Promotion of Ontological Comprehension: Exposing Terms and Metadata with Web 2.0

Andrew Gibson Katy Wolstencroft Robert Stevens

University of Manchester
School of Computer Science, Kilburn Building,
Oxford Road, Manchester, UK

+44 161 275 0649

andrew.p.gibson@manchester.ac.uk

ABSTRACT

Knowledge artifacts that have been labeled as ontologies have many different qualities and intended outcomes. This is particularly true of bio-ontologies where high demand has led to a rapid growth in the number of these artifacts. Good communication between the human agents involved in the life cycle of ontologies is essential for the ontologist to encode the right knowledge in the ontology. Not only this, but it should be encoded such that subsequent retrieval of the knowledge from the ontology by any agent can be clear and precise. The ontologist can encode ontological statements, for interpretation by a computer agent, or meta-ontological statements, for interpretation by human agents. We consider how the current communication between agents and ontologies produces drawbacks that add to the considerable overheads associated with ontology development. We describe the processes of communication between human agents and ontologies as Ontology Comprehension. We then suggest how these processes could be augmented, particularly with the use of Web 2.0 ideas. By exposing and enhancing the social interactions involved in ontology comprehension, development overheads are potentially reduced and the prospect of ontology sharing and reuse is improved.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *representations,* r*epresentation* languages.

General Terms

Design, Human Factors, Standardization, Languages.

Keywords

Ontologies, Semantic Web, Web 2.0, OWL, Ontology Comprehension.

1. INTRODUCTION

The technologies of the Semantic Web [6] have been centrally conceived, specified and designed with recommendations by the W3C[1]. This next generation Web promises to transform the information Web into a machine computable utopia for semantically described data and information. Despite the development of the technologies, there is, however, only little evidence of the materialization of the Semantic Web (or Webs). Simple RDFS vocabularies such as Friend of a Friend have provided small views on the potential of the Semantic Web [9]. Rich ontological views supported by reasoning have appeared in applications [27, 30, 31], but less so in the Web itself, and when they do, they often represent unconnected niche pockets of interest.

In contrast, Web 2.0 is in the here and now, in use by large interconnected user communities, and is ever growing as more people adopt and contribute to various community efforts. To try and specify Web 2.0 would almost be a contradiction in terms, and restricting its users with strong recommendations would be seen as an attempt to unnecessarily limit the creativity of those who have something new to try. Taxonomies give way to folksonomies, letting the user mark-up things lightly on the Web rather than specify a typed URI. The technologies of Web 2.0 were not specified; they evolved out of clear and present needs of users to connect with one another. The principles of Web 2.0 grow out of a mixture of hindsight and insight to current practice, and revolve around online community building, quick and easy linking, unlimited customization in the hands of the masses. In this article we use 'Web 2.0' to refer to these principles rather than any specific technology.

It has not gone unnoticed however that the artifacts, such as vocabularies and ontologies, that will support the Semantic Web need populating [25, 26], and for this to happen, both the technology and the nature of ontology building need to be accessible to the masses. Similarly in the computer science view of knowledge artifacts such as ontologies inherently have this community aspect—they are shared conceptualizations that aim to enable both human and computational interoperation of diverse resources at a semantic level.

The simplicity and robustness of HTML fuelled the growth of the current Web, but the highly-specified nature of the technologies in the Semantic Web recommendations suggests that the semantic side of the development, delivered through ontologies, will be driven mostly by experts. In this way, it is key that somehow this barrier of complexity is lowered through creating an easier user

Copyright is held by the author/owner(s).
WWW 2007, May 8--12, 2007, Banff, Canada.

[1] http://www.w3.org/2001/sw/

experience, and that the motivators that are driving Web 2.0 are harnessed to promote uptake of Semantic Web ideas.

In this paper we consider the social and communication dependent aspects of the ontology development life cycle, and identify problems encountered by people with specific roles of interaction. From this, we suggest that a clear, layered separation is made between statements in ontologies that are logical and those that are linguistic, supporting annotations on the ontology. In doing so, the annotations can be exposed to the collaborative aspects of Web 2.0, promoting light discussion at the level of natural language about the meanings of terms, whilst leaving the heavier encoding of knowledge into OWL as a task for ontologists.

2. ONTOLOGIES AND DEVELOPMENT

The central premise of the Semantic Web is enabling computational processing of Web resources through knowledge artifacts. The W3C have provided the Resource Description Framework (RDF) and the Web Ontology Language (OWL) recommendations. The latter, particularly in its OWL-DL variant, is offered as a means of building robust property based descriptions with a logical underpinning that can be used to provide vocabulary for describing Web content, but also support reasoning across Web content [20]. Such ontologies are to be the semantic backbone for linking resources in the Semantic Web. Additionally, these ontologies are to represent knowledge of domains, and have the virtues of being sharable and reusable. As yet, it is difficult to find an ontology that could be said to have been designed to fit the criteria for enabling a Semantic Web by being domain general and rich in content. One prominent example of an ontology approaching these criteria is the Foundational Model of Anatomy (FMA) [12, 23]. The FMA could be said to be more of a true domain ontology (or reference ontology) than any other in bio-medicine. However, even the FMA has barriers to the Semantic Web goals of sharing and reuse because of its large size, perhaps because it was developed in Frames and later converted to OWL.

