Issues with evaluating and using publicly available ontologies

Yannis Kalfoglou
Advanced Knowledge Technologies (AKT)
School of Electronics and Computer Science
University of Southampton, UK
y.kalfoglou@ecs.soton.ac.uk

Bo Hu
School of Electronics and Computer Science
University of Southampton, UK
bh@ecs.soton.ac.uk

ABSTRACT

The proliferation of ontologies in the public domain and the ease of accessing them offers new opportunities for knowledge sharing and interoperability in an open, distributed environment, but it also poses interesting challenges for knowledge and Web engineers alike. In this paper we discuss and analyse those challenges with emphasis on the need to evaluate publicly available ontologies prior to use. We elaborate on a number of issues ranging from technological concerns to strategic and political issues. We drawn our experiences from the field of ontology mapping on the Semantic Web, a necessity that enables many of Semantic Web's proclaimed features.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic Networks*; D.5.3 [ ]: Information Interfaces and PresentationGroup and Organization Interfaces[Web-based Interaction]; K.8.1 [**Personal Computing**]: Application Packages—*Freeware/Shareware*

Keywords

ontology, ontology evaluation

General Terms

ontology certification, emergent semantics, ontology management

1. INTRODUCTION

Before ontologies became popular, knowledge engineers hardly ever had to work with more than one ontology at a time. Even in cases where multiple ontologies were used (see for example, [5]), these were mostly controlled experiments [27] in moderated environments [9]. Nowadays however, the practice is somewhat different. Modern trends in knowledge management dictate that we should expect to work more and more within distributed and open-ended environments like the Web, and its ambitious extension, the Semantic Web (hereafter, SW). That fact alone, has had a significant impact on ontology engineering; the most prominent changes we observe are: (a) sourcing ontologies, which is far easier today than it was in the recent past, (b) the need to use more than one ontology as a result of aiming to achieve knowledge sharing in a distributed environment, like the SW, (c) knowledge engineering practice is difficult to enforce when dealing with outsourced ontologies, (d)the sociopolitical context of an open-ended distributed environment needs to be taken into account when using ontologies.

These phenomena are intimately connected with knowledge representation (hereafter, KR) and have an impact on KR practice. But, we are not interested to analyse them in depth as they are too broad, KR specific, and peripheral to our discussion. Rather, we would like to elaborate on a practical, engineering issue that affects the use of ontologies in pragmatic cases on the SW: evaluation. For the sake of narrowing down the argument we distinguish between two phases of evaluation: during design and development and prior to use. We will not discuss the former phase as it is an ontology development issue. The literature provides us with methodologies (see, for example, the early work in [10]) and experience reports on how ontology engineers deal with evaluation issues prior to releasing ontologies. Recent reports from large scale research projects also distinguish between different phases of evaluation (see, for example, [14]). The phase that we are interested in, is the one we believe poses subtle challenges: evaluation of ontologies before use in an application on the SW. One important point of clarification here is that we are referring to ontologies that we have not authored but merely outsourced from the public domain, and that these could well have been evaluated before made public. But, we still face the dilemma of whether to further evaluate them or blindly trust them and put to use straightaway. This is the dilemma we analyse in this paper.

We first discuss the trigger that brought to bare these challenges: how ontologies became so easy to access and retrieve from the public domain and why it matters for engineers (section 2). We then focus on more technical and specific issues and we discuss our experiences with evaluation issues drawn from SW application cases we were involved (section 3), before presenting a specific case on ontology mapping with regard to two large scale medical ontologies 3.1. In section 4 we elaborate on various ways of alleviating the problem and we conclude the paper in section 5 with a draft roadmap for further work in this area.

WWW2006, May 22–26, 2006, Edinburgh, UK.
.

2. ON PUBLICLY AVAILABLE ONTOLOGIES

Historically speaking, ontologies - in their modern computer science reincarnation - emerged in the Artificial Intelligence (hereafter, AI) community as a means to share knowledge between knowledge bases. In those days, ontologies were hard to built, very specialised, formal in nature, and even harder to find in the public domain. In the mid nineties, communal input made frameworks like Ontolingua ([9]) popular and ontologies begun to appear in the public domain. The emergence of the SW and the technologies that underpin it (like, for example, the W3C's OWL family[1] of languages or the RDF format[2]), along with the portrayed role of ontologies on the SW (see the visionary article in [4]), triggered an unprecedent influx of ontologies to the public domain.

