

Quality Assurance of Medical Ontologies

J. E. Rogers

BioHealth Information Group, School of Computer Science,
University of Manchester, Manchester, United Kingdom

Summary

Objective: To review the literature concerning the quality assurance of medical ontologies.

Methods: scholar.google.com was searched using the search strings (+ontology + "quality assurance") and (+ontology + "evaluation/evaluating"). Relevant publications were selected by manual review. Other work already familiar to the author, or suggested by other researchers contacted by the author, were included. The papers were analysed for common themes.

Results: Four broad properties of an ontology were identified that may be quality-assured: philosophical validity, compliance with meta-ontological commitments, 'content correctness', and fitness for purpose. Each published methodology addressed only a subset of these properties. 'Content' may be divided into domain knowledge content, and metadata describing either the provenance of domain knowledge content, or relationships between it and lexical information (e.g. for display and retrieval). 'Correctness' (whether of domain knowledge content or metadata) may also be further subdivided into truth, completeness, parsimony and internal consistency.

Conclusions: Understanding of how to assure the quality of ontologies, or evaluate their fitness for specific purposes, is improving but remains poor. A combination of methodologies is required, but tools to support a comprehensive quality assurance programme remain lacking.

Perfect quality of an ontology is not provable and may not be desirable: an ontology compliant with all current philosophical theories, following necessary ontological commitments, and with entirely 'correct' content, may be too complex to be directly usable or useful. The extent to which an ontology's fitness for purpose is predicted or influenced by its other properties remains to be determined. Field studies of ontologies in use, including interrater effects, are required.

Keywords

Terminology, quality control, information systems

Methods Inf Med 2006; 45: 267–74

Introduction

Controlled vocabularies are recognised in many informatics domains as a necessary component if data and systems interoperability is to be achieved. Such vocabularies seek simultaneously to provide both sufficient expressivity to support primary data collection about individual data subjects, and the capability to support less well specified – or unspecified – secondary uses involving *post hoc* merger and analysis of data from multiple different data subjects. For either purpose, simple lists of agreed terms are not generally sufficient: classification of the terms as a means to aggregate data is usually a central functional requirement.

Medical terminologists have traditionally constructed controlled vocabularies for their domain by attempting to enumerate and classify *pre hoc* exhaustive, static lists of all the concepts required. Since the mid 1990s, however, biomedical terminologists have increasingly rejected this approach in favour of a faceted or compositional structure whereby complex concepts are represented *ad hoc* as a structured composition of simpler concepts (e.g. Gene Ontology, ICPC, Clinical Terms Version3, SNOMED 3). The most advanced medical terminologies such as SNOMED-CT[®] and OpenGALEN attempt to go still further, employing a class of knowledge representation paradigms now commonly known as 'logic-based' ontologies based on 'description logics' [1–7]. These support both *ad hoc* composition and classification.

This paper considers how the quality of such logic-based ontologies might be assessed. To place the discussion in context, a short summary is first presented of the limitations of traditional clinical terminology technologies and how logic-based ontologies address these limitations. A review of prior work on ontology quality assurance is

then presented. The final section proposes a framework within which any methodology purporting to assure or control quality of an ontology may, itself, be assessed.

Problems with Medical Classifications

Medicine and life sciences are characterised by very large and descriptive domains. Two serious challenges arise: delivering a highly expressive terminology that can be used consistently, and organising all descriptions supported by that terminology into a classification.

In pursuit of expressivity, entirely manual attempts to construct comprehensive controlled clinical term lists have produced large corpora, but these have still proved unsatisfactory. Even very large pre-enumerated lists remain inadequately expressive for users, and prone to interrater variability in use [8].

Organisation of the terms also remains problematic. Most actively used schemes aim primarily to support data aggregation for statistical purposes and thus are organised as a monohierarchy to avoid double counting. However, *different* statistical studies require a *different* monohierarchy, whilst new clinically important data abstraction applications such as decision support require a polyhierarchy. For these reasons biomedical terminologies in development are now more commonly structured as polyhierarchies, conceived as the conflation of multiple distinct statistical monohierarchies with one or more abstraction polyhierarchies.

Constructing polyhierarchical life science classifications by hand is, however, highly error-prone. Our own experiments with even relatively small, manually maintained polyhierarchical lattices of only a few hundred entities suggest typically between

5% and 10% of entities are missing at least one valid subsumption relationship with another entity in the lattice [9-11].

Ontologies

The history of ontology as a branch of philosophy stretches back to the ancient Greeks. Here, it means the metaphysical study of the nature and essential properties and relations of all beings, or of the principles and causes of being.

This paper is concerned with a second meaning coined more recently within the artificial intelligence and knowledge representation communities: the study of how to represent those objects, concepts and other entities assumed to exist in some area of interest, and the relationships that hold among them. In the context of biomedical terminologies and classifications, ontologies aim to analyse and represent both the explicit and implied concepts used within a particular biomedical discipline, and the relationships between those concepts [12].