In computer science, what are called ontologies covers a broad range of knowledge artifacts. Glossaries, vocabularies, thesauri, informal and formal ontologies (both in language and ontological discrimination) are all used at various points in the Semantic Web. Different levels of expressiveness (sometimes called formality) come from the purpose and demands of the ontology being developed [28]. These demands can be considered with increasing levels of expressiveness from very "light-weight" term lists, thesauri, dictionaries or hierarchies up to "heavy-weight" with very expressive constraints [10, 25]. OWL-DL offers a formal language and can be used to build rich, logical representations of descriptions of what exists; it can also be used, in various forms, to develop other forms of knowledge artifact while still retaining strict language semantics in the representation, but weakening the ontological distinctions made in

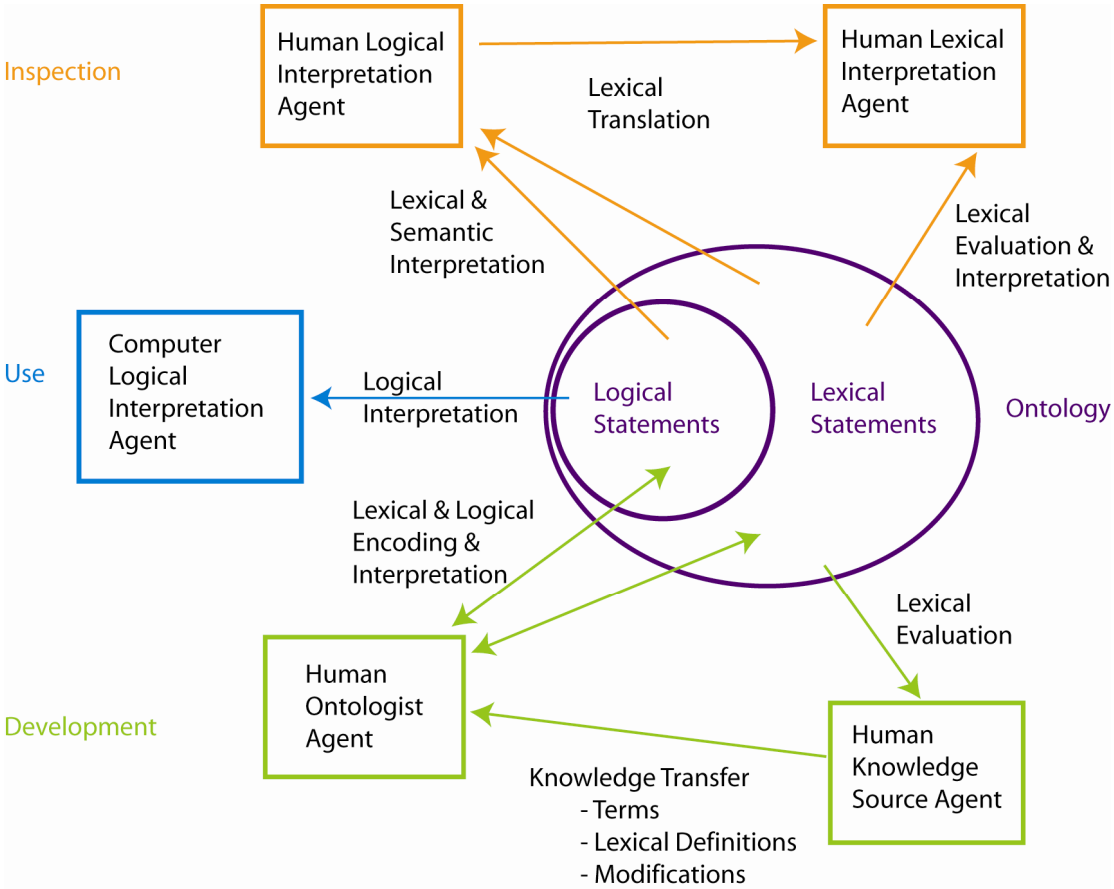

Figure 1: Ontology Comprehension: Current model of interactions between various agents and an ontology, as described in Section 3. The human agents are not necessarily different individuals, but rather are separated here by the roles fulfilled in the development and inspection processes.

the knowledge artifact.

Building OWL-DL logic based ontologies is a difficult process [21] and reaching a community consensus is hard, especially in complex domains such as biology, where knowledge for making ontological distinctions can be incomplete. These issues need to be addressed if ontologies are to play their role in the Semantic Web. Here, we are mostly interested in the aspect of reaching a community consensus. Focus is often placed on the aspect of collaborative ontology building, that is, a group of people working directly with one ontology. We do not aim to discuss this type of system, as we see such systems as expert systems for logic-savvy ontologists rather than currently being suitable for "the masses". Much more work needs to be done on enabling true collaboration in logic based ontologies. Instead, we currently envisage a core of expertise for logic encoding supported by people conceptualizing and gathering linguistic material. We acknowledge that there is a wealth of methodologies that address certain aspects of the ontology development lifecycle [10, 29] and evaluation [8, 24], good reviews of these fields can be found in the references. For the purposes of this paper, we wish to focus on the social interactions during these processes rather than the processes themselves.

3. ONTOLOGY COMPREHENSION

We learn from the field of software engineering that effective reuse of elements of object oriented frameworks is reliant on many levels of understanding from the point of view of the programmer [4, 15]. In software engineering, improving these levels of understanding is known as "software comprehension", and we extend the principles to ontology development. We outline *ontology comprehension* as the interaction between human agents and the knowledge expressed in an ontology.