Nowadays, the Google culture for seeking information on the Web has influenced other specific information seeking, like for example, searching for ontologies on the Web with tools like SWOOGLE[3]. Furthermore, large scale European and US projects and distinct research centres in this field have begun publishing a plethora of ontologies on a variety of domains. Popular ontology building tools (like Protege) and their communities also make and share a lot of ontologies in the public domain. In short, if one needs to find an ontology, it won't be a problem!

Despite the abundance of ontologies, however, we see this phenomenon more skeptically when it comes to an engineer's point of view. There are certain assurances and features that an engineer would like to see in a publicly available ontology before putting it to use. As we said in the introduction, we are focussing on a specific scenario where we outsource an ontology from the public domain and we use (or more technically speaking, re-use) it in an application. In such a case, an engineer would like to know specific background information about the ontology. This sort of information could be found (semi-) automatically by using, for example, a catalogue style characterisation of ontologies that was originally proposed in [28] and further elaborated in [20] or by calling upon advanced techniques, like ontology search and ranking [1]. Assuming that this step is achievable, the engineer would then like to have some sort of a formal, and mechanised way of verifying the validity of the ontology. That could range from the straightforward check of syntactic soundness of an ontology (which is easily done with popular syntax checkers, like OWL validators) to the much harder conceptual correctness which involves thorny issues of KR and domain coverage. While the former is at an advanced stage with a variety of tools available the latter is still at an early stage of research [15]. If we further assume that this step will also be achievable, then we would have reach at a stage where the remaining bits of plugging-in the outsourced ontology are algorithmic and mundane, given the import facilities of most ontology editors and APIs.

The crux of the problem is that the script outlined above is a hypothetical one. It is based on assumptions that are hard to materialise on the SW, as we speak. For instance, we do not have the experiences needed to develop and deploy catalogue style characterisations of ontologies[4]. Even worse, we do not have consensual views on what should go in those catalogues and how ontologies should be classified. Furthermore, the work on search and ranking is at an early stage of research but it could prove to be a promising one. The most difficult issue though, seems to be the verification of conceptual correctness of an ontology. It is a problem identified in the nineties (referred to as "metaphysical consistency" in [12]) and prevails in today's engineering efforts when we try to achieve interoperable systems. Ensuring that an ontology covers the domain at question adequately and the conceptualisation is a credible reflection of the real world is a task that is difficult to automate, prone to constant changes and should involve practitioners and domain experts, not only engineers. Experience reports highlight the necessity for such a collaboration of different stakeholders, for example, in an interdisciplinary meeting on semantic interoperability [18], it was argued that:

> [...]domain ontologies need to be built and vetted by domain experts and scientists, as those built by computer scientists were usually rejected.

We will elaborate on the role of domain experts and community input in section 4. Regarding the verification of the syntactic soundness of an ontology it seems that it is a straightforward engineering task, given the plethora of validators available.

How do all these issues affect us, the engineers, in practice? In the next section, we will try to answer this question by instantiating some of these issues in the context of building, (re-) using, and deploying ontologies in real world scenarios on the SW.

3. OUR EXPERIENCES

Five years ago, the UK's Engineering and Physical Sciences Research Council (EPSRC) funded an Interdisciplinary Research Collaboration (IRC) consortium comprising of five leading British Universities to research Advanced Knowledge Technologies (AKT)[5]. AKT's original focus was on how knowledge engineering and AI techniques can improve Knowledge Management (KM) practices. Over the years, we adopted our main focus to the SW. As part of our research agenda, we designed, developed and deployed ontologies in large contexts and in support of real world cases[6]. Our experiences with evaluation issues developed around the CAS (Computer science AKTive Space), a dedicated portal that allows semantically enriched exploration of a domain [25], the construction of a scalable RDF storage system, 3Store [13], and ontology mapping technologies, like CMS (CROSI Mapping System[7] - [16]) and IF-Map (Information Flow based ontology mapping - [17]). In particular, when developing CAS we encountered the problem of *verifying the appropriateness* of externally sourced ontologies. Similarly, 3Store development had to deal with the problem of *referential integrity* (or co-reference resolution as we call it) for external resources. In the ontology mapping domain, we faced the problem of *trusting the external ontologies* and *checking*

[1]http://www.w3.org/2004/OWL/
[2]http://www.w3.org/RDF/
[3]http://swoogle.umbc.edu/

[4]One could argue that Ontolingua was such a catalogue but the ones we are envisaging need to be richer in content.
[5]More on www.aktors.org
[6]See, for example, http://www.aktors.org/technologies/
[7]http://sourceforge.net/projects/ontologymapping

their validity in order to produce credible mappings. We elaborate on each of these issues in subsections 3.1, 3.2 and 3.3, respectively.

