Establishing and Harmonising Ontologies in an Interdisciplinary Health Care and Clinical Research Environment*

Barry Smith[1] and Mathias Brochhausen[2]

[1]Department of Philosophy and Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo
[2]Institute for Formal Ontology and Medical Information Science, Saarland University

Ontologies are being ever more commonly used in biomedical informatics and we provide a survey of some of these uses, and of the relations between ontologies and other terminology resources. In order for ontologies to become truly useful, two objectives must be met. First, ways must be found for the transparent evaluation of ontologies. Second, existing ontologies need to be harmonised. We argue that one key foundation for both ontology evaluation and harmonisation is the adoption of a realist paradigm in ontology development. For science-based ontologies of the sort which concern us in the eHealth arena, it is reality that provides the common benchmark against which ontologies can be evaluated and aligned within larger frameworks. Given the current multitude of ontologies in the biomedical domain the need for harmonisation is becoming ever more urgent. We describe one example of such harmonisation within the ACGT project, which draws on ontology-based computing as a basis for sharing clinical and laboratory data on cancer research.

1. Introduction : Using ontologies in eHealth environments

This paper aims to provide an overview of some important recent developments in ontological engineering in healthcare and in clinical research. Ontology-based eHealth applications have become ever more popular in recent years, and they are gradually replacing the terminology-based artefacts of an earlier generation. We can in fact distinguish three (overlapping) phases in this development:

1. a phase in which work on terminology and coding schemes was dominated by the influence of library science (with classifications, which often had their origins in earlier printed dictionaries, oriented towards the cataloguing and indexing of published literature),

2. a phase in which such work was dominated by the influence of database design and software technology (with classifications focused on the need to describe and promote

* Preprint version of: Barry Smith and Mathias Brochhausen, "Establishing and Harmonizing Ontologies in an Interdisciplinary Health Care and Clinical Research Environment", in: B. Blobel P. Pharow and M. Nerlich (eds.), *eHealth: Combining Health Telematics, Telemedicine, Biomedical Engineering and Bioinformatics on the Edge* (Global Expert Summit Textbook, Studies in Health, Technology and Informatics, 134), IOS Press, Amsterdam, 219-234.

access to data, and programmers sometimes making information management decisions for a domain – biomedicine – about which they often had very little understanding)

3. a phase in which biologists are becoming increasingly involved in ontology development, resulting in an increasing orientation towards the biological reality, including the reality on the side of the patients, which texts and data describe.

Phase 1 is illustrated most clearly by MeSH [1], the vocabulary of Medical Subject Headings developed and maintained by the US National Library of Medicine for the indexing and retrieval of literature. MeSH is a highly successful and useful terminology resource; but its definitions and hierarchical organisation fall short of manifesting the sort of logical structure which would enable it to be used to maximal effect in supporting automatic reasoning. Phase 2 is illustrated by the HL7 Reference Information Model [2] and by the artefacts based thereon, which have been criticized for drawing an inadequate distinction between data and reality [3]. Phase 3 is illustrated by the Gene Ontology (GO) [4] and by the other ontologies within the Open Biomedical Ontologies (OBO) Foundry initiative [5], which are currently undergoing a coordinated process of incremental reform in the direction of greater formal rigor and of greater faithfulness to biological reality [6].

There are several components which in our view mark out ontologies from their terminological predecessors (we focus here exclusively on ontologies developed to serve the needs of biological and clinical researchers and healthcare practitioners):

- a logical structure which can support algorithmic processing,
- a concern for the reality to which the terms in an ontology relate (so that the ontology rests on a clear distinction between entities in reality and the documents or data entries used to represent them),
- a concern for the interoperability of ontologies developed for the representations of related domains of entities,
- a coherent strategy for quality assurance, based on user feedback and empirical testing, for update and maintenance in light of scientific advance and for evolutionary improvement of the ontology as a whole.

We will argue in what follows that a core aspect of ontology assessment consists in establishing that these four components are indeed realized – so that it is as if determining the quality of an ontology is in fact the other side of the coin from determining what should be called an ‘ontology’ and what should not.

2. Biomedical ontologies and realism

2.1 Recent developments in biomedical terminologies and ontologies

The increasing focus on biological reality is illustrated not only in the ontologies of the OBO Foundry but also in initiatives which play a more established role in eHealth circles.

In general, the realist holds that reality and its constituents exist independently of our (linguistic, conceptual, theoretical, cultural) representations thereof. Realism in a scientific context goes hand in hand with the doctrine of fallibilism, which holds that all our theories and classifications are subject to revision, precisely because we can always learn more about the independently existing reality towards which our scientific investigations are directed [7].