Figure 1 outlines the interactions between various agents and an ontology that are considered in this section. There are two main modes in which ontology comprehension is important:

1. Development mode. Ontology development requires that there is efficient interaction between experts that represent the knowledge of the domain in the scope of the ontology (domain experts) and the ontologist that is responsible for the construction and continued maintenance of the ontology. Here we assume a model where, for a specific ontology development exercise, there is a limited cohort of domain experts that are involved with an ontologist.
2. Inspection mode. Ontology inspection is a light evaluative process that an agent will go through the ontology to quickly assess whether or not that ontology is of good quality and whether what it contains is suitable for some specific needs of the inspector.

What follows is an outline of task models that highlight how currently, the interactions of agents involved with ontologies leads to discrepancies in ontology comprehension.

3.1 Task Model 1: Ontology Development

We consider early ontology development as a process that begins with the lightest possible knowledge structure, essentially a term list, and subsequently moves up through levels of complexity and expressiveness of the types discussed in [10]. This happens socially as well as in the ontology as all those involved in the development become more familiar with scope. At the beginning of the ontology development life cycle, the ontologist (assuming they have no domain knowledge) will usually rely on the domain expert to provide a core set of terms from the domain of interest as a starting point. The initial scope of the ontology, rather than being rigidly defined, is often roughly determined from the initial term list and this will get refined as things move on. At this early stage it is necessary for the domain expert to be able to quickly assess if the terms are appropriate. As things are, the easiest way to do this is for the domain expert to be able to access the ontology for themselves and browse the hierarchy of terms, whilst checking and adding in textual annotations for the terms, as well as any comments about the specific or contextual use of any of the terms.

The ontologist will be using one of the commonly available ontology development tools such as Protégé-OWL[2], Swoop[3] and OBO-Edit[4]. All of these tools are centered on the user interacting with a class hierarchy view, which the ontologist will be building from the terms given to them by the domain expert. At this stage, the domain expert will primarily be concerned with having the correct term-definition pairs represented in the proto-ontology. Decisions regarding the class hierarchy signal the beginning of a slightly more complex level of expressivity, as the ontologist will be making assertions between classes about subsumption relationships [14]. This is especially true of OWL ontologies, and such decisions do not necessarily need to be considered for simpler controlled structured vocabularies in which hierarchical relationships "broader than" and "narrower than" are possible. The ontologist may also start to guide the domain expert in how to transfer knowledge regarding some of the more fundamental object properties such as part-hood.

At some point, the domain experts need to let the ontologists start to make even more expressive assertions in the ontology that they may not necessarily understand the implications of for themselves. This signals the next stage of ontology development, in which the balance shifts so that the ontologist starts to refine the assertions in the ontology. Instead of being instructed and guided by the domain expert, the ontologist now needs to ask careful questions of the domain expert. The aim of these questions should be to extract the intrinsic meaning of the terms that the domain expert has provided so that the ontologist can encode these meanings into the ontology using more and more expressive restrictions and axioms. Significantly, unless the domain expert has had training in understanding the meanings of logical assertions of ontologies, they will still primarily rely on the lexical annotations and definitions when evaluating the ontology.

Once the content of the ontology has begun to stabilize (i.e. there are fewer major revisions in the content of the ontology being made) it will be made available to a wider audience. This can signal a whole new critical process of revision for the ontology. In the next section we will consider what sort of interactions may occur between different agents and ontologies when they are first encountered.

Eventually, the increase in the content of the ontology, both lexical and logical, should start to level off as the content and the intended scope, at which time further structural modifications may be made, such as modularization, which could happen once

[2] http://protege.stanford.edu/

[3] http://code.google.com/p/swoop/

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

the micro-organization of the knowledge in a domain has become clear. A publicly available and relatively stable ontology has a new set of requirements, for which the topics of ontology evolution and change management address [19]. Change management of ontologies has been considered in a technological sense for some time, and it should be clear that changes to a publicly available ontology need to be transparent. However, there is a growing trend for including extra hierarchical structures into the ontology that represent deprecated classes (e.g. [30]). The need to do this is obvious; it is less so how to do it neatly and ontologically. Versioning etc. are all parts of the ontology life-cycle that have no really, consistent support.

The following discrepancies in ontology comprehension should be clear from this section.

1. Discrepancies in Early Development
    a. The most convenient means of constructing, looking at and sharing the early term list is, unusually, from within an ontology file, which implies some hierarchical structure.
    b. Early revisions of the ontology are experimental for the ontologist, yet are still subject to inspection and lexical evaluation from the domain expert.
    c. Domain experts, having looked directly at revisions of the ontology file, may be resistant to subsequent major changes in structure and terminology by the ontologist as knowledge is disambiguated.
2. Emerging Discrepancies
    a. Inclusion of information regarding deprecated classes into the class hierarchy of the ontology.
3. Communicative Discrepancies
    a. Discussions between the domain experts about terminology that are potentially crucial for ontology comprehension are lost or are completely separated from the ontology itself.
    b. Discussions between the domain experts and the ontologists about disambiguation of terms are lost or are completely separated from the ontology itself.
    c. Potential for misinterpretation of logical aspects of the ontology by the domain experts through exposure to the logical component.

3.2 Task Model 2: Ontology Inspection

Ontologies are complex entities. If any ontology is going to get used by someone other than the person or group that implemented it, there has to be a way in which it can be decided whether or not it is an appropriate ontology for the task in hand [18]. Currently, this inspection process is difficult because of the paucity of ontologies available, and the fact that many have been designed for a specific purpose. Also, the discrepancies listed in 3.1 result in a general lack of information that can aid effective inspection and overall ontology comprehension.