3.1 Verify appropriateness

CAS application, an award-winning application for the SW[8], is using the AKTive portal and AKTive support ontologies[9]. We used external ontologies to enrich the ones we had at our disposal in order to get a more accurate view of the domain at question. As we wanted to experiment with re-use of ontologies in a real world case, we opted for external ontologies that we had no authority over. That introduced us to the challenge of making sure that the ontology we wanted to re-use is appropriate for the domain at question. Even if CAS is concerned with the domain of computer science in academia, a domain that we are familiar with, outsourcing ontologies that claim to model computer science in academia was not easy. Subtle differences in representation of real world concepts and inaccuracies were difficult to spot. Solutions to this problem ranged from semi-automatic heuristics to "clean" the imported ontologies to more mundane tasks involving many manual checks by engineers [25]. In cases where we decided to re-use ontologies that we had authority over, like the locally developed computer science ontologies, we had to spend a lot of time debating representational issues and reconciling diverse real world conceptualisations.

In the latter case, evaluation was performed at both phases as introduced in section 1, that is, at design and development time (since developers were easy to reach and part of the team) and at deployment and re-use time (since we had full control and authority of the SW application that was using the ontologies). The situation was somewhat different in the former case. It wasn't easy to evaluate external ontologies due to lack of editorial authority. Furthermore, we could only test the ontology once we imported it in the existing ones. Overall, we found that verifying the appropriateness of an ontology is not an easy to automate task and requires plenty of time and technical compromises. Our experiences shown that evaluating externally defined ontologies, specifically when verifying their appropriateness, is an issue that should be considered prior to use them as automating the evaluation task looks highly unlikely with the current state-of-the art.

3.2 Co-reference resolution

An issue closely related to the aforementioned, was that of using external data. In the 3Store scenario, we had to deal with millions of externally sourced RDF triples. These were instances deemed necessary to operationalise the CAS ontologies. Defining all those instances though, is a cumbersome task and as it is natural in an application domain like the SW, we outsourced them. For instance, our case involved harvesting computer science related information from a wide variety of resources, ranging from university departments to funding organisations data. The crux of the problem with all that externally defined data, is in the inconsistent way of referring to an instance. In the database world, this has been identified in the past as the referential integrity problem, however, on the SW we coined the phrase "co-reference resolution" to reflect its SW-specific nature [2]. Ensuring that the instances we access and use do not have duplicates and co-references could be resolved without causing inference chains to collapse, is a problem we had to deal with at an early stage. It is an evaluation issue, but this time we are called upon to evaluate the data (or instances) themselves, not the ontologies that will instantiate. It is a much harder problem than we originally anticipated, mostly due to the sheer size of data sets we had to deal with (millions of instances), and the diversity of resources we harvested them from. Our solution has been a number of semi-automatic scripts that use adaptable heuristics in order to spot similarities, or otherwise, among seemingly duplicate instances [2]. As before, this issue should also be considered early especially if the re-use case involves instances that will be outsourced.

3.3 Trust and validity check

In the ontology mapping domain, our experiences with evaluation provided us with a difference challenge: not only the ontologies we needed to map were externally defined, but the mapping exercise was an ad-hoc one. In contrast, the problems we encountered with evaluating external ontologies in the CAS case, were easier to tackle as we new where and how those ontologies will be used. That helped us to devise a tailored evaluation strategy where changes were easier to make. In the mapping case though, the goal was to map ontologies. We did not have a clear idea of who and how will use the produced mappings. Hence, evaluating the externally defined ontologies was a problem as we had to treat them as a "black box". We briefly analyse some specifics with regard to an instantiation of the mapping case.

A specific case: FMA vs. OpenGALEN: As part of the annual ontology contest event[10] we had to map two large medical ontologies, the FMA (Foundational Model of Anatomy) and OpenGALEN (the open source version of GALEN ontology). FMA ontology describes the domain of human anatomy and it aims to provide "a reference ontology in biomedical informatics for correlating different views of anatomy, aligning existing and emerging ontologies in bioinformatics" [24]. However, there are two notable facts regarding the syntactic and modelling idioms of FMA and existing results from previous efforts in trying to align FMA and GALEN. As far as the former is concerned, the OWL version we had to work with was a result of translation from Protege. Previous work has shown that this result is not always a faithful representation of the original FMA Protege model. For instance, it has been reported that FMA DL constructs are often ill-defined and they lead to inconsistencies when a reasoner parses the ontology [11]. Consistency checking for FMA is an acknowledged problem though, even by its authors: "[...] feedback from these investigators revealed an aggregate of a few hundred errors, many of which related to spelling and only a few to cycles in the class subsumption and partonomy hierarchies." [24].