At various points in this communication we discuss developments around the SNOMED vocabulary. For the moment we note that recent revisions of SNOMED CT [8, 9], too, bear evidence of an increasing realist orientation, as is illustrated for example by SNOMED's deactivation of concepts involving the qualifier 'not otherwise specified' (NOS), such as:

262686008 Brain injury NOS (disorder),
162291000 Eye symptom NOS (finding)
162035000 Indigestion symptom NOS (finding)

Already Cimino in his famous "Desiderata" essay [10] had counseled against the use of this and similar qualifiers. As argued in [11], such terms capture, not the reality on the side of the patient, but rather a certain feature of a state of *knowledge* about such reality on the side of the healthcare practitioner. The documentation of both sorts of information is, of course, crucially important to the construction of an adequate health record. But if coding schemes are to support algorithmic reasoning in ways valuable to clinical and translational research, then we believe that it is no less important that a clear distinction be drawn between the two sorts of information. An unknown living organism (SNOMED: 89088004) is not a special sort of organism, just as a presumed viral agent (SNOMED: 106551006) is not a special sort of virus.

2.2 Increasing formal rigor

Another trend running in parallel with increasing concern for realism in ontology circles is a concern for increasing formal-logical sophistication of medical terminologies and related artefacts. Enhanced formal rigor of medical ontologies is still occasionally resisted by a school of thought which argues that medical knowledge is too intuitive and depends to too great an extent on subjective experience and local traditions to allow the creation of scientifically-based terminologies. (Medicine is an 'art' and not a 'science'.) We believe that this argument rests mainly on aspects of medical practice which are predestined to become increasingly recognised as being outdated with the growth of molecular medicine and of associated biomedical technologies.

The expansion of formal methods is illustrated most clearly in the growth of the Semantic Web, and in the work of the W3C, for example through its Healthcare and Life Sciences Interest Group [12], as also in the development of description logic infrastructures for vocabularies such as GALEN [13], SNOMED-CT and the National Cancer Institute Thesaurus [14].

Work on the OBO format (formerly the GO format) [15], the logico-linguistic idiom favored by many biologists for purposes of ontology development, is also witnessing an increasing concern with issues of formal rigor. One significant product of this is that there now exist bi-directional convertors which can automatically transform OBO ontologies into the OWL-based format used by the Semantic Web [16]. The goal of these and a series of related endeavors is to find ways to harvest greater formal rigor in order to allow for the exploitation of new possibilities of algorithmic reasoning [17] to support both biomedical research and clinical care. Increasingly the development of OBO ontologies and their application in annotations is serving as an important channel for the expansion of the Semantic Web in the life science domain.

2.3 The SNOMED initiative

It must be admitted that the attitude of clinical professionals towards ontologies is still somewhat ambivalent. Certainly ontology-based systems are viewed as bringing the promise of intriguing new possibilities in the biomedical informatics and health IT arenas. Such systems are seen as providing the possibility of transforming existing shallow coding schemes such as ICD 9, still used primarily for billing purposes, into more coherent representations of biomedical reality which might be used for purposes of research, for clinical decision support or for the gathering of more useful and more detailed public health statistics. On the other hand however – as Rector *et al.* have pointed out – systems such as those based on description logics can be hard to understand for clinical users [18]. The more formally rigorous the system, the more expensive it is to develop and maintain and the greater the costs incurred in training its users.

The most ambitious initiative to address these problems is currently being mounted by the International Health Terminology Standards Development Organisation [19], which is seeking to establish the SNOMED CT vocabulary as an international master terminology for the entire domain of biomedicine with a description logic backbone. The goal is one of comprehensive coverage of the entire domain of medicine in a multiplicity of languages, starting out from the basis of an English-language vocabulary which already comprehends more than 357,000 concepts and has partial versions in other languages. In addition the SNOMED vocabulary is mapped to other important existing standards, including the widely used ICD classifications of the World Health Organisation.

A major advantage of SNOMED CT is the comprehensive reach, which it secures through some 21 hierarchies:

Clinical findings	Procedure
Body structure	Anatomical concepts (Body Structure)
Morphologies (Body Structure)	Organism
Physical Force	Substance
Specimen	Social context

Attributes	Context Dependent categories
Physical object	Events
Environments and geographical location	Observable entity
Qualifier value	Staging and Scales
Special concept	Pharmaceutical / biologic product
Record artefact	

2.4 Problems with SNOMED in the clinical setting

An issue which has still not been satisfactorily resolved, however, is the degree to which the introduction of sophisticated broad-coverage terminologies such as SNOMED CT into the hospital environment will involve costs in training and implementation which would be so great that they could not be justified by compensating rewards. Billing needs are catered for by simpler terminologies. The rewards of using a more rigorous and comprehensive terminology do indeed promise to be of great significance for example for clinical decision support and for more adequate public health data. By providing common structure and terminology, the SNOMED CT international master vocabulary would go far towards providing a single data source for review and also bring the benefits of less redundant data and easier opportunities for longitudinal studies and meta-analysis and for ensuring consistency of data across the lifetime of the patient and from one healthcare institution to the next. The use of SNOMED CT would allow in addition the use of common tools and techniques, common training and a single validation of data. But the fact that so few healthcare institutions have embraced SNOMED CT for clinical coding seems to suggest that incentives are still missing for the considerable investments which would be needed to harvest these benefits [20].