It is hard not to liken an ontology inspection process to some sort of evaluation. What we describe here is fairly close to *ontology selection* [24], except that ontology inspection is more of a browsing process, driven by what access there is to comparative information between several ontologies. Selection has much better defined initial parameters for the desired outcome, and can give a more targeted outcome. We do not wish to label this inspection as an evaluation however, as we do not make the assumption that the inspector will be following any pre-determined criteria, and if they are, that they are rational criteria.

The ontology inspection process is short lived, and for many people's goals, the choice of beginning a new ontology that they know will satisfy their criteria is more favorable than editing an existing one. However, such inspections can quickly be deemed fruitless when the term searched for turns out not to be defined by logical statements in an ontology. This is a common occurrence, as such 'classes' can be placeholders for future development or intrinsically defined terms where no logical definition was thought necessary. Ontologies can be intensely developed in one particular area where immediate goals are important, yet there is no way to effectively discover this other than through thorough browsing. For the goals of the Semantic Web, it is imperative that such information required to carry out this inspection process be made as clear as possible for the inspector, such that we do not see immense reproduction of individual effort and no clear "shared conceptualizations".

The domain knowledge these ontologies describe can require a considerable amount of understanding for anyone trying to inspect them. There are several ways in which this can be the case.

1. The domain knowledge encoded may be outside the experience of the inspector, or in a different context to what was expected. The inspector may not be able to tell if the knowledge represented is valid because it is not within their expertise, and will need to seek help and advice from a domain expert.
2. The knowledge may be appropriate, but encoded with axioms and restrictions that the inspector may not be able to accurately interpret as real world meaning, such that they have to find the advice of an ontologist.
3. The ontology may have been written for a specific purpose. The inspector may not be able to tell whether this is the case, and could therefore assume that the first or second scenario above is true, unless it is possible to seek advice from the original authors or find a resource containing this information.

The three scenarios above are serious issues for the future of ontologies in the Semantic Web. Most ontologies are developed as part of projects, and projects are usually pragmatic in terms of their goals. Hence, people build these ontologies as application ontologies that serve the immediate needs of the project. There is no perceivable immediate benefit for a project to develop a more general domain ontology in tandem with an application ontology, and so it does not happen. Consequently the Semantic Web goals of sharing and reuse become much harder, as people will tend to assess these application specific ontologies as too specific for a new purpose, as they see that they will need to invest effort in its re-engineering. Another danger here is that with so many application ontologies being developed, that inspectors always start to assume that unusual features of ontologies are the result of the needs of an application, and dismiss the ontology as

potentially unusable. What is really needed is for the inspector to be sure what sort of artifact they are looking at by having easy access to certain parameters.

In the Semantic Web vision, the first course of action for an ontologist would be to verify the existence or non-existence of a domain ontology with close or overlapping scope to the ontology they are to develop. This process will be laborious if it relies on the current practice of downloading ontologies and browsing them to see if they are at all reusable. In response to this, technologies such as Swoogle [16] and AKTiveRank [2] are starting to provide access to online ontologies through page ranking and other analytical methods to establish potential target ontologies. However, these technologies have been criticized for ignoring the meaning of concepts and also relations [24]. Furthermore, we note that the results returning from these searches are whole OWL files, free and independent of contextual information. For example, a Swoogle search for "Protein" has in its top hits an ontology used in an educational tutorial (that in this case is evident from its URL), which is by no means intended to be a shared or reusable resource, but none the less is discovered and accessible.

Those inspecting ontologies can find themselves in an isolated situation where Web searches and personal inspection of an ontology or its documentation are the only means to ontology comprehension. It has already been recognized that the Web has enormous potential for social organization and engagement. In ontology comprehension, for example, it offers the means of asking those who know. It also, as Wikipedia has shown, offers the means by which elements that aid ontology comprehension can be developed. Having concluded the need for ontology comprehension, we now explore what is necessary for such a facility.

The following discrepancies in ontology comprehension should be clear from this section:

1. Discovery Level Discrepancies
    a. Targeted discovery based on search for terms rather than meanings
    b. Ontologies are discoverable independently of statements of purpose, scope etc.
    c. Searches may discover anything from tutorial OWL files, programmatic OWL fragments, application ontologies, outdated or unmaintained ontologies etc.
2. Ontology Level Discrepancies
    a. Statements of scope, purpose, expressivity etc are often missing altogether, or require extra searches to discover them.
    b. Discussions that have affected overall ontology development are not recorded
    c. Minimal opportunities to interact with the development team
3. Term Level Discrepancies
    a. Feeding in from Section 3.1, ontologies need exploring in the development environment to assess appropriateness of terms.
    b. No indication without exploration of the level of effort put into different areas of an ontology.

4. DESIDERATA FOR SEMANTIC ONTOLOGY COMPREHENSION

Section 3 highlights the social and communicative discrepancies that prevent an effective amount of ontology comprehension that is required for the uptake of the Semantic Web goals of ontology sharing and reuse. This section cross-analyses these discrepancies to produce some desiderata that can be considered for future systems. Whilst all types of data in and about an ontology may be considered 'ontological', we specify 'Meta-Ontological Data', 'Ontological Metadata' and 'Logical Statements' as clearly identifiable parts. For information contained within ontology files that is only for human interpretation of the encoded semantic content, we use the idea of Meta-Ontological data. For data specific to an individual ontology that is necessary for interpret and inspection across the whole structure and history of development, we use the idea of Ontology Metadata. The 'logical statements' in an ontology constitute the remainder of the content.