Leaving aside this fact of life (as it is natural for an ontology that big and so close to human practice to be inconsistent), we point to a couple of syntactic idioms that we found interesting when parsing the ontology with our Jena-based CMS system. Firstly, the rather unusual use of unique frame IDs for class names (`<owl:Class rdf:ID>` constructs) and the textual description of a class in an `rdfs:label` construct. We also noticed some unusual uses of references to

[8] http://challenge.semanticweb.org/

[9] Accessible online from www.aktors.org/ontology

[10] http://oaei.ontologymatching.org/

frame IDs. For instance, the declaration of "arterial supply" as an object property: `<owl:ObjectProperty rdf:ID= "arterial_supply" rdfs:label="arterial supply">` is used in other parts of the ontology where it refers to a `rdf:resource` which points to a different resource:
`<arterial_supply rdf:resource= "#frame_14586"/>`. Tracing that frame ID leads us to a definition of a "Tissue" class, and not the "arterial supply": `<owl:Class rdf:ID="frame_14586" rdfs:label="Tissue">`. The definition of an instance (with frame ID 14586) of an object property ("arterial supply") that is a class ("Tissue") could lead to modelling misunderstandings and confusion (although, syntactically speaking, it is allowed in some versions of OWL).

Going back to our argument for the notable facts, we found that previous efforts for aligning FMA to GALEN reported rather controversial results. For example, in [30], the authors employed two different alignment methods to map FMA to GALEN. Some of their findings are questionable from the semantics point of view: for example, it was reported that "Pancreas" in FMA matches "Pancreas" in OpenGALEN with 1.0 similarity value which "indicates a perfect match" [30]. When we looked carefully at the definitions of "Pancreas" in both ontologies we saw that "Pancreas" is defined as a class in FMA (`<owl:Class rdf:ID="frame_12280" rdfs:label="Pancreas">`)
whereas in GALEN (OpenGALEN) as an instance of class "Body Cavity Anatomy"
`<owl:Class rdf:ID="Body_Cavity_Anatomy">`
`<rdfs:subClassOf`
`rdf:resource="#OpenGALEN_Anatomy_Metaclass"/>`
`<Body_Cavity_Anatomy rdf:ID="Pancreas">`
Even if OWL semantics allow to map an individual to a class (when dealing with OWL Full), such an alignment is misleading especially when we consider the high level of abstraction for the "Pancreas" class in OpenGALEN. It seems that the "lexical phase" parsing used in [30] was the main contributor to this high similarity value when relatively little structure information was taken into account. As a final comment on the case, we also point the reader to observations made by the FMA authors when trying to validate mapping results and differences in terminologies with these two ontologies: "[...]the reasons for the differences have not yet been explored, but at least some of them may be the different contexts of modelling. GALEN represents anatomy in the context of surgical procedures, whereas FMA has a strictly structural orientation." [24].

Our experiences with the mapping case between FMA and OpenGALEN might be specific to these two ontologies but the observations we drawn are generic and highlight a problem with evaluating externally defined ontologies: how to trust these ontologies and how to check their validity. From what we learned from the FMA vs. OpenGALEN case, specialist knowledge is needed to verify not only the correctness of mapping results - in our case - but most importantly the original ontologies. This is not an easy task though as the report in [24] suggests where there was a clear disagreement over the ontology among domain experts. We have also seen reports in the literature that emphasize the difficulties of evaluating medical ontologies [6], mostly due to the specialists terminology but also because of deep disagreements among medic professionals.

In the next section we discuss possible ways of alleviating the problem by using: certified ontologies, a community oriented approval and maintenance scheme, and techniques for ranking and cataloguing ontologies.

4. WAYS TO ALLEVIATE THE PROBLEM

One of the trendy solutions to old and current problems, especially in the context of the SW, is the use of community input to engineering tasks. This is also heralded as *emergent semantics* in the literature. We reflect on the positive and negative aspects of engaging user communities in the ontology evaluation tasks.

4.1 Engaging user communities

The concept of *emergent semantics* is seen as both a challenge and an opportunity for SW engineering. It is an opportunity if it is manifested properly in order to regulate a community's vocabulary and help it evolve to a stable version. However, *emergent semantics* communities work on the principle of self-organision and it is more likely to reach local consensus first, before achieving the desired inter-community interoperability. This is a necessity before any community input can be re-used in similar cases.