Such an analysis will usually go further than, for example, the simple observation that an existing terminology has listed entities such as ‘endocrine disease’ and ‘thyroid cancer’ together with a relationship between them, such that one is a kind of the other. Rather, the thyroid gland itself will be identified as an entity whose existence is implied by many phrases in the terminology, and the endocrine system as another, even when no specific code existed for either entity in the original scheme. Similarly, the entities of disease and cancer are identified. The ontological analysis would further represent that the thyroid was, structurally, *part_of* the endocrine system, that cancer *is_A* disease, and that endocrine disease and thyroid cancer may be defined respectively as ‘disease *with_locus* endocrine system’ and ‘cancer *with_locus* thyroid gland’.

By these steps, the detailed reasons *why* ‘thyroid cancer’ is classified as a kind of ‘endocrine disease’ are made explicit. A common motive for such deconstructive ontological work is to use the newly explicit information to infer new relationships, or validate existing ones.

Logic-based Ontologies

Logic-based ontologies, formed by combining an ontology as described above with rules for reasoning over it, offer a unified solution to address many of the problems of traditional medical classifications. This solution comprises an explicit model of the conceptual content of the medical domain (the ontology), expressed in a syntax with specified semantics (a *logic-based* ontology), that can be reasoned over by a computer algorithm (a description logic engine). This approach promises both greatly increased expressivity – users can create *ad hoc* an indefinitely large number of ‘composed entities’ as structured combinations of simpler entities and relations already in the ontology – and a simultaneous reduction in the manual effort required to curate and organise the resulting dynamic corpus. This reduction in effort arises primarily because – within a carefully delimited subset of first order logic – description logic engines automate the process of examining the explicit semantics of any given new expression, comparing it with the semantics of all other previously encountered expressions, and thus deriving an automatic classification of the new expression with respect to the space of entities already encountered [13].

Quality Assurance of Medical Ontologies

Cimino’s desiderata for modern biomedical terminology construction commend an approach based on concepts and formal semantics [14]. But whilst a logic-based framework should improve the accuracy and speed of any inference, logic itself does not guarantee no errors can arise: “garbage in, garbage out” still applies. Logic and semantics may be necessary to the success of the enterprise, but they are not sufficient. Whilst Cimino recognises the primacy of content, the desiderata for its quality assurance – those systematic actions necessary to provide adequate confidence that ontology content meets the needs and expectations of the customer – remain to be proposed or agreed.

As a logic-based ontology grows larger, so the number of possible inferences about it grows as a combinatorial explosion. It is therefore impossible to inspect and affirm all possible inferences within logic-based ontologies of useful scale, particularly if one must also detect when correct answers are obtained for the wrong reasons. Instead, as Goble observes [15], we can only attempt to validate both the initial axioms and the soundness of the reasoning algorithm, and claim that we know by inference that no incorrect result could be derived.

For the purposes of this paper, these two verification tasks (the axioms, and the reasoner) are treated as orthogonal and only axiom verification is considered in further detail. In reality, however, these tasks are not entirely orthogonal: a lesser reasoner may either require explicit statement of axioms that only a more powerful reasoner can infer or (if also based on a lesser formalism) be unable to import those explicit axioms expressed in a more powerful formalism. Therefore, although a reasoner’s soundness can be expressed without reference to any particular ontology, validation of any specific set of ontological axioms should properly be qualified by the minimum formalism and reasoning capability for which that validation holds.

Whilst many publications exist reporting the tradeoffs between the soundness, expressiveness and computational tractability of description logic algorithms (e.g. [16]), Schulz observed in 1998 that the medical informatics literature was weak concerning the problems of evaluating and quality assuring the knowledge content of large and complex ontologies [17]. He postulated this may be because coded clinical data had previously been used mainly in aggregate, with individual errors smoothed out: any residual error could be tolerated as long as population trends could be followed across time. Gómez-Pérez observed in 1994 that no guidelines existed by which to evaluate ontologies in general [18]. In 2001, she noted that the field of evaluating ontologies was still only just emerging, and suggested this was perhaps in part due to low levels of interest in the issue within the ontology engineering community, and in part because the tools to support ontology evaluation were not yet developed [19, 20].

Objectives

This paper reviews the literature on assuring the quality of ontologies both specifically relating to medicine and in general. A framework for evaluating such programmes is proposed.

Method

(+ontology +“quality assurance”) and (+ontology +“evaluation/evaluating”) were presented as search strings to scholar.google.com. Eighteen out of 124 returned papers were selected following manual review. Other work already familiar to the author, or suggested by prominent researchers contacted by the author, were included. The results are grouped into those reporting work specifically with medical terminologies, and those from the wider ontology engineering community.

Results

Medical Terminologies and Ontologies

Schulz [17, 21] described some of the quality control processes used in the construction of the UK's Clinical Terms Version 3 (CTV3). These included lexical tools to suggest classifications, but also a periodic automatic classification of the concept space, according to the explicit semantic definitions and using what amounts to a primitive description logic algorithm. The inferred and asserted hierarchies were subsequently automatically compared to detect certain forms of misclassifications.