4.1 Separating the Ontological from the Meta-Ontological

Ontologies come with a considerable amount of meta-ontological information (or should do so) which is used by the human to assess and see the intended use of that ontology. Much of this meta-ontological information is linguistically orientated. These meta-ontological extensions to the ontology itself are meaningless strings to the computer, and in this respect are unnecessary in so far as the computational goals of the Semantic Web are concerned. We know that this meta-ontological information is necessary, but we also see that it is not convenient to access; it lacks the human resources that often make the most of such material, as in Section 3.2 where a lack of a single access point means that secondary information needs to be sought out manually.

In reality, we have a chance to design and build support for the meta-ontological in the light of current experience. OWL has virtually no support, apart from some *ad hoc* solutions, for carrying meta-ontological knowledge. We would advocate such a separation of the ontological from meta-ontological and this is where a Web 2.0 approach could help.

Our current scenario places too much reliance on assessment through simple linguistic inspection of, for instance, terms. These are labels for concepts and a simple assumption of lexical matching implying conceptual matching is dangerous. For example, in biology, it might seem safe to assume that hepatocyte and liver cells are the same thing. In fact, cells in the liver include hepatocyte cells, but also include adipocyte cells. Hepatocytes make the liver the liver, but there are other cells too.

Ontologies are only intuitively discoverable through the identification and inspection of the appropriate individual terms. Even the construction of linguistic definitions can leave ambiguous meanings for those inspecting an ontology, with no real way to find out how those definitions were converged upon. Even with logical definitions, we still rely upon natural language labels. The aim of languages such as OWL-DL is, however, to minimize potential ambiguities through logical descriptions.

Overall, there should be a synergy between logical and linguistic definitions.

Non-ontologist domain experts will attach intrinsic meaning to terms by drawing on their internal knowledge and the context in which a term is used. It is possible to restrict the intrinsic meaning of a term using the consensus of a domain, so long as it is stated in the context of the purpose of the controlled vocabulary. Interpretation of meaning in these controlled vocabularies still requires a human agent and the knowledge is logically inaccessible to a computer agent.

Thus, we define the inline linguistic portions of ontology files as meta-ontological data. These include anything that a human agent would use for the translation of specific complex logical statements into meaning (including links to other meanings) but are also intrinsically meaningless to the computer. Primarily, these are:

- Terms
  - The specific string by which the logical meaning is labeled, usually considered as the real meaning.
  - E.g. (from celltype ontology) 'subsidiary cell'
- Synonyms
  - Any number of labels that refer to the same meaning.
  - E.g. (cont'd) 'accessory cell'
- Definitions
  - Short, concise description of the meaning, including links to other terms.
  - E.g. (cont'd) 'An epidermal cell associated with a stoma and at least morphologically distinguishable from the epidermal cells composing the groundmass of the tissue'
- Annotations (examples of)
  - Longer, more verbose descriptions.
  - Examples of how the term is used.
  - Explanations of contextual use for the term.
  - Links to term provenance.
  - E.g. (cont'd) DBXREF - TAIR:0000296

Achieving the separation of this meta-ontological layer allows for the consideration of how to manage this mostly linguistic information. This separation is our major desideratum and from this flows the means by which Web 2.0 can provide a platform to expose meta-ontological information and harness and extend the range of group activities.

4.2 Promoting Social Interaction – A Meta-Ontological Workspace

Explicit logic based ontologies for the Semantic Web are going to need to capture implicit knowledge with axioms and restrictions. Yet, unless the experts with the knowledge all manage to learn how to interpret complex logical statements, there needs to be a workspace in which implicit knowledge can be discussed and defined lexically within expert groups. In other words, terms and term linked information can essentially exist independently of the formal environment of ontologies. This implicit knowledge can be used by ontologists as a resource. With such a resource, development of early stage ontologies will not require the construction of formal hierarchies until a critical amount of implicit knowledge has been collected in these more lightweight resources. Also, multiple hierarchies for different purposes could be constructed from the same resource, reusing the collected knowledge in a way not possible in file-oriented development. The ontologists have a way to interact with the domain experts as a community to perform tasks such as the disambiguation of terms before they have been encoded in an ontology, reducing the chance that major revisions of ontological structure will be required. As this resource is shared and linkable, project and domain contexts for terms can be established. These contexts can be used by both the ontologists and the domain experts to traverse the gap into discussions that involve other groups, and discover overlapping scopes more intuitively. Additionally, these resources would provide ideal testing grounds for lexical research (e.g. [7]) that should lead to future improvement on methodologies for these workspaces.