One additional factor is that SNOMED CT is marked by a number of internal structural problems (for example gaps in the terminology, a lack of compositional structure, shortfalls in consistency from one part of the vocabulary to another) which detract from its appeal to novice users and provide obstacles to its efficient application in coding [11, 21, 22]. In our view it is still the case that too little effort is being invested in attempts to decrease the costs involved in adoption of terminologies such as SNOMED as a basis for clinical coding by improving the degree to which such problems are addressed, and it seems to us that the major existing healthcare terminology resources still lack coherent strategies for incremental improvement.

2.5 Towards evidence-based ontology development

We believe that at least part of what must be involved in any such strategy is the development of an evidence-based evolutionary methodology for quality assurance of ontologies – a methodology whose application can at one and the same time both enhance the degree to which ontologies constitute a realistic representation of reality and create a more intuitive and more easily maintainable framework for clinical coding. The ideal result of the implementation of such

a strategy would be a framework which is both biologically accurate and able to supply its users with a view of clinical reality which coheres with their expectations of how this reality should look. Such an outcome would, we believe, not merely save time in coding and raise the accuracy, breadth, and depth of coverage of the results; it would also enhance the degree to which the systems in question can be used by the clinician and researcher for genuinely useful purposes.

We have argued in a number of prior publications [23, 24, 25] that reasoning on the basis of information that comes closer to an adequate picture of reality has the potential to provide the basis for more valuable results, whether in decision support, meta-analysis, or trial management, than reasoning on the basis of representations which confuse features of our data or knowledge with features of the reality on the side of the patient. We will describe below a realism-based project in the domain of post-genomic clinical trials that is designed to yield benefits of just this sort [26].

3. Harmonisation and quality management in ontology development

In the development of the OBO Foundry, and also in some of the more mature initiatives within the framework of the Semantic Web, we can witness a third important trend – in addition to those of *greater realism* and *greater formal rigor* – a trend towards the *harmonisation of ontologies* with a view towards ensuring their interoperability. The goal of such harmonisation is to bring about a situation in which the coverage of ontologies can be increased in stepwise fashion across ever broader domains of biomedical reality while at the same time preserving the advantages of consistency and of formal rigor [4, 6, 27]. This trend is at the opposite pole from that of SNOMED, which relies on the idea of a single broad-coverage master terminology, in that it seeks to draw in systematic fashion on the benefits of modularity while ensuring extendibility of coverage through interoperation of its separate modules.

3.1 Benefits of harmonisation

Today it is generally agreed that the goal of ontology development should be, not to develop one single ontology covering the entirety of what exists, but rather to find ways in which ontologies covering different domains of reality can be developed in tandem with each other in a way which allows exploitation of the benefits of division of labor and pooling of expertise. Experts in given domains should clearly be the ones to bear the burden of developing and of maintaining the ontologies in those domains. Experience has demonstrated also that experts are willing to invest considerable resources to this end in return for the benefits – analogous to those yielded through participation in the open source software movement [28] – of contributing to the improvement of a valued community resource. But experience suggests also that domain experts need assistance in ontology development in the form of guidelines which tell them which direction to take in their work. This is so especially where ontologies must be created *ab initio*, in areas where the need for controlled vocabularies for data annotation is only now beginning to be acknowledged. Guidance is needed by those new to ontology as to successful methodologies above all to ensure

the development of ontologies which will interoperate with those which already exist in neighboring domains.

Such interoperation should also serve to ensure combinability of terms when composite terms need to be formed for specific application purposes. One problematic feature of the SNOMED vocabulary is its non-compositional character, illustrated for example by 'assay for X' terms, such as

SNOMED 55534003: macrophage migration factor assay,

where the corresponding 'X' term is missing from the vocabulary. SNOMED thereby allows the simple coding of information about *macrophage migration factor assays*, but no correspondingly simple coding of *macrophage migration factors* themselves. The OBO Foundry ontologies, in contrast, embrace a deliberate policy of ensuring compositionality [29]. Indeed compositionality of terms is used as a methodology to support coherent ontology development, as for example in the Foundry's Infectious Disease Ontology (IDO), which provides a repertoire of those basic component terms, such as 'host', 'pathogen', 'vector', which are used in all infectious disease domains, and works with researchers on specific diseases with a need to form specialized ontologies for different combinations of pathogen, host and vector, to create the corresponding extensions as far as possible through simple composition [30].