Schulz noted that, in examining the computed result for errors, humans were better at noticing misclassifications than missed classifications. He also commented that assuring high quality semantic decomposition was a skilled and expensive activity, and effort expended towards it needed to be traded off against the need to apply effort to other important properties of a finished termi-

nology, such as ‘synonym purity’ and ‘hierarchy induced ambiguity’.

Rogers [10, 11] described a cross-validation approach where the semantic definitions of CTV3 were re-expressed within a richer ontology (GALEN) and description logic framework, following which the concept space was automatically re-classified, and the newly inferred hierarchy compared with that originally asserted. The re-expression of CTV3 was achieved using a methodology based on an intermediate representation and subsequent transformations to normalise the semantics of candidate expressions, taking into account metamodel conventions and preferred forms. The comparison phase of the study detected errors of omission within both the CTV3 source corpus and the GALEN ontology, as well as highlighting where genuinely different world views led to disagreement over the correct classification of a term (e.g. whether thymectomy should no longer properly be classified as an endocrine procedure, and whether oophorectomy should be).

Campbell [22] advocated the need not only for processes to ensure quality, but also metrics to monitor the effect of those processes. Spackman describes the use of ‘lexically suggested logical closure’ (LSLC) as a metric to monitor the development of SNOMED RT [23]. Advanced lexical tools suggested candidate semantic relationships between concepts. These were manually reviewed and either approved or rejected. A metric of the quality improvement process was thus derived: the proportion of all candidate relationships that had been reviewed, tracked over time.

Both Campbell and Schulz observed that, whilst build-time QA processes and metrics are useful and important, the scale of the overall undertaking meant they could never entirely prevent run-time problems such as achieving consistent interpretation of semantics across multiple end users. They believed that a significant part of the QA could only be done after the terminology was in use, to be performed by recruiting real clinical end users to feed back errors as they were encountered in use.

Three papers compared established clinical terminologies and newer schemes in development with respect to their ability to ex-

press typical clinical concepts [24-26]. Each study involved groups of clinicians attempting to represent typical clinical data using the old, and new, terminologies. A variety of scoring systems were described in order to distinguish between where the terminologies on trial provided an exact match, no match or some intermediate level of expressivity.

Rogers reported an approach in which an analysis of the routine coding behaviour of small groups of clinicians over a period of time was compared with the aggregated behaviour of a larger group of similar clinicians over the same period, in order to detect statistically significant different recording patterns that may indicate systematically idiosyncratic interrater differences [27]. He cited such differences as one barrier to the successful implementation of an experimental decision support software system to support the review of repeat prescribing in primary care [8].

Ceusters reported a combined approach using proprietary ontological and linguistic reasoning to detect inconsistencies in SNOMED-CT [28]. One component of the approach creates a probe construct, and returns a set of concepts within SNOMED-CT computed to be semantically similar to the probe, together with a measure of the semantic distance between the probe and each member of the similar set. Extreme values for semantic distance may indicate concepts only deemed similar because they were incorrectly modelled. A second component reclassifies SNOMED-CT content within an independent ontology (LinK-Base™) and compares the newly inferred hierarchy with the received SNOMED CT hierarchy.

Bodenreider describes a methodology for the indirect detection of various types of ontological errors in SNOMED-CT through a structural analysis of the subsumption hierarchy of its classes in an ontology [29]. For example, instances where an entity has only a single child indicate where Linnaen classification principles are not being followed: either the child is redundantly the same as the parent, or if it has true differentiae then at least one other unstated child must exist with alternative differentia and the set of all children should exhaustively cover the domain of the parent.

Cornet and Arts set out four requirements that an ontology should satisfy, including that its content should be correct, complete and not contradictory [30, 31]. In the context of this paper's division of the overall task into validating the axioms, and validating the reasoner, Cornet's fourth requirement (any process reasoning over the ontology should be competent and efficient) is properly part of the latter task and not considered further here.

ISO TC215 WG2 has approved ASTM E2087 as a technical specification (TS17117) describing high level indicators of quality in controlled medical terminologies [32]. These include whether the terminology includes redundant, ambiguous or inconsistent concepts, whether the terminology has a stated purpose and whether its coverage supports that purpose, as well as a variety of other properties including length of update cycle, mappings to other schemes, local extensibility, expressivity with respect to a standard corpus and whether an effective user interface to it can be constructed.

General Ontologies

Gómez-Pérez identified five properties of an ontology's content: consistency, completeness, conciseness, expandability (the effort needed to extend an ontology without invalidating it) and sensitiveness (the extent to which the validity of an ontology may be affected by its subsequent extension) [18, 33, 34]. In more recent work, Gómez-Pérez categorises possible content errors into circular definitions, partition errors, inconsistent semantics, incomplete classification and redundancy [19].

Uschold put forward a unified methodology for ontology engineering [35]. This included an evaluation phase comprising testing against general ontological criteria such as clarity, consistency, coherence, extensibility and reusability, as well as application-specific criteria including whether it satisfies informal competency questions [36].