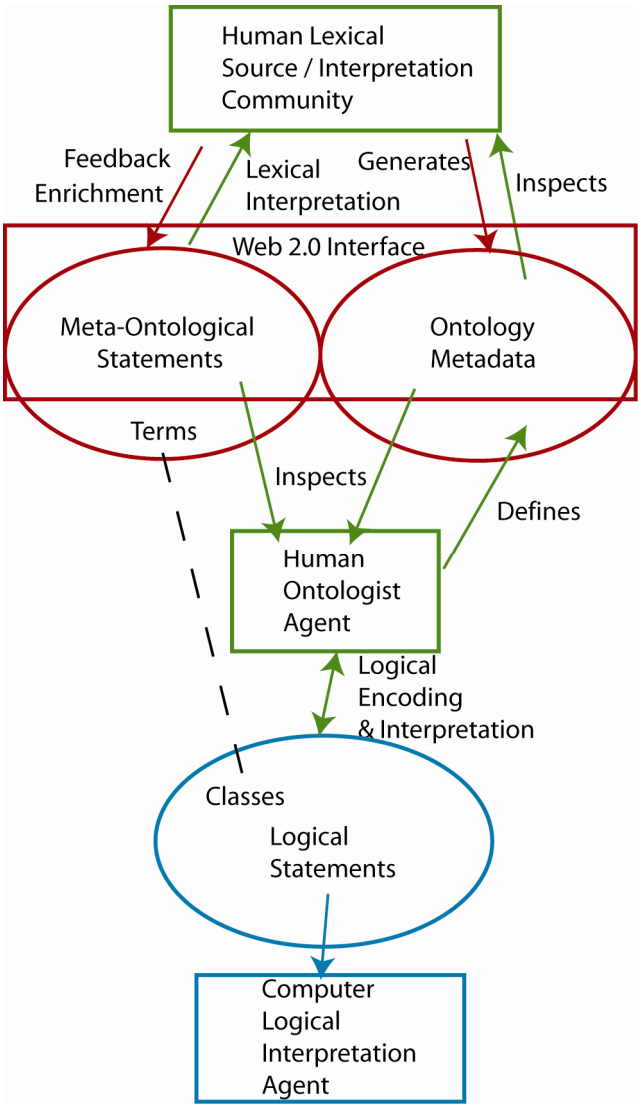

Figure 2: An augmented form of ontology comprehension. Ontology Metadata and Meta-Ontological statements have been separated from the Logical Statements and has been exposed to a community using Web 2.0 principles.

Generating discussion of implicit meaning may sound a little like cutting the domain expert out of the ontology construction process. It should in fact considerably reduce the overhead of ontology development by shifting the discussions based around

intrinsic meanings of terms and which terms are the most appropriate to use away from the attention of the ontologists. It is important not to make the division too wide, as there is a risk that bias could creep in from the ontologists as the domain experts would be unable to assess the implications of certain restrictions and axioms. In terms of feedback to the domain expert from the ontologist, we recognize that there is a need for some sort of consistent translation methodology that can generate accurate textual definitions from logical statements, but we consider this outside of the scope of this article.

It should be clear that such discussion workspaces would be well suited for Web 2.0 style systems. These workspaces should promote the creation of lexicons in which a group of experts can start to add in and inspect lexical information. In this implicit view, it is the terms that are the focus of discussion, not the ontological interpretation, which are two different goals that sometimes get confused during ontology development. Within the workspace, the terms can be discussed, and annotated with textual definitions, comments about usage, links to synonymous terms, requests for clarification etc. Helium was, for instance, discovered in 1894. Of course it was the category of Helium that was discovered, not the instances of the helium atoms (which presumably have existed much before 1894). This is an example of meta-class statements that are part of the ontology. They are class level statements, but those that are well suited to this linguistic, community style of interaction.

The purpose of targeting Web 2.0 as a base for this meta-ontological data is not to completely remove this type of information from the view of the ontology, we merely seek to relocate it so that the incredibly social nature of the definition of knowledge can be coupled with an environment that is equally socially driven (see Figure 2). Modularization of ontologies is seen to be one of the keys to making ontologies viable for the Semantic Web vision, and as such, import mechanisms exist that support the combination of different sources. Lexicons developed by groups could be given URIs, as could all of the terms described in them. Knowledge held in WordNet [17] style lexical resources could be linked using online URIs in a similar way to imported online ontologies taking advantage of well established methods for dealing with words and their meanings at the lexical level.

4.3 Promoting Ontology Sharing and Reuse: An Ontological Metadata Workspace

The production of ontologies that can be effectively shared and reused is a major step towards achieving the goals of the Semantic Web. There are significant barriers to these goals in our current model of ontological comprehension. We have highlighted how ontologists and domain experts alike need to inspect ontologies to assess whether they are appropriate for their needs. Currently, the information that would be necessary to effectively conduct this investigation is hard to find, and does not always come in the same format.

An ontological metadata workspace would provide access to whole-ontology level information for ontologies necessary to carry out light evaluative processes. A collaborative Web 2.0 approach to ontologies would see 'ontology profiles' that include clear statements about the purpose and scope of the ontology and information regarding its status. Ontologies would clearly be labeled as domain and application ontologies to help evaluation, and subsequently, when application ontologies are derived from domain ontologies, this can be marked up and become visible such that ontology level provenance, a history of where everything in an ontology originated from and how it changed over time, can start to be built up.

Introducing a strong community aspect would encourage those developing ontologies to start using tagging, thereby linking up their ontologies to particular domains and projects. Domain ontology construction could be promoted by using ranking systems where inspectors can rate how useful the ontology was in terms of what was expected, assuming that more general ontological models will fit the requirements of more people.

OWL has been made popular for use as an ontology language because of the publicity of the Semantic Web, accessibility of tools for creating OWL ontologies and the fact that it is useful beyond the scope of the Semantic Web. OWL has been used for a lot of purposes, and searching for ontologies based on the content of their files seems like it may be unsustainable as the number of files grows faster than the number of useful ontologies.

Efficient inspection of ontologies can be limited by a large size of ontologies. The current tendency is to build larger ontologies, as the tool support and methodologies for modularization have been slow off the mark until recently. As we learn more about the implications and methods of modularization [22], ontologies can become more manageable, reducing the amount of evaluation cost per ontology. This of course will require better indexing, along with information about how each ontology has been used/imported, perhaps leading to a 'shopping cart' model for highly modular ontology construction.