The general strategy, embraced also by the CARO Common Anatomy Reference Ontology [31], is to develop small reference ontologies for well-specified domains and to extend these ontologies to create larger ontology frameworks for specific application purposes by combing terms from different ontologies in what are called 'cross-products' [32]. This strategy contributes to ensuring comparability of the separate ontologies, and therefore also to guaranteeing alignment of the data annotated in their terms. It serves at the same time, again, to provide common guidelines for the developers of the specialized ontologies and to allow the lessons learned by early adopters of the strategy to be passed on to their successors as the guidelines become incrementally refined. [33].

The OBO Foundry is a systematic realization of this strategy. Following the model of the Gene Ontology, ontologies are created for specific domains on the basis of standards which have been accepted in advance by separate groups of ontology developers because they are designed to secure interoperability of their separate ontologies. The Foundry thereby provides an evolving suite of orthogonal basic science ontologies for rigorous annotation of different kinds of experimental data. At the same time it provides rules for the creation of cross-product terms on the basis of terms from its constituent ontologies joined together via relations formally defined in the Foundry's Relation Ontology (RO) [34]. These rules are applied as a means of removing the arbitrariness involved in the informal cut-and-paste strategies for term-composition embraced by more traditional terminologies. On the one hand orthogonality of the source ontologies goes a

long way to ensuring a unique choice for constituent terms where complex term formation is needed; on the other hand the formal definitions of the RO help to ensure unambiguous meaning of the results of this combination. Because all complex terms are required to be defined as cross-products of more basic terms, compositionality, with associated benefits for automatic reasoning, is ensured.

By providing regimented sources and templates for term composition the Foundry is, we believe, helping to avoid the bottlenecks currently created for example in the case of SNOMED CT development, where each new term must be approved for inclusion in the ontology, through a multi-stage committee process, on the basis of intuitive rules rather than of formal principles.

3.2 Harmonisation efforts in pre-existing systems

Independently of the success of either the OBO Foundry initiative or of SNOMED CT's efforts towards international standardisation, it is already clear that in developing ontology-based applications in the biomedical field, account must be taken of a large number of pre-existing terminologies, controlled vocabularies and ontologies, some of which – such as MedDRA, ICDx, LOINC, OMIM, and other constituent vocabularies of the UMLS Metathesaurus [35] – already have the status of de facto standards. Increasingly it is becoming clear that it will be necessary to achieve progressive integration of such representations, too, and to this end strategies must be found to bring about an incremental, evidence-based, process of harmonisation. This will need to be achieved, in part at least, on the one hand by ensuring internal formal coherence of each representation, and on the other hand by maximizing their conformity with the results of on-going scientific research, ideally as this is documented within external gold standard reference ontologies such as are proposed within the Foundry framework. Unfortunately the need for such radical harmonisation has still not been generally recognised, though small steps in the necessary direction can be witnessed.

3.3 Challenges to harmonisation

The different ontologies, terminologies and other means of knowledge representation in the biomedical domain are often governed by different attitudes towards reality, towards the representation of reality, and towards the most effective practical use of representations in data management. The resultant multiplicity of approaches poses severe challenges to harmonisation.

A key challenge is that of preserving coherence as the reach of the ontologies becomes ever further extended and the ontologies themselves become ever more complex. Ways will need to be found to ensure that this extended reach and complexity does not act to the detriment of what has already been achieved within given healthcare institutions or disciplinary communities. The introduction of ontology-based technology or of global resources such as SNOMED CT should not lead to a corruption or dilution of quality standards already established, and it should not lead to already working terminological solutions developed to meet specific local needs becoming

overwhelmed by the needs of conformity with larger frameworks. To achieve this end within the OBO Foundry techniques are being delivered to create slimmed-down versions or 'views' of larger ontologies designed to achieve specific local purposes while ensuring consistency with the larger framework [36, 37].

The harmonisation of ontologies must be to some degree centrally organised through directives which enjoy consensus support and are clearly documented in such a way as to address the needs of a variety of different types of audience and to secure their willingness to participate. But if harmonisation is not to bring negative consequences it must be effected in a stepwise fashion, with careful precautions to ensure that existing solutions are not jeopardized.

4. Utilizing upper level ontologies for harmonisation put this earlier with stuff on harmonisation

Providing an upper level overarching ontology framework for reality representation is a basic feature of harmonisation. According to the Standard Upper Ontology (SUO) working group of IEEE:

> An upper ontology is limited to concepts that are meta, generic, abstract and philosophical, and therefore are general enough to address (at a high level) a broad range of domain areas. Concepts specific to given domains will not be included; however, this standard will provide a structure and a set of general concepts upon which domain ontologies (e.g. medical, financial, engineering, etc.) could be constructed [38].