Examples of communities of that sort are, for example, mainly Web users who begun using social software such as FOAF[11], Flickr[12] or del.icio.us[13] few years ago for leisure and as a socialising medium. This kind of software and social networks have seen an unprecedented growth recently with at least a couple of dozens of tools like that. The interesting development that makes this new area appealing for engineers[14], is that it could be used to alleviate some of the tough problems engineers face today with the Web, and the SW. The premise of the argument is that as communities interact and formed up, their members are contributing, knowingly or unknowingly, to communal knowledge assets. Although this old adage has been acknowledged in the past in neighbouring domains, like communities of practice [29], in today's online social networks on the Web has a different form.

It is possible to use current technology, like machine learning and text engineering, to extract the meaning (semantics) underlying information exchanged in these networks between their members, and then collectively represent it as a community's underlying semantic model. This way of uncovering the semantics used by a community's members is often described as *emergent semantics*. The difference with traditional knowledge engineering is that we no longer have a pre-defined semantic model (in the form of an ontology) of a domain to which a community adheres to, but rather, we have an emergent model that surfaces while a community functions and its members interact.

This promise has been put to work, to a certain extent, in certain scenarios and communities. For example, [31] presents an approach where input from the community could be used to validate ontology mapping results, whereas in [7] a range of user driven applications are discussed. But, to the best of our knowledge, it has not been applied yet to more complex and subtle domains and tasks where a stable

[11]www.foaf-project.org
[12]www.flickr.com
[13]http://del.icio.us/
[14]See, for example, the first ever dedicated event to social software: http://sw.deri.ie/ jbreslin/foaf-galway/

and reliable semantics are a necessity. However, early steps towards this directions are outlined and examined in [21].

We also identify some weak areas of *emergent semantics* that could be potential show-stoppers, if not considered properly. *Emergent semantics* promise relies on a relatively smooth and uniform representation of a community's interest. But, real world practice in similar domains tells us that communities use a variety of norms and manifestations [3]. If *emergent semantics* are to be extracted from a community's log, then we'd better have a stable log with a standard vocabulary to work with. The more variations we encounter in the semantics used, the more difficult it will be to extract them in a distributed environment like the SW.

Another area where little work has been done is to resolve possible conflicts on representation of common knowledge used by these communities. It is not uncommon for similar communities to use seemingly similar representations of the same domain concepts. Even if the sheer number of members is a strong indicator that the concepts used are prevailing and should be part of a prospective ontology, the fact that there could be slight variations in representing these concepts means that we might have to do some sort of mapping and alignment first before using them. We might even have to identify potentially conflicting concepts and we do not currently have the technology to do conflict resolution in an automatic manner [2].

Emergent semantics are also heavily engineered using machine learning technology and other statistical techniques, like OLAP. We do not despise the use of these technologies, but we are sceptical about the practicability of deploying machine learning to capture and extract prevailing semantics from a community's log of information exchanges in an environment like the Web. Supervised learning methods are probably the most reliable ones, yet, they are the ones that need a lot of input from an engineer to ensure that the right learning data set is fetched into the system, update the learning strategy, maintain the learning rules, etc. All these are time consuming tasks and require specialised knowledge of not only machine learning, but also of the domain at question. We do not see how this could scale up at the level of the Web or SW, when we will have to deal with thousands of outsourced domain concept definitions where only a human expert in that domain will be the most qualified person to interpret the concept properly and attach the correct semantics to it.

Therefore, many practitioners turned to unsupervised learning methods, that require the least possible input. However, they also have their limitations as the domains to which they have been applied with considerable success, are well understood and a universal set of initial training data and rules can be defined. But in domains where a wide variety of concepts could be learned and extracted, this is not the case.

Despite the advantages and disadvantages of using *emergent semantics* we mentioned above, we are cautiously optimistic about them. We believe that *emergent semantics* will continue to grow as more and more people will be drawn into these online communities using social software. We would like, therefore, to make the most out of their interactions and information exchanges on the Web and SW by capturing the semantics underlying their actions with respect to the pertaining problem we discuss in this paper: evaluation.

Communities alone though, will not be able to provide us with practical input with respect to evaluation unless we have ways of regulating and vetting their input. One way that this could be mechanised is with the use of certification.