More recently, the EC-funded WonderWeb consortium recommends in DOLCE, OCHRE and BFO a range of philosophically coherent distinctions that should exist

between the high-level classes or relationships in any well-formed ontology [37, 38]. These works proscribe a principled division of entities in the world into categories such as universals, perdurants or endurants, rigid or anti-rigid, quality or quale. These categories may be linked by a similarly proscribed set of semantic relationships such as proper-part, overlaps or partially-contained-in.

Donnelly has proposed formal theories of locative, partonomic and containment relations and described how two existing large scale anatomical ontologies (*OpenGALEN* and the Foundational Model of Anatomy) are logically and philosophically ambiguous with respect to those theories [39, 40].

Several researchers have described tools and methodologies (e.g. *OntoClean*) by which an ontology may be examined for compliance with such principles [41-45]. Welty reported an experiment in which the performance of an intranet search engine was improved after the *OntoClean* methodology [46] had been applied to its previously unprincipled ontology.

Brewster describes an approach and algorithms by which the expressivity of an ontology may be automatically compared with the set of terms or relations extracted from a representative corpus from the same domain, using natural language processing information extraction technologies [47]. Where the ontology can not express terms or relations encountered in the analysed corpus, this may indicate that it is incomplete.

The EC-funded KnowledgeWeb Network of Excellence [48] observed in 2005 that no robust methodology exists for determining which ontology provides the best fit for a given purpose, such as for the semantic web. Several possible methodologies are described, including *OntoMetric*, *Natural Language Application Metrics (NLAM)*, *OntoClean* and *Evalexon*.

Discussion: Themes Identified

Several different types of ontology error were identified in the literature review. Following a thematic analysis, I propose below four distinct properties of an ontology with in which errors may occur:

Philosophical Rigour

The philosophy community advocate that the design phase of any ontology should include a proper treatment of philosophical principles in order to avoid the propagation of significant errors during its implementation. Outputs of contemporary philosophy research over the last decade, aimed specifically at biomedicine, include upper level ontologies such as DOLCE and BFO, formal theories of mereonomy, and methodologies such as *OntoClean* for applying these principles *post hoc* to existing ontologies. However, despite 2500 years of philosophy research [49], some notions central to practical medical ontologies remain robustly contentious. Cognitivist philosophers would represent that an organ *is_made_of* tissue; realists reject the validity of *is_made_of* and hold that an organ *is_identical_to* its tissue. Realists further reject the notion of 'concept' as defined by ISO and which underpins biomedical ontology works such as UMLS and SNOMED CT, in favour of 'universals' [50].

Ontological Commitment

Most notions can be represented in more than one logical and/or semantic form, but humans easily recognise these to be equivalent. By contrast, consistent manipulation of representations and data by logical computational systems is dependent on one possible form – or a group of forms – being designated 'canonical'. Whilst some transformations from variant to canonical form can be achieved through logic alone, other variant forms are not equivalent in any logical sense: transforming "Fixation of femur by means of insertion of pins" to an arbitrarily preferred form such as "Insertion of pins to fixate femur" [51] requires the application of metamodel conventions. Ensuring that all present and future ontology content is consistent with such conventions is an ontological commitment that must be rigorously applied both within the ontology itself, and across all applications that rely on the ontology for interoperability [11].

Content Correctness

Whilst most reported methodologies restrict themselves to the quality and coverage of the knowledge content represented directly in the ontology, it is also appropriate to evaluate the content metadata. At least two distinct metadata types may be considered:

- annotations concerning the provenance or epistemology of content (e.g. information about authors, literature citations, knowledge review dates, valid jurisdictions),
- lexical annotations of the content (e.g. display and search strings).

In evaluating content quality, Cornet and Gómez-Pérez [19, 30] each provided some clarification of what it means for content to be correct or incorrect. Building on their work, I propose an expanded matrix for evaluating the ‘correctness’ of ontology content (both knowledge and metadata) with respect to our understanding of the real world:

- Is the content true?
Types of error: errors of fact, partition errors
e.g. ‘structure of labial vein’ erroneously declared as a kind of ‘structure of vein of head’ [28]. The island of Aruba simultaneously classified as an island of France, England and The Netherlands.
- Is the content complete (is every truth that can be represented actually represented, or able to be inferred)?
Types of error: errors of omission, ambiguous representation
e.g. although the testis is an endocrine gland, testicular surgery is not normally given as a subclass of endocrine surgery in clinical classifications [10]; the thymus gland secretes immunorestorative hormones, but some medical classifications do not classify it as an endocrine organ, although they classify thymosins as ‘thymic hormones’.
- Is the content concise (parsimonious)?
Types of error: semantic redundancy – explicit inclusion of axioms that could be inferred by a particular reasoner
e.g. some part of the heart may be stated to be both a kind of heart structure and a kind of cardiovascular organ structure,

though the latter is inferable from the former.

