OntologySummit2014 Communique

Ref. Final Outline (ver. 1.0) from MichaelGruninger / LeoObrst - 2014.03.18-15:46 EDT
- at: http://interop.cim3.net/forum/ontology-summit-org/2014-03/msg00027.html

Here is the proposed draft for this year's Communique
Please provide feedback about the overall structure, and any ideas about further refinement.

1. Introduction, Scope, Motivation

Since the beginnings of the Semantic Web, ontologies have played key roles in the design and deployment of new semantic technologies. Yet over the years, the level of collaboration between the Semantic Web and Applied Ontology communities has been much less than expected. Within Big Data applications, ontologies appear to have had little impact.

Ontology Summit 2014 provided an opportunity for building bridges between the Semantic Web, Linked Data, Big Data, and Applied Ontology communities. On the one hand, the Semantic Web, Linked Data, and Big Data communities can bring a wide array of real problems (such as performance and scalability challenges and the variety problem in Big Data) and technologies (automated reasoning tools) that can make use of ontologies. There is a particular emphasis on Web, i.e. making sense of knowledge distributed over the Web. This is in contrast to, say, using a local OWL reasoner on a small ontologies, where the only "Web" aspects are using IRIs as symbol names, and employing inference rules based on an open world assumption. On the other hand, the Applied Ontology community can bring a large body of common reusable content (ontologies) and ontological analysis techniques.

Three focus areas arose from the Summit.

1. How can the Semantic Web and Big Data communities share and reuse the wide array of ontologies that are currently being developed?
2. How are ontologies actually being used in Semantic Web and Big Data applications, and how does this differ from existing applications of ontologies?
3. Identifying and overcoming ontology engineering bottlenecks.

2. Sharable and Reusable Ontologies

Semantic technologies such as ontologies and related reasoning play a major role in the Semantic Web and are increasingly being applied to help process and understand information expressed in digital formats. Indeed, the derivation of assured knowledge from the connection of diverse (and linked) data is one of the main themes of Big Data. One challenge in these efforts is to build and leverage common semantic content while avoiding silos of different ontologies. Examples of such content are whole or partial ontologies, ontology modules, ontological patterns and archetypes, and common, conceptual theories related to ontologies and their fit to the real world. However, crafting of whole or partial common semantic content via logical union, assembly, extension, specialization, integration, alignment and adaptation has long presented challenges. Achieving commonality and reuse in a timely manner and with manageable resources remain key ingredients for practical development of quality and interoperable ontologies.

2.1 Why is there not more Ontology Reuse?

Despite development of such things as foundational top-level ontologies (such as DOLCE and BFO), and the availability of broad domain models (such as SWEET) as starting points, the amount of reuse seems quite low in practice. We can examine several possible reasons for this situation and determine whether or not they present fundamental obstacles to ontology reuse.

2.1.1 Mismatch

One potential reason for little reuse is that the required ontologies simply do not exist, that is, the ontologies that have been designed do not satisfy the needs of users with new applications. Determining whether or not an existing ontology meets the needs of a user leads to the discussion of the ontology lifecycle (the topic of Ontology Summit 2013), in which we consider ontologies in the context of requirements developments, ontology analysis, design, evaluation, and deployment. In particular, users need to understand how the requirements for an ontology can be captured using techniques such as competency questions. There are many opportunities for reuse, but a domain and its competency questions must be understood first. Often, reuse fails because it is attempted before the requirements, underlying concepts and assumptions (driving the creation of the content) are fully understood. These help guide useful selection of what to reuse from the massive supply now available. Thus, there may be ontologies that can be reused, but users do not recognize that the existing ontologies do in fact meet their needs.

2.1.2 Finding Mr. Right Ontology

Another possibility is that the ontologies exist but it is difficult to find them. Where can users find this content? Efforts such as Linked Open Vocabularies (LOV) and the Open Ontology Repository are beginning to address this issue. Of course, more than a simple registry of ontologies is needed -- there must also be ways of organizing and annotating the ontologies with the appropriate metadata so that users can find the ontologies that match their requirements (as discussed in the preceding section). In addition to notions such as provenance (captured by efforts such as OMV), such metadata will also need to include the competency questions, ontological commitments and design decisions which were used in the development of the ontology.

2.1.3 This Ontology is Too Big ...

Like Goldilocks and the Three Bears, perhaps the ontologies exist but they have issues that prevent them from being reused by particular end users.

An ontology may incomplete, that is, it might not satisfy all the requirements for a particular application. Existing ontologies are usually insufficient for an application and must be extended. In this regard, it is important to remember the role of competency questions in the selection of what to reuse. If users are able to match their competency questions with the competency questions supported by the existing ontologies, they can better determine how the ontologies can be extended to satisfy all of the requirements.

An ontology may be too big, since the user only needs parts of the ontology, and this leads to the problem of supporting partial reuse. An obvious approach to this problem is modularity, but the modularization of existing ontologies itself remains a difficult problem. The assembly, extension, specialization, integration, alignment and adaptation of small modular ontologies needs to become part of the ontology development methodology. Ontology repositories may also be able to provide more explicit support for the modularization of ontologies as they are uploaded.

Finally, an ontology may not be in the representation language that a user expects, so that even if the ontology meets all the requirements as captured by competency questions, it does not meet the additional requirements that arise from the role that the ontology plays in overall system design and deployment. In this case, it is important to recognize that reuse of an ontology can occur across languages. For example, given an ontology in an expressive language such as Common Logic, we can specify weaker versions, or fragments of the ontology in other representation languages, such as RIF, OWL, and RDF. Each of these fragments can then be reused by a wider variety of applications. In particular, applications on the (big) Web of Data can profit from using lightweight ontologies and methods. These lightweight definitions can provide focused ontological commitment, and still afford the benefits of complete semantics and support reasoning. The idea is to find and use ontology parts that are appropriately expressive.

2.1.4 Integration

Reuse also requires integration of multiple ontologies, and the integration problem can be just as difficult as designing a new ontology. A key technique is the creation of "integrating" modules that merge the semantics of the reused components.

Ontology mapping plays a key role in reuse when there are multiple ontologies that can potentially be used. Understanding how different ontologies in the same domain (e.g. multiple time or process ontologies) are related to each other is an essential part of determining whether or not one ontology can be integrated with others, even in cases where the terminologies are not the same.

Integration arises most acutely in the variety problem with Big Data, and ontologies can tackle variety by aiding the annotation of data and metadata. Data sets will differ in completeness of metadata, granularity and terms used. Ontologies can reduce some of this variety by normalizing terms and filling in absent metadata. An