Upper level ontologies can provide not merely basic categories ensuring good ontology organization but also a set of tested principles that can be re-used by others in the development of specific domain ontologies. The Basic Formal Ontology (BFO), which serves as upper level kernel of OBO Foundry ontologies, rests on a basic distinction between continuants and occurrents. The former are entities in reality that endure (continue to exist) through time. They persist self-identically even while undergoing changes of various sorts. The latter *occur*, which means that they have, in addition to their spatial dimensions, also a fourth, temporal dimension. Occurrents (for example processes) unfold through a period of time in such a way that they can be divided into temporal parts or phases. They have a beginning, a middle, and an end [7]. Continuants, in contrast (for example organisms), exist in full at any time at which they exist at all, while at the same time gaining and losing parts in the course of development and growth.

Using an upper level ontology can foster harmonisation by providing a uniform and coherent approach to reality representation at the topmost level of organisation. It is at the lowers levels, however, that we will find those terms which predominate in practical uses of the ontology. General criteria of the sort embodied in an upper level ontology provide useful tools when organizing these lower level terms. [39]

An upper ontology thus stands to a domain ontology in roughly the same relation as mathematics to physics. We need to prove mathematical theorems only once, and we can thereafter use these theorems over and over again in different physical theories. Similarly, a major advantage of an upper level ontology is its status as a tested resource, whose re-use prevents time-consuming re-development of those meta-level structures which are needed by domain scientists to organize their ontology resources, but which embody principles of which these domain scientists will likely have an imperfect grasp.

5. Methodologies in ontology development

5.1 What does 'ontology' mean?

We hold that the hypothesis of realism is fundamental to the realisation of the goal of evidence-based harmonisation in ontology development, and that the still widely popular 'conceptualist' alternatives to this hypothesis in fact constitute obstacles to success in its achievement because the conceptualist can point to no benchmark against which such success could be measured.

The conceptualist view, still popular in knowledge engineering and AI circles, sees ontologies as representations of what are called 'concepts', which means, roughly, units of knowledge (or of meaning) in the mind of human beings [36]. The definitions of Gruber [40] and Studer *et al.* [41] are concept-based definitions of ontology in this sense. Here, in contrast, we propose the following definition:

> An ontology is a representation of the universals or classes in reality and of the relations existing between these universals or classes.

Universals are the real invariants or patterns in the world apprehended by the specific sciences. The relation between universals and particulars is one of *instantiation*. Universals are multiply instantiated: they exist at different places and times in the different particulars which instantiate them [7]. Universals are designated by general terms in ontologies such as 'dog' or 'cell' or 'oophorectomy' or 'diabetes'. 'Dog' is the name of a universal which is instantiated by my dog Fido and by your dog Rover. There are however also general terms which do not designate universals, such as 'dog owned by the Emporer' or 'patients with diabetes in the Homburg University Hospital'.

In these terms we can propose the following specification of the relation between universals and what, in Semantic Web circles, are called 'classes'. In our idiolect a *class* is a collection of all and only those particulars to which a given general term applies. Where the general term in question refers to a universal, then the corresponding class, called the *extension* of the universal, comprehends all and only those particulars which as a matter of fact instantiate the corresponding universal [42].

Our realist view is based on a distinction between three levels of reality:

1. the ideas, thoughts in our minds which form representations of specific portions of reality
2. those representational artefacts (including ontologies, textbooks, and so forth) which we develop to make these mental representations concretely accessible to others.
3. reality itself, which serves as the target of these mental and physical artifacts

We believe that success in ontology development depends on keeping clear the distinction between these three levels [42] and on recognizing that the reality which our representations are developed to represent exists independently of these representations themselves. Only in relatively rare cases (for example in the ontology of psychiatry) is this reality inside our heads, but even there it is possible to keep the three levels clearly distinct.

5.2 Methods of ontology development

A realist paradigm in ontology development brings the need to foster the creation of gold standard ontologies which reflect current scientific understanding and serve both as models of good practice and also as benchmarks against which the correctness of other ontologies can be gauged. Such gold standard ontologies, the Foundational Model of Anatomy [43] is our paradigm example, should be not merely in good order as they stand as representations of their selected domain of reality; they should also employ state-of-the-art practices in order to ensure that they are well-maintained and updated as knowledge advances.

Such gold standard ontologies must, we believe, be developed and maintained by experts in the corresponding domains. Techniques of ontology development via natural language processing (NLP) as applied for example to textbook literature sources produce results which still fall far short of the necessary formal rigor and scientific accuracy. Such techniques would, if they could be successfully developed, bring tremendous benefits in the biomedical domain, where ontologies and other terminology resources may be very large and may need to be updated rapidly in response to large-scale changes in our underlying scientific knowledge. Increasingly, therefore, we anticipate that NLP tools will provide valuable assistance to ontology-based research. We do not, however, anticipate that such tools will themselves be capable of being used in the creation of ontologies which can serve in the role of gold standard along the lines described. Indeed we believe that gold standard ontologies will themselves provide an indispensable presupposition to their successful application to other purposes.