Perhaps one of the most motivating factors for achieving this desire for more effective inspection is the aspect of learning. Once it becomes easy to empirically see what constitutes a good and useful ontology, then these features get propagated and discussed. As has been noted in [1], the viral spread of understanding how to write HTML was in part because existing HTML could be inspected and copied. Also, the effect of newly written HTML was instantly verifiable in a Web browser. It is harder to have this sort of verification with ontologies, and there are a lot of conflicting styles of ontology development with no consensus of what is 'right', If the Ontological Metadata Workspace were to be realized, then a hub of comparable, commented and marked-up ontologies could develop in a much quicker and consistent fashion than the solitary efforts that are currently the norm.

5. BIO-ONTOLOGIES: EXPERIENCES AND PERSPECTIVES

While our discussion in this article is most pertinent to the notion of the Semantic Web as a whole, it originates from the discipline of bioinformatics. Biologists were early adopters of the Web as a means of disseminating data and the tools for their analysis. These data and tools are developed in a highly autonomous manner and consequently they are beset by both syntactic and semantic heterogeneities. Bioinformaticians have seen ontologies as a means to create common understandings for human and computers about the meaning of data in their distributed resources in a life science Semantic Web [13]. The DNA sequences of different organisms, for example have a common representation, but this is not so for the functional knowledge associated with those sequences. So, the sequences can be interpreted by humans and computers, but not what is known about those sequences.

Consequently, biologists have created ontologies to describe, for instance, the functional attributes of DNA and proteins [3, 11].

Bioinformatics has, therefore, much Web accessible data described by ontologies. The W3C have recognized a nascent Semantic Web in this domain in the development of the Health Care and Life Sciences SIG[5]. It is a significant feature of the move towards ontologies in this sector that it is biologists who build these tools, with some guidance from ontologists. Whilst this community has not made great use of OWL, but its own representation, OBO[6], it still provides a good representation of Semantic Web activities.

The OBO ontologies have significant standing in biological communities, and it is perhaps the community building aspect that fuels this standing, as it includes:

- A large number of centrally available OBO ontologies[7]
- The OBO-Edit OBO ontology development tool that is specifically designed by a working group of users.
- A committee, the OBO-Foundry[8], that has been set up and has produced a set of principles for new OBO ontologies to aspire to, including the promise of textual definitions for all terms and good documentation for all ontologies.
- The OBO file format, for which the primary goals include human readability and ease of parsing together with a syntax that makes them exportable as OWL.
- Pages on the SourceForge[9] open source software development site, which includes the potential for project information, forums, downloads and issue tracking by which suggestions for new terms and modifications can be submitted.

Contributors to OBO are starting to pull together as a virtual community by pooling its resources on the Web. The Gene Ontology [3] saw a phenomenal growth in the number of terms it contained through user interaction alone that is well documented [5], such was the demand for the resource to represent so many researchers. Since then, the trend has continues as more and more biological domains aim to be represented by OBO.

The caveat for the relative success of OBO has probably been similar to that of Web 2.0 over Semantic Web (so far). Formality and methodology have temporarily made way for ease of use and ease of interaction. Interestingly, the majority of the OBO ontologies clearly state that they are "structured controlled vocabularies", which require nothing like the expressive power of OWL, and little in the way of knowledge engineering because the statements linking things do not require it. This is not for any other reason than nothing more complex than this is required, OBO ontologies are used for marking biological data so that they can be linked if they are annotated in the same way. Primarily, these ontologies contain a hierarchy of terms denoting 'is_a' relationships. Less often but still common are 'part_of' relationships, and occasionally other properties key to biology such as 'develops_from'. Despite having the full expressivity of OWL available in the OBO 1.2 syntax, there is little evidence to suggest that the developers in this community either see the need or have the will to take on this level of expressivity in their knowledge.

Perhaps then, this community can be a model for the future of ontology development on the Web. Quick and easy development of terms by engaging the user, employing Web 2.0 design principles to forge more coordinated communities for development of Semantic Web technologies. Web 2.0 has the capability to expose all of the 'light' lexical issues and some basic assertions of linking meaning to terms. 'Heavier' more expressive assertions in OWL are in the domain of the ontologist, who can be informed by the interactions they can have with domain experts and other ontologists through Web 2.0 communities.

6. DISCUSSION

We propose the construction of ontology specific resources, using the Web as a platform, which specifically deals with the management of lexical meta-ontological aspects of ontology development together with the management of ontology metadata. The applications of Web2.0 are geared towards harnessing these types of community interaction, which is precisely the sort of interaction that is not supported in the current model of ontology development. Dealing with meta-ontological data in downloadable ontology files and disparate descriptions of ontology metadata on development sites is prohibitive to a more universal appreciation of ontology design and implementation.

A centralized resource for sharing OWL resources would act as a hub for community learning, sharing and reusing of ontology resources, bringing together ontology users and builders in a way that is currently not possible. Designing ontologies by consensus in such workspaces would encourage best practice and speed up the uptake of the more complicated Semantic Web technologies, starting with OWL and the knowledge that is to be contained within. At the same time the system would provide a measure of control, ensuring that the dangers of misinterpreting the powerful semantics of OWL by untrained eyes are avoided. Having the community built lexical resources is the beginning of an opportunity to link up ontologists with a more specific system that can refer to the online lexical corpus.

The widespread realization of the Semantic Web will depend on the production of ontologies that can be effectively shared and reused, but in order to achieve this, the overheads of ontology development and ontology comprehension have to be considerably reduced. The OBO community/consortium has effectively demonstrated the advantages of lowering these overheads by engaging a community of domain experts in ontology development. OBO ontologies, however, are for human interpretation, so the true Semantic Web vision of human and computational understanding is not addressed. At the same time, highly expressive OWL-DL ontologies, for both computer and human interpretation *are* being produced, but largely in isolation. We propose a solution here which would bridge the gap between these approaches and effectively enable the same type of domain expert community engagement for formal ontologies.