4.2 Towards certified ontologies

In the knowledge engineering domain, the issue of certification has been debated in the past [26] in the context of certified knowledge bases. Recently, the issue of using ontologies as a commodity, and the commercial interest it has attracted has also been debated [22]. We have also witness efforts that aim to certify and validate domain specific ontologies, like the work of [8] with medical ontologies.

All these representative pieces of work emerge from different contexts and application domains but point to a workable approach: evaluation could be done by professionals and adhere to standards and practices approved by recognized bodies of prominence. We should also point to efforts that already exist in the commercial world, especially those that apply to the Web. For example, the commercial importance of the Web and the volume of trading online brought us technologies like SSL certificates for encrypting financially sensitive information and certification mechanisms like VeriSign's "verified by" trademarked certificates.

Similarly, at the syntactical level, some of the W3C family of languages, and other products related with the consortium's efforts, have clearly identifiable stickers on compatible web pages ("XHTML checkers", etc.) pointing to syntax validators and checkers or simply stating conformance to a standard.

Despite these activities though, the certification of ontologies, especially with respect to evaluation remains an issue largely unresolved and ignored by big standardisation bodies. We might have witness high profile efforts in ontology development, like the commercialisation of CyC [19] or the IEEE sponsored work on SUO[15] but this does not mean that we have evaluation bodies that provide certificates of ontology quality assurance.

The problem with issuing certificates of ontology quality is two fold: on the technological level, we do not have a clear idea of what quality criteria and tests ontologies should satisfy in order to be accredited. On the political level, there is an issue of authority. Who will certify ontologies and how? How trustworthy will that organisation be and what, if any, will be the costs of certification. Will there be licensing issues and restrictions of use with respect to the ontology? How likely it is to reach at a standardisation level when talking about ontology evaluation?

Experience and industry reports on standardisation tells us that standards are hard to debate, difficult to enforce in an open-ended environment, hard to reconcile conflicting commercial interests, and take years to materialize. But, for ontology evaluation efforts to have more credible profile some sort of standardisation would be needed. One way of combining the strengths of emergent semantics we reviewed before and ideas from commercial efforts on certification and standards could be to use simple cataloguing technologies, like ranking.

4.3 Classification and Ranking

in [1] the authors report on early efforts to come up with ranking mechanisms that allow us to classify ontologies according to their usage. Their domain of application is on

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

searching for appropriate ontologies but the ranking mechanism is simple and could be adopted to support evaluation. Assuming that a community is willing to participate in a common effort to rank ontologies, such an approach could provide us with a majority's view on what is best and what to avoid. This is the premise of the ranking approach.

We do however, have certain issues to resolve before making it practical for evaluating ontologies: (a) how to monitor and regulate rankings in an open-ended environment? reports that examined well crafted commercial efforts on using communal ranking (like for example the ebay feedback mechanism) has shown that it is easy to deceive authoritative systems in order to achieve personal gains [23] (in the case of ebay feedback, a positive one could mean better deals for auctioneers). (b) what sort of features in an ontology users will be called upon to evaluate? that issue is related to the certification content discussed above and we see efforts such as in [28] as an early step towards a consensual set of features that evaluated ontologies should demonstrate. Furthermore, in [20] a more detailed and extensive list of characteristics for ontology classification has been proposed. (c) will all participating users have equal opinion weighs? for example, in the case of the FMA ontology, should an anatomist's opinion have greater importance than an ontology engineer's? common sense might dictate that he should, but there might be subtle KR related issues that only the ontology engineer will be qualified to resolve.

5. CONCLUSIONS

Evaluation of ontologies themselves, as opposed to evaluation of ontology tools that was the theme of all previous EON workshops, is a difficult issue. It cannot be seen as orthogonal to other ontology development and use issues. Especially not in an environment like the Web, and the SW. The promise of accessing, retrieving, and re-using a variety of ontologies in these environment necessitates an evaluation strategy that is (a) open to users, transparent in nature and with references to the standards it adheres to or certificates it holds, (b) amendable, easy to change and adopt to different use cases, (c) domain specific, and reflect opinions of various stakeholders, not only of ontology engineers.

But, these are hard to achieve goals. In the short to medium term we should look for mid term solutions that we can build and experiment with, before engaging to long term evaluation research. In the last part of the paper, we elaborate on a rough roadmap of the short to medium future. Standards and certification is an area that needs more work. In fact, when it comes to ontology evaluation, it is in its infancy. However, we want to avoid the painfully slow process of standardisation. There are lessons learnt and experiences we can build upon. For example, in the context of the IEEE SUO effort, there have been debates on using ISO standards to evaluate the content and appropriateness of ontologies[16]. Despite the fact that views and opinions expressed there are subjective, it is a good start.