- Is the content internally consistent?
Types of error: contradictory axioms, semantic duplication, circular definition
e.g. the same ontology may include a concept called ‘traumatic unilateral amputation’ and a different one called ‘unilateral traumatic amputation’ [28]; endocrine surgery may be defined as ‘surgery performed by endocrine surgeons’ and endocrine surgeons as ‘surgeons that perform endocrine surgery’.

Fitness for Purpose

An ontology that complied with all the current philosophical recommendations, defined and complied with all necessary ontological commitments, and whose content was entirely correct, may yet be too complex to be directly usable or useful by anybody except the very highly trained [11, 52]. A significant property of an ontology, therefore, must be its fitness for purpose.

Clinical ontologies are increasingly being either co-opted or designed from scratch to function as one software component in a complex information management software system intended to support and individualise clinical care as it happens, rather than to monitor clinical care in aggregate after the fact. In this context, mainstream software development methodologies to test and evaluate fitness for purpose should be applied to the engineering of ontologies both as discrete components and as components fully integrated within larger systems. A complete discussion of such methodologies is outside the scope of this paper (see e.g. [53, 54]) but should include as a minimum both functional and usability field testing.

Discussion: a Framework for Ontology Quality Assurance

The preceding analysis suggests a methodology whereby programmes to assure, or improve, the quality of specific ontologies may themselves be quality-assured. Such sys-

tems may be scrutinised to determine whether they address all quality attributes identified above: philosophical validity, meta-ontology compliance, content truth, content completeness, content parsimony, internal consistency of content, and fitness for purpose. Such an evaluation, however, must be sympathetic to the significant barriers that stand in the way of successfully assuring the quality of some attributes.

Barriers to Assuring Content Quality

Any review of an ontology quality assurance programme must recognise two inescapable limitations that compromise our ability to quality assure the content of an ontology as ‘correct’. The first and primary limitation is that no gold standard exists, or will ever exist, representing the truth, the whole truth and nothing but the truth. This inherently denies us the possibility to automatically cross-check specific ontologies against such a gold standard in order to exhaustively list their errors. As a consequence, whilst ontologies may be cross-checked against each other (provided no commercial or intellectual property barrier exists to this activity), strategies to manually inspect and evaluate the content of individual ontologies remain necessary whilst the correctness of all other ontologies remains unknowable. The second limitation follows from the first: manual quality improvement strategies are known to be inaccurate (e.g. due to reporter fatigue), but we can never determine their precision or recall in the absence of such a gold standard.

Given these limitations, the content of an ontology can never be proven to be either true, complete or internally consistent. Further, there is no possibility to determine exactly how incorrect an ontology is. Finally, if no individual ontology can be assigned a measure of content correctness relative to a gold standard, it is correspondingly difficult to interpret any comparison of correctness between ontologies. Statements of the form ‘ontology A is more correct than ontology B’ have only limited value: better than worse is not necessarily good.

Notwithstanding that any practical ontology is unlikely to be correct, it is also un-

likely to be either maximally concise or complete. Indeed, a certain amount of content redundancy may be deliberately included to allow for more efficient implementations, whilst maximal completeness is seldom cost-effective: much of the content would never be used.

Limitations of Quality Metrics

Though absolute or comparative quality measures of ontology content may be precluded as described in the previous section, the absolute rate of change in quality may be more accurately determined. For example, it may be calculated as the number of axioms later manually identified to be false as a percentage of all those stated to be true at a prior point in time [22, 55]. However, whilst any positive detection rate for a given ontology inspection strategy is certainly evidence that content quality is being improved, this improvement is from an immeasurable baseline value towards an indeterminate goal. Although error detection rates may reduce with time, this does not necessarily mean the absolute number of all types of content error left to find is falling at the same rate; when no errors are found, this does not imply none exist.

A further problem with quality metrics is that they typically describe the quality of the entire ontology in aggregate, whereas in practice the quality of an ontology may vary widely across the subdomains within it. This may occur for example where a large ontology is constructed by multiple authors, with each individual author carrying responsibility for one area of the ontology but with different skill. Uneven quality across an ontology also necessarily arises because all ontologies have finite coverage, typically comprising a high quality 'core' covering a central chosen domain of coverage, but necessarily referencing small numbers of entities from multiple peripheral domains that – pragmatically – can only be represented with lesser quality (for example, an ontology of diabetes will reference anatomical concepts, but is unlikely to require a detailed or highly principled anatomy ontology for this purpose) [11].

If measuring ontology content quality is difficult, the interpretation of such metrics

is also problematic. How much quality is required? What types of error are more serious, particularly with respect to fitness for purpose if this is the only absolute requirement? Whilst pragmatic compromises on quality as an ideal are common – for example allowing redundant or incomplete content – little is known about the trade-off between quality and operational performance. Both semantically redundant and factually incomplete content expose an ontology to the racing certainty that purely logical algorithms for detecting semantically equivalent expressions will fail some of the time, whilst incomplete content also brings the risk of incomplete inferred classification (post-coordination).