7. ACKNOWLEDGMENTS

Funding for this work was through BBSRC grant BBS/B/17156.

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

[6] http://www.geneontology.org/GO.format.obo-1_2.shtml

[7] http://obo.sourceforge.net/

[8] http://obofoundry.org/

[9] http://sourceforge.net/

8. REFERENCES

[1] Alani, H., Position paper: ontology construction from online ontologies. in *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland, 2006), ACM Press, New York, NY, 491-495.

[2] Alani, H., Brewster, C. and Shadbolt, N., Ranking Ontologies with AKTiveRank. in *International Semantic Web Conference*, (Athens, GA, USA, 2006).

[3] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. Gene Ontology: tool for the unification of biology. *Nat Genet*, *25* (1). 25-29.

[4] Austin, M.A., III and Samadzadeh, M.H., Software comprehension/maintenance: an introductory course. in, (2005), 414-419.

[5] Bada, M., Stevens, R., Goble, C., Gil, Y., Ashburner, M., Blake, J.A., Cherry, J.M., Harris, M. and Lewis, S. A short study on the success of the Gene Ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, *1* (2). 235-240.

[6] Berners-Lee, T., Hendler, J. and Lassila, O. The Semantic Web. *Scientific American*, *284*. 34-43.

[7] Bodenreider, O., Burgun, A. and Rindflesch, T.C. Assessing the consistency of a biomedical terminology through lexical knowledge. *International Journal of Medical Informatics*, *67* (1-3). 85-95.

[8] Brank, J., Grobelnik, M. and Mladenic, D., A survey of ontological evaluation techniques. in *Conference on Data Mining and Data Warehouses*, (Ljubljana, Slovenia, 2005).

[9] Brickley, D. and Miller, L. FOAF vocabulary specification, 2005.

[10] Corcho, O., Fernandez-Lopez, M. and Gomez-Perez, A. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, *46* (1). 41-64.

[11] Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, *6* (5). R44.

[12] Golbreich, C., Zhang, S. and Bodenreider, O. The foundational model of anatomy in OWL: Experience and perspectives. *Web Semantics: Science, Services and Agents on the World Wide Web*, *4* (3). 181-195.

[13] Good, B.M. and Wilkinson, M.D. The Life Sciences Semantic Web is Full of Creeps! *Brief Bioinform*, *7* (3). 275-286.

[14] Guarino, N. and Christopher, W. Evaluating ontological decisions with OntoClean. *Commun. ACM*, *45* (2). 61-65.

[15] Kirk, D., Roper, M. and Wood, M., Identifying and addressing problems in framework reuse. in, (2005), 77-86.

[16] Li, D., Tim, F., Anupam, J., Rong, P., Cost, R.S., Yun, P., Pavan, R., Vishal, D. and Joel, S. Swoogle: a search and metadata engine for the semantic web *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM Press, Washington, D.C., USA, 2004.

[17] Miller, G.A. WordNet: a lexical database for English. *Communicationd of the ACM*, *38* (11). 39-41.

[18] Noy, N.F., Guha, R. and Musen, M.A., User Rating of ontologies: Who will rate the raters? in *AAAI 2005 Spring Symposium on Knowledge Collection from Volunteer Contributors*, (Stamford, CA, USA, 2005).

[19] Noy, N.F. and Klein, M. Ontology Evolution: Not the Same as Schema Evolution. *Knowledge and Information Systems*, *V6* (4). 428-440.

[20] Pulido, J.R.G., Ruiz, M.A.G., Herrera, R., Cabello, E., Legrand, S. and Elliman, D. Ontology languages for the semantic web: A never completely updated review. *Knowledge-Based Systems*, *19* (7). 489-497.

[21] Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H. and Wroe, C. *OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns*, 2004.

[22] Rector, A.L. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL *Proceedings of the 2nd international conference on Knowledge capture*, ACM Press, Sanibel Island, FL, USA, 2003.

[23] Rosse, C. and Mejino, J.L.V. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, *36* (6). 478-500.

[24] Sabou, M., Lopez, V., Motta, E. and Uren, V., Ontology Selection: Ontology Evaluation on the Real Semantic Web. in *WWW2006*, (Edinburgh, UK, 2006).

[25] Schaffert, S., Gruber, A. and Westenhaler, R., A Semantic Wiki for collaborative knowledge formation. in *Semantics*, (Vienna, Austria, 2005).

[26] Shadbolt, N., Hall, W. and Berners-Lee, T. The Semantic Web revisited. *IEEE intelligent systems*, *21* (3). 96-101.

[27] Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A. and Brass, A. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, *16* (2). 184-186.

[28] Uschold, M. Knowledge level modelling: concepts and terminology. *The Knowledge Engineering Review*, *13* (1). 5-29.

[29] Vrandecic, D., Pinto, S., Tempich, C. and Sure, Y. The DILIGENT knowledge processes. *Journal of Knowledge Management*, *9* (5). 85-96.

[30] Whetzel, P.L., Parkinson, H., Causton, H.C., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., Sansone, S.-A., Taylor, C., White, J. and Stoeckert, C.J., Jr. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, *22* (7). 866-873.

[31] Wolstencroft, K., Lord, P., Tabernero, L., Brass, A. and Stevens, R. Protein classification using ontology classification. *Bioinformatics*, *22* (14). e530-538.