6. Ontology based clinical research initiatives

6.1 The ACGT project

Recent years have seen a number of initiatives resting on the use of ontologies to facilitate cross-linkage of clinical research activities among institutions and communities of researchers. One such initiative financed by the European Union within its 6th Framework Programme is the

Advancing Clinico-Genomic Trials on Cancer (ACGT) project. The goal of this project is to enable the rapid sharing of data gained in both clinical trials and associated genomic studies. In order to meet this goal ACGT is providing a GRID structure designed to transmit the data between different groups of users in real time according to need, with data integration being achieved by means of an ontology-based mediator [44].

The ontology-related efforts of the ACGT project provide an example of one large-scale effort to create new ontologies useful to clinical research, and it provides a valuable testbed for learning lessons about successful and non-successful strategies for their integration with existing clinically relevant ontology resources.

When considering the development of an ontology-based information-sharing system for the cancer domain, the National Cancer Institute Thesaurus (NCIT) is a terminology resource of obvious importance. Yet, there are a number of drawbacks preventing the use of the NCIT itself as ontology for the ACGT project, in part because its formal resources are too meagre for our purposes, with only a fraction of NCIT terms being supplied with the formal definitions of the sort required by its official description logic framework. The NCIT contains only one relation, namely the subtype relation (*is_a*), as contrasted with the plurality of formally defined relations included, for example, within the OBO Relation Ontology. Further, the NCIT (like the UMLS source vocabularies from which it is derived) is marked by a number of problems in its internal structure and coverage [45, 46], including problems in the treatment of *is_a*. A small example can be found in its treatment of 'Organism', which includes among its subtypes for example 'Other Organism Groupings', so that we have

Other Organism Groupings *is_a* Organism [14].

6.2 The ACGT Master Ontology

In light of such problems the ACGT consortium developed its own Master Ontology (MO) to address the goal of data integration for the domains of clinical studies, genomic research and clinical cancer management and care. The ontology was constructed in modular fashion, with Clinical Trial and Patient Management Ontology modules designed to be reused for different clinical domains. The ontology has thus far been developed manually, in order to secure the high standards of knowledge representation outlined in the foregoing.

One basic principle of ontology development is that ontologies include only what is general (classes, universals), and thus not particulars (instances, tokens). Hence the ACGT MO does not include real world instances but only universals. It also embraces principles of good practice designed ensure a proper treatment of the *is_a* relation. First it insists on a formal rule according to which

A *is_a* B if and only if all instances of A are also instances of B.

This rule guarantees the transitivity of the is_a relation (so that from A is_a B and B is_a C, we can infer A is_a B) and thus allows application to the corresponding statements a simple but nonetheless very useful kind of reasoning. The rule can also be used to ensure a coherent structure of the backbone is_a hierarchy (taxonomy) by ruling out those informal is_a relations still used in a number of terminology resources, as for example in SNOMED:

cow *is_a* class mammalian
kingdom animalia is_a organism

and so on.

Second ACGT MO embraces the rule of single inheritance, designed to guarantee that the backbone *is_a* hierarchy of the ontology should be a genuine hierarchy in the sense that each child term should have at most one single parent, according to the rule:

if A *is_a* B and B *is_a* C, then B and C are identical.

The central aim is to avoid the sorts of polysemy ('*is_a* overloading'), and associated errors, that often results where multiple inheritance is allowed. This rule is designed also to support the sort of modularity of ontologies, and associated engineering benefits, captured in Alan Rector's project of normalisation [47]. As Rector points out, the restriction to single inheritance in the hierarchy asserted within the ontology is perfectly compatible with use of the ontology to infer a polyhierarchy as required. In Rector's view, managing a large body of complex definitions becomes easier using a single inheritance hierarchy. This it is easier to make changes in a simple hierarchy when changes are needed. There are also human factors. Experience has shown that people make mistakes with when they have the freedom to deviate from the principle of single inheritance.

The basic principles of the development of the ACGT MO have been derived directly from BFO. The ACGT MO is aligned with OBO Foundry ontologies such as the FMA [48] and GO. It also incorporates slightly modified versions of existing medical classifications such as the TNM system [49] in ways designed to enhance their interoperability with other ontologies in the system.