How much quality is required, and the consequences of any trade-off, seem likely to differ for different intended applications. But determining the nature of any potential mismatch between the quality of an ontology and the needs of a specific application is often difficult: in the absence of a standard means of either representing or measuring ontology quality (both in terms of what is required, and what is available), the best information regarding an ontology's quality – and of how quality varies within that ontology – may be little better than an inchoate mental impression known only to its author(s) but that can not be communicated, whilst the ontological needs of a given application may be similarly held only in the mind of its developer.

Fitness for Purpose

Most of the reported operational (as opposed to structural) evaluations of clinical terminologies or ontologies have attempted to measure their expressiveness – their ability to support typical clinical recording tasks. However, a common feature of most of these studies is that they are small scale, and conducted under laboratory conditions. Reports of large scale field evaluations are lacking; this may in part reflect the familiar problem that confidence and adoption of any new product is typically low until it has been extensively field tested, but such testing is impossible without high uptake.

The interpretation of the results of such operational studies as exist is also problematic: what levels of expressivity, and inter-rater agreement, are actually required for a given ontology to be useful, or acceptable, to a given group of end users and for a specific purpose?

Finally, consider that the true value of a clinical ontology is increasingly only evident when it is tightly integrated within a complex healthcare information system. If that entire system fails or is rejected, how are we to characterise the specific nature of any deficiencies in its ontology component and the extent of their contribution to the general failure? Friedland has reported some initial steps towards a framework for analysing failures in knowledge representation and reasoning systems [56].

Conclusion

The problem of assuring the quality of ontologies has been recognised for over a decade. However, the medical and non-medical ontology engineering communities have yet to define, much less to regularly practise, a comprehensive and systematic methodology for assuring, or improving, the quality of their product.

Whilst software tools such as WebODE [57], OntoEdit and ODEval [48] are beginning to address the need for tools to support the entire ontology engineering process, including quality assurance, these tools currently address only a subset of the possible properties of an ontology identified in this review (e.g. in WebODE's case, only its philosophical characteristics) and thus are still short of a comprehensive solution.

A significant and insoluble problem for ontology engineering and quality assurance is the absence of a gold standard against which to objectively, systematically or exhaustively determine either the correctness of an ontology's content at a point in time, or the effects of any quality improvement methodology. Even if such metrics could be calculated, little is known of how quality trades off against either the effort required to build and maintain an ontology, final user acceptability of it, or its practical performance.

Fitness for purpose is considered by some to be the only quality indicator that really matters, but its direct measurement and interpretation is problematic. In this paper I have described several other candidate indicators of quality (philosophical rigour, ontological commitment and correctness of content), yet the extent to which any of these predicts fitness for purpose in the field is unknown.

In the continued absence of a significant body of usability or reliability field studies, the clinical terminology wars will be obliged to continue raging largely by reference only to increasingly specialist theories of ontology construction. The debate would benefit greatly from being informed, and balanced, by analytical studies (rather than case reports) of practical field experiences.

In particular, the performance of clinical information systems driven by ontology-based data will depend ultimately not on the theoretical or in-laboratory properties of that ontology, but the practical ability of clinical users to wield it. Inter alia, therefore, research is required to characterise the epidemiology and implications of interrater variability amongst ontology end-users, and to determine those ontological properties that assist (or are a prerequisite for) strategies to manage interrater variability and those that actively cause it.

Acknowledgement

This work is supported by the 'SemanticMining' Network of Excellence, funded by the European Commission Information Society Technologies programme, project no. 507505.

