklyar, N.: Survey of existing bio-ontologies. Technical report 5/2001, Department of Computer Science, University of Leipzig.(http://lips.informatik.uni-leipzig.de:80/pub/2001-30/en)
53 Gangemi, A., Pisanelli, D. and Steve, G.: Ontology Integration: Experiences with Medical Terminologies. In N. Guarino (ed.), *Formal Ontology in Information Systems*, 163-178. IOS Press, 1998.
54 Flett, A., Casella dos Santos, M. and Ceusters, W.: Some Ontology Engineering Processes and their Supporting Technologies, in: Gomez-Perez, A. and Benjamins, V. R. (eds.) *Ontologies and the Semantic Web, EKAW2002*, Springer 2002, 154-165.