We also see an increasing interest in using *emergent semantics* and engaging user communities. That could prove to be a useful and practical input to the evaluation problem. The commercial interest in ontologies nowadays also brings us closer to certification and standards. As academic and neutral interest stakeholders we should inform possible attempts for certification as to what the quality features that ontologies need to exhibit should be and leave the prolong arguments on how to enforce them to the politicians. Licensing is also an issue that should be considered closely with evaluation. Appropriate licensing should provide certain assurances on evaluation.

The practical research questions on what sort of evaluation technology we need should be part of the ontology development and language communities. The SW community at the moment, focusses on applications and infrastructure issue. Having closed a successful cycle on developing languages to materialise the SW, researchers and practitioners are focussing on attracting commercial and public interest by demonstrating SW technology and its advances. But, evaluation of ontologies, a cornerstone for achieving the full potential of the SW, is not complete yet. We believe that this is wrong. We need to advance the current methods for evaluation, some of which have been demonstrated in the evaluation of ontology tools through meetings like EON, and extend them to include evaluation of ontologies themselves.

Last, but not least, in an era where user communities matters the most, we need to raise the awareness of this issue and demonstrate its importance. As researchers, we need to share experiences, good and bad, on related efforts and learn from each others mistakes. Open source and publicly available tools should be on the the agenda so that we can reach to a consensus quicker. We should not be afraid to constructively critique and despise ill-defined ontologies as this will raise the quality standards. Most importantly, we should work with examples, tools, and use cases that are easy to replicate in neutral settings.

[16]see message thread in: http://suo.ieee.org/email/msg12376.html

6. ACKNOWLEDGMENTS

This work is supported under the Capturing, Representing, and Operationalising Semantic Integration (CROSI) project which is sponsored by Hewlett Packard Laboratories at Bristol, UK. The first author is also supported by the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC) project which is sponsored by the UK EPSRC under Grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of the EPSRC or any other member of the AKT IRC.

7. REFERENCES

[1] H. Alani and C. Brewster. Ontology Ranking Based on Analysis of Concept Structures. In *Proceedings of the 3rd International Confernece on Knowlegde Capture (K-Cap'05), Banff, Canada*, pages 51–58, Oct. 2005.

[2] H. Alani, S. Dasmahapatra, N. Gibbins, H. Glasser, S. Harris, Y. Kalfoglou, K. O'Hara, and N. Shadbolt. Managing reference: Ensuring Referential Integrity of Ontologies for the Semantic Web. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), Siguenza, Spain*, pages 317–334, Oct. 2002.

[3] H. Alani, K. O'Hara, and N. Shadbolt. ONTOCOPI: Methods and Tools for Identifying Communities of Practice. In *Proceedings of the 2002 IFIP World Computer Congress, Montreal, Canada*, Aug. 2002.

[4] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May 2001.

[5] P. Borst, H. Akkermans, and J. Top. Engineering Ontologies. In *Proceedings of the 10th Knowledge Acquisition for Knowledge Based Systems Workshop,Banff, Canada*, 1996.

[6] W. Ceusters, B. Smith, A. Kumar, and C. Dhaen. Mistakes in Medical Ontologies: Where do they come from and how can they be discovered? In *Proceedings of Workshop on Ontologies in Medicine, Rome, Italy*, oct 2003.

[7] M. Dzbor, H. Takeda, and M. Vargas-Vera, editors. *End User Aspects of the Semantic Web (UserSWEB'05)*, volume 137 of *WS*. CEUR, may 2005.

[8] G. Eysenbach. An Ontology of Quality Initiatives and a Model for Decentralized, Collaborative Quality Management on the (Semantic) World Wide Web. *Journal of Medical Internet Research*, 3(3):e34, 2001.

[9] A. Farquhar, R. Fikes, and J. Rice. The Ontolingua Server: a Tool for Collaborative Ontology Construction. *International Journal of Human-Computer Studies*, 46(6):707–728, June 1997.

[10] M. Fernandez, A. Gomez-Perez, and N. Juristo. METHONTOLOGY: From Ontological Arts Towards Ontological Engineering. In *Proceedings of the AAAI-97 Spring Symposium Series on Ontological Engineering, Stanford, CA, USA*, pages 33–40, Mar. 1997.

[11] C. Golbreich, S. Zhang, and O. Bodenreider. Migrating the FMA from Protege to OWL. Technical report, jul 2005. In notes of the 8th International Protege Conference, Madrid, Spain.