6.3 Clinical trial management, and ontology harmonisation within the ACGT project

The central goal of the ACGT project is to put clinicians in the driver's seat, thus ensuring that all project efforts are in the service of patient care. The ACGT MO has been developed in close collaboration with clinicians utilizing existing Clinical Report Forms (CRFs) to gather documentation on the universals and classes in their respective target domains. All versions of the ontology have been reviewed by clinical partners who have proposed changes and extensions

according to need. In this process the problem of handling an ontology with more than 1300 classes became apparent. This led to the decision that ACGT should aim to provide tools to view the ontology in user-friendly ways. The basis for these efforts is a clinical view of the ACGT MO (resting, as it were, on the full ontology running behind the scenes), a view based on tracking the workflows common in clinical practice thereby encapsulating the clinician's approach to a medical problem.

A group of IT specialists, clinicians and ontologists in ACGT has proposed the development of a novel ontology-based system to administer clinical trials [26] called ObTiMA (for Ontology-based Trial Management for ACGT). ObTiMA allows each clinical trial administrator to create automatically an ontology-based Clinical Data Management System [50] tailored to the needs of each given trial. One core intended functionality of ObTiMA, currently in its test phase, is an ontology-based tool for the generation of clinical report forms called CRF Creator. The idea here is, that instead of mapping the data in existing clinical databases to external ontologies, the data that is collected will be classified in terms of the ontology from the very start. ObTiMA will support the clinician in both planning and management of clinical trials. In addition, it is planned to serve as a tool for the maintenance of the ACGT MO itself, in a fashion designed to ensure just the kind of tight connection between ontology and empirical investigation that is the key to evidence-based ontology development. As trial administrators propose new terms to be submitted for review by the ontology's curators, this provides a way in which advances in biomedical knowledge can become automatically incorporated into the system and so made available to all its users distributed across a plurality of diverse institutions. It also brings about updates of clinical report form templates in such a way that at least certain aspects of legacy data generated in terms of earlier versions of the ontology can become updated automatically.

7. Conclusion

The need for ontology harmonisation, based on principles-based ontology evaluation, is now accepted in a number of influential ontology circles. But these efforts need to be still more intensively pursued. The potential users of ontology-based tools in the eHealth domain still need to be convinced that such tools will be not only easy to use but also useful in their work. Ontology development and evaluation efforts must therefore always rest on close collaboration with the intended users.

The successful ontologies in the biomedical domain all work in the same way [6]. Researchers working in a given domain have data; they need to make this data available for semantic search and algorithmic processing; and to achieve this end they take steps to create a consensus-based ontology for annotating (describing) their data, working with ontologists who help them to ensure that their ontology can interoperate with ontologies already created for neighboring domains. Experience suggests that the most reliable way to create an ontology for such purposes is on the basis of a collaboration between experienced ontologists working to explicit, tested guidelines with domain expects working to address real data annotation needs.

References

[1] http://www.nlm.nih.gov/mesh.

[2] http://www.hl7.org.

[3] Smith B, Ceusters B (2006) HL7 RIM: An Incoherent Standard (MIE 2006), *Studies in Health Technology and Informatics*, vol. 124:133–138.

[4] http://www.geneontology.org.

[5] http://obofoundry.org.

[6] Smith B, Ashburner M, Rosse C, et al. (2007) The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology*; 25 (11):1251-1255.

[7] Grenon P, Smith B, Goldberg (2004) Biodynamic Ontology: Applying BFO in the Biomedical Domain. In: Pisanelli DM (ed.): *Ontologies in Medicine*. IOS Press, Amsterdam:20-38.

[8] http://www.snomed.org/snomedct.

[9] Ceusters W, Spackman KA, Smith B (2007) Would SNOMED CT benefit from Realism-Based Ontology Evolution? In Teich JM, Suermondt J, Hripcsak C. (eds.), American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy, Chicago IL, 2007:105-109.

[10] Cimino JJ (1998) Desiderata for controlled medical vocabularies in the Twenty-First Century. *Methods Inf Med*; 37(4–5):394–403.

[11] Bodenreider O, Smith B, Burgun A. (2004) The ontology-epistemology divide: A case study in medical terminology. FOIS (Formal Ontology and Information Systems) 2004: 185-95.

[12] http://www.w3.org/2001/sw/hcls/.

[13] http://www.opengalen.org/.

[14] http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do.

[15] http://www.geneontology.org/GO.format.shtml.

[16] http://www.berkeleybop.org/obo-conv.cgi.

[17] Ruttenberg A, Clark T, Bug W, et al. (2007) Advancing translational research with the Semantic Web. *BMC Bioinformatics*, 8.

[18] Rector AL, Zanstra PE, Solomon WD, Rogers JE, Baud R et al. (1999) Reconciling Users Needs and Formal Requirement: Issues in Developing Re-Usable Ontology for Medicine. *IEEE Transactions on Information Technology in BioMedicine* 2(4):229-242.

[19] http://www.ihtsdo.org/.

[20] www.hiww.org/smcs2006/talks/Rector.ppt.