References

- Borgida A, Patel-Schneider PF. A Semantics and Complete Algorithm for Subsumption in the CLASSIC Description Logic. *J Art Intell Research* 1994; 1: 277-308.
- Borgida A. On the relative expressive power of Description Logics and Predicate Calculus. *Artificial Intelligence* 1996; 82: 353-67.
- Rogers JE, Roberts A, Solomon WD, van der Haring E, Wroe CJ, Zanstra PE, Rector AL. GALEN Ten Years On: Tasks and Supporting tools. In: Patel V, et al. (eds). *Proc MEDINFO 2001*. IOS Press 2001; pp 256-60.
- Bada M, McEntire R, Wroe C, Stevens R. GOAT: The Gene Ontology Annotation Tool. In: *Proc 2003 UK e-Science All Hands Meeting*. Nottingham, UK; pp 514-9.
- Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press 2003. ISBN: 0521781760.
- Spackman KA. Normal forms for description logic expressions of clinical concepts in SNOMED RT. *Proc AMIA Symp 2001*: 627-31.
- Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. *Proc AMIA Symp 2002*: pp 712-6.
- Rogers JE, Wroe CJ, Roberts A, Swallow A, Stables D, Cantrill JA, Rector AL. Automated quality checks on repeat prescribing. *Br J Gen Pract* 2003; 53 (496): 838-44.
- Wroe CJ, Stevens R, Goble CA, Ashburner M. A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. *Pacific Symposium on Biocomputing 2003*; 8: 624-35.
- Rogers J, Price C, Rector A, et al. Validating Clinical Terminology Structures: Integration and Cross-Validation of Read Thesaurus and GALEN. In: *Proc AMIA Symp Orlando FL 1998*. Philadelphia, PA: Hanley & Belfus Inc.; 1998. pp 845-9.
- Rogers JE. Development of a methodology and an ontological schema for medical terminology. MD Thesis, University of Manchester 2005. <http://www.cs.man.ac.uk/mig/people/jeremy/papers/2004-Rogers-MD-Thesis.pdf> (visited Aug 9, 2005).
- Gruber TR. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Int Journal of Human-Computer Studies* 1993; 43: 907-28.
- Chandrasekaran B, Josephson JR, Benjamins VR. What are ontologies, and why do we need them? *IEEE Intelligent Systems* 1999; 14: 20-6.
- Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Methods Inform Med* 1998; 37: 394-403.
- Goble C, Wroe C. The Montagues and the Capulets. *Comp Funct Genom* 2004; 5: 623-32.
- Baader F. Restricted Role-value-maps in a Description Logic with Existential Restrictions and Terminological Cycles. *CEUR Workshops Proc Description Logics 2003*. <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-81/> (visited Aug 9, 2005).
- Schulz EB, Barrett JW, Price C. Read Code Quality Assurance: From Simple Syntax to Semantic Stability. *J Am Med Inform Assoc* 1998; 5 (4): 337-46.
- Gómez-Pérez A. From Knowledge Based Systems to Knowledge Sharing Technology: Evaluation and Assessment. *Knowledge Systems Laboratory Abstract KSL-94-73* (1994). http://www.ksl.stanford.edu/KSL_Abstracts/KSL94-73.html (visited Aug 9, 2005).
- Gómez-Pérez A. Evaluation of Ontologies. *Int J Intell Systems* 2001; 16: 391-409.
- Fernández-López M, Gómez-Pérez A. A survey on methodologies for developing, maintaining, evaluating and reengineering ontologies. *Ontoweb consortium: Deliverable 1.4* (2002). <http://ontoweb.org/About/Deliverables/D1.4-v1.0.pdf> (visited Aug 9, 2005).
- Schulz EB, Barrett JW, Price C. Semantic quality through semantic definition: Refining the Read Codes through internal consistency. *J Am Med Inform Assoc* 1997; 4 (Symp Supp): 615-9.
- Campbell KE, Tuttle MS, Spackman KA. A "lexically-suggested logical closure" metric for medical terminology maturity. *Proc AMIA Symp Philadelphia: Hanley and Belfus; 1998*. pp 785-9.
- Spackman KA, Campbell KE. SNOMED RT: A Reference Terminology for Health Care. *Proc AMIA Symp 1997*; pp 640-4.
- Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. *J Am Med Inform Assoc* 1996; 3: 224-33.
- Brown PJB, Odusanya L. Does Size Matter? – Evaluation of Value Added Content of Two Decades of Successive Coding Schemes in Secondary Care. In: Bakken S (ed). *Proc AMIA Symp Philadelphia: Hanley and Belfus; 2001*. pp 71-5.
- Brown PJB, Warmington V, Laurence M, Prevost AT. Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice. *BMJ* 2003; 326: 1127.
- Rogers JE, Wroe CJ, Roberts A, Rector AL, Swallow A, Stables D, Cantrill JA. Feasibility and Reliability of Computerised Review of Long Term Prescribing: Final Report (2002). <http://www.cs.man.ac.uk/mig/projects/old/prescribing-indicators/FinalReportRTF.zip>. Supplementary Report pp 88-96 (visited Aug 9, 2005).
- Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. In: Fieschi et al. (eds). *Proc MEDINFO 2004*. IOS Press 2004; pp 482-6.
- Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: A Case Study in SNOMED CT. In: *Proc KR-MED 2004*. Whistler: pp 12-20. <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-102/> (visited Aug 15, 2005).
- Cornet R, Abu-Hanna A. Description logic based methods for auditing frame-based medical terminology systems. *Artificial Intelligence in Medicine* 2005; 34: 201-17.
- Arts DGT, Cornet R, de Jonge E, de Keizer NF. Methods for evaluation of medical terminological systems; a literature review and a case study. *Methods Inf Med* 2005; 44: 616-25.
- ISO/TC215 WG 3 Standard Specification for Quality Indicators for Controlled Health Vocabularies 2000 July. Report No.: TS17117.
- Gómez-Pérez A. Some Ideas and Examples to Evaluate Ontologies. *Knowledge Systems Laboratory Abstract KSL-94-65* (1994). http://www.ksl.stanford.edu/KSL_Abstracts/KSL-94-65.html (visited Aug 9, 2005).
- Gómez-Pérez A. Criteria to Verify Knowledge Sharing Technology. *Knowledge Systems Laboratory Abstract KSL-94-65* (1995). <http://www.ksl>