[12] A. Gomez-Perez. Towards a Framework to Verify Knowledge Sharing Technology. *Expert System with Applications*, 11(4):519–529, 1996.

[13] S. Harris and N. Gibbins. 3store: Efficient Bulk RDF Storage. In *Proceedings of the ISWC'03 Practicle and Scalable Semantic Systems (PSSS-1), Sanibel Island, FL, USA*, Oct. 2003.

[14] J. Hartman, P. Spyns, A. Giboin, D. Maynard, R. Guel, M.-C. Suarez-Figuera, and Y. Sure. Methods for Ontology Evaluation. NoE Deliverable D1.2.3, KnowledgeWeb EU NoE - FP6 507482, jan 2005.

[15] Y. Kalfoglou, H. Alani, M. Schorlemmer, and C. Walton. On the Emergent Semantic Web and Overlooked Issues. In *Proceedings of the 3rd International Semantic Web Confernece (ISWC'04), LNCS 3298, Hiroshima, Japan*, pages 576–591, Nov. 2004.

[16] Y. Kalfoglou and B. Hu. CMS: CROSI Mapping System - Results of the 2005 Ontology Alignment Contest. In *Proceedings of the K-Cap'05 Integrating Ontologies workshop, Banff, Canada*, pages 77–84, Oct. 2005.

[17] Y. Kalfoglou and M. Schorlemmer. IF-Map: an Ontology Mapping Method based on Information Flow Theory. *Journal on Data Semantics*, 1:98–127, Oct. 2003. LNCS2800, Springer, ISBN: 3-540-20407-5.

[18] Y. Kalfoglou, M. Schorlemmer, M. Uschold, A. Sheth, and S. Staab. Semantic Interoperability and Integration. Seminar 04391 - executive summary, Schloss Dagstuhl - International Conference and Research Centre, Sept. 2004.

[19] D. Lenat. Cyc: A Large Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11), 1995.

[20] A. Lozano-Tello and A. Gomez-Perez. ONTOMETRIC: A Method to Choose the Appropriate Ontology. *Journal of Database Management*, 15(2):1–18, 2004.

[21] N. Noy, R. Guha, and M. Musen. User Ratings of Ontologies: Who will Rate the Raters? In *Proceedings of the AAAI 2005 Spring Symposium on Knowledge Collection from Volunteer Contributors, Stanford, CA, USA*, 2005.

[22] K. O'Hara and N. Shadbolt. Issues for an Ontology for Knowledge Valuation. In *Proceedings of the IJCAI'01 workshop on E-Business and the Intelligent Web, Seattle, WA, USA*, Aug. 2001.

[23] P. Resnick and R. Zeckhauser. Trust Among Strangers in Internet Transactions: Empirical Analysis of ebay's Reputation System. *Advances in Applied Mircroelectronics*, (11), 2002.

[24] C. Rosse and J. Mejino. A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36:478–500, 2003.

[25] N. Shadbolt, N. Gibbins, H. Glaser, S. Harris, and M. Schraefel. CS AKTive Space or How we learned to stop worrying and love the Semantic Web. *IEEE Intelligent Systems*, 19(3):41–47, May 2004.

[26] N. Shadbolt, K. O'Hara, and L. Crow. The Experimental Evaluation of Knowledge Acquisition Techniques and Methods: History, Problems, and new Directions. *International Journal of Human-Computer Studies*, 51:729–755, 1999.

[27] M. Uschold, M. Healy, K. Williamson, P. Clark, and S. Woods. Ontology Reuse and Application. In N. Guarino, editor, *Proceedings of the 1st International Conference on Formal Ontology in Information Systems(FOIS'98), Trento, Italy*, pages 179–192. IOS Press, June 1998.

[28] M. Uschold and R. Jasper. A Framework for Understanding and Classifying Ontology Applications. In *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods(KRR5), Stockholm, Sweden*, Aug. 1999. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-18/.

[29] E. Wenger. Communities of Practice: The Key to Knowledge Strategy. In E. Lesser, M. Fontaine, and J. Slusher, editors, *Knowledge and Communities*, pages 3–20. Butterworth-Heinemann, 2000. first published 1999.

[30] S. Zhang, P. Mork, and O. Bodenreider. Lessons Learned from Aligning two Representations of Anatomy. In *in Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation, Whistler, BC, Canada*, pages 102–108, 2004.

[31] A. Zhdanova. Towards Community-Driven Ontology Matching. In *Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP'05), Banff, Canada*, pages 221–222, oct 2005.