[21] Ceusters W, Smith B, Kumar A, et al. (2004) Ontology-based error detection in SNOMED-CT®. *Proc Medinfo 2004*: 482-6.

[22] Ceusters W, Smith B, Kumar A, et al. (2004) Mistakes in medical ontologies: Where do they come from and how can they be detected? *Studies in Health Technology and Informatics;*102: 145-64.

[23] Köhler J, Munn K, Rüegg A, et al. (2006) Quality control for terms and definitions in ontologies and taxonomies. *BMC Bioinformatics*;7:212-20.

[24] Smith B, Ceusters W (2005) An ontology-based methodology for the migration of medical terminologies to Electronic Health Records. *Proc AMIA Symp 2005*:704-8.

[25] Rosse C, Kumar A, Mejino JLV, et al. (2005) A strategy for improving and integrating biomedical ontologies. *Proc AMIA Symp*:639–43.

[26] Weiler G, Brochhausen M, Graf N, Hoppe A, Schera F, Kiefer S (in press) Ontology Based Daa Management Systems for Post-Genomic Clinical Trials within an European Grid Infrastructure for Cancer Research. *29th IEEE EMBS Annual International Conference, Lyon*.

[27] Ceusters W (2006) Towards A Realism-Based Metric for Quality Assurance in Ontology Matching. In: Bennett B, Fellbaum C (eds) Formal Ontology in Information Systems. Proceedings of FOIS-2006. Amsterdam:321-332.

[28] Webber S (2004) The Success of Open Source, Cambridge, MA: Harvard University Press.

[29] Mungall CJ (2004) Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics* 5, 6-7:509-520.

[30] http://www.bioontology.org/wiki/index.php/Infectious_Disease_Ontology.

[31] Haendel, M *et al.* (in press) CARO: the Common Anatomy Reference Ontology. In: Burger et al. (eds): *Anatomy Ontologies for Bioinformatics*, New York: Springer.

[32] Hill DP, Blake JA, Richardson JE, Ringwald M (2002) Extension and integration of the Gene Ontology (GO): Combining GO vocabularies with external vocabularies. *Genome Res*. 12:1982-1991.

[33] http://sourceforge.net/mailarchive/forum.php?forum_name=obo-crossproduct.

[34] http://obofoundry.org/ro.

[35] www.nlm.nih.gov/pubs/factsheets/umlsmeta.html.

[36] http://www.geneontology.org/GO.slims.shtml.

[37] Detwiler LT, Brinkley JF (2006) Custom Views of Reference Ontologies. *Proceedings, American Medical Informatics Association Fall Symposium*, Bethesda, MD:909.

[38] http://suo.ieee.org.

[39] Smith B (2006) From Concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies. *Journal of Biomedical Informatics* 39(3):288-298.

[40] Gruber TR (1993) A Translation Approach to Portable Ontologies. *Knowledge Acquisition*, 5:199-220.

[41] Studer R, Benjamins VR, Fensel D (1998) Knowledge Engineering: Principles and Methods *Data & Knowledge Engineering*, 25(1-2):161-198.

[42] Smith B, Kusnierczyk W, Schober D, Ceusters W (2006) Towards a Reference Terminology for Ontological Research and Development in the Biomedical Domain. *KR-MED 2006*.

[43] Rosse C, Mejino JLF (in press) The Foundational Model of Anatomy ontology. In: Burger A et al. (eds.): *Anatomy Ontologies for Bioinformatics*. Springer, New York.

[44] Tsiknakis M, Brochhausen M, Nabrzyski J, Pucaski L, Potamias G, Desmedt C, Kafetzopoulos D (in print) A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of post-genomic clinical trials on Cancer. *IEEE Transactions on Information Technology in Biomedicine*, (Special issue on Bio-Grids).

[45] Ceusters W, Smith B, Goldberg L (2005) A Terminological and Ontological Analysis of the NCI Thesaurus. *Methods of Information in Medicine* 44:213-220.

[46] Kumar A, Smith B (2005) Oncology Ontology in the NCI Thesaurus. *Artificial Intelligence in Medicine Europe* (AIME 2005) (Lecture Notes in Computer Science 3581):213-220.

[47] Rector A (2003) Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. *K-CAP'03*, October 23-25, 2003, Sanibel Island, Florida, USA:121-8.

[48] http://sig.biostr.washington.edu/projects/fm.

[49] Wittekind C, Meyer HJ, Bootz F (2005) TNM. Klassifikation maligner Tumoren, Springer, Heidelberg.

[50] Graf N, Weiler G, Brochhausen M, Schera F, Hoppe A, Tsiknakis M, Kiefer S (2007) The Importance of an Ontology Based Clinical Data Management System (OCDMS) for Clinico-Genomic Trials in ACGT (Advancing Clinico-Genomic Trials on Cancer), November 1-3, SIOP Mumbai, Mumbai.