- stanford.edu/KSL_Abstracts/KSL-95-10.html (visited Aug 9, 2005).
35. Uschold M. Building Ontologies: towards a unified methodology. In: Proc Expert Systems. 16th Annual Conference of BCS SG on Expert Systems 1996.
 36. Grüninger M, Fox MS. Methodology for the Design and Evaluation of Ontologies. In: IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, April 13, 1995.
 37. Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A. Wonderweb: Deliverable 18. Ontology Library. WonderWeb consortium (2003). <http://wonderweb.semanticweb.org/deliverables/D18.shtml> (visited Aug 9, 2005).
 38. Grenon P, Smith B, Goldberg L. Biodynamic Ontology: Applying BFO in the Biomedical Domain In: Pisanelli (ed). Ontologies in Medicine. Amsterdam: IOS Press; 2004. pp 20-38.
 39. Donnelly M, Bittner T, Rosse C. A Formal Theory for Spatial Representation and Reasoning in Biomedical Ontologies. *AI Med* 2005; 36 (1): 1-27.
 40. Donnelly M. Containment Relations in Anatomical Ontologies. In: Friedman C, Ash J, Tarczy-Honoch (eds). Proc AMIA 2005 Annual Symposium Washington DC. pp 206-10.
 41. Guarino N, Welty C. Evaluating ontological decisions with OntoClean. *Communications of the ACM* 2002; 45 (2); 61-5.
 42. Gangemi A, Guarino N, Masolo C. Sweetening ontologies with DOLCE. In: Gomez-Perez, Benjamins (eds). EKAW 2002. Proc 13th Int Conf Knowl Eng & Knowl Mangmt. Ontologies and the Semantic Web. Heidelberg: Springer Verlag; 2003. pp 166-81.
 43. Flett A, Casella dos Santos M, Ceusters W. Some Ontology Engineering Processes and Their Supporting Technologies. In: Gomez-Perez, Benjamins (eds). EKAW 2002. Proc 13th Int Conf Knowl Eng & Knowl Mangmt. Ontologies and the Semantic Web. Heidelberg: Springer Verlag; 2003. pp 154-65.
 44. Casalla de Santos M, Dhaen C, Fielding M, Ceusters W. Philosophical scrutiny for run-time support of application ontology development. In: Proc FOIS 2004, Turin.
 45. Fielding JM, Simon J, Ceusters W, Smith B. Ontological theory for ontological engineering: biomedical systems information integration. In: Principles of Knowledge Representation and Reasoning. Proc KR2004, Whistler. AAAI Press 2004; 116; 647-52.
 46. Welty CA, Mahindru R, Chu-Carroll J. Evaluating Ontological Analysis. In: Proc ISWC-03 (2003). CEUR-WS vol 82. <http://CEUR-WS/Vol-82> (visited Aug 9 2005).
 47. Brewster C, Alani H, Dasmahapatra S, Wilks Y. Data Driven Ontology Evaluation. In: Proc Int Conf Language Resources and Evaluation. Lisbon, Portugal, 2004.
 48. Hartman J, Spyns P, Giboin A, Maynard D, Cuel R, Suárez-Figueroa C, Sure Y. KnowledgeWeb D1.2.3 Methods for Ontology Evaluation. (2005). <http://www.starlab.vub.ac.be/research/projects/knowledgeweb/KWeb-Del-1.2.3-Revised-v1.3.1.pdf> (visited Aug 9, 2005).
 49. ECOR European Centre for Ontological Research. <http://www.ecor.uni-saarland.de> (visited Aug 9, 2005).
 50. Smith B, Ceusters W, Temmerman R. Wüsteria. In: Proc MIE2005, Geneva. *Stud Health Technol Inform* 2005; 116: 647-52.
 51. O'Neil M, Payne C, Read J. Read Codes Version 3: A user led terminology. *Methods Inf Med* 1995; 34: 187-92.
 52. Spackman KA, Reynoso G. Examining SNOMED from the Perspective of Formal Ontological Principles: Some Preliminary Analysis and Observations. in Proc KR-MED 2004 Whistler; 12-20 <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-102/> (visited Aug 15 2005)
 53. Kitchenham B, Pflieger SL. Software Quality: The Elusive Target. *IEEE Software* 1996; 13 (1): 12-21.
 54. Kitchenham B, Jones L. Evaluating software engineering methods and tool (Twelve part series). *ACM SIGSOFT Software Engineering Notes* 1996-8; 21: 1; 11-14 to 23; 5: 21-24.
 55. CAP (College of American Pathologists). SNOMED Clinical Terms User Guide. July 2004 Release.
 56. Friedland NS, Allen PG, Witbrock M, et al. Towards a quantitative, platform-independent analysis of knowledge systems. In: Principles of Knowledge Representation and Reasoning. Proc KR2004, Whistler. AAAI Press 2004; 116: 507-15.
 57. Fernández-López M, Gómez-Pérez A. The integration of OntoClean in WebODE. In: CEUR Workshop Proceedings. Amsterdam, The Netherlands 2002; 62: 38-52. <http://CEUR-WS.org/Vol-62> (visited Aug 9, 2005).

Correspondence to:

J. E. Rogers
 BioHealth Information Group
 School of Computer Science
 University of Manchester
 Manchester, M13 9PL
 United Kingdom
 E-mail: jeremy.e.rogers@manchester.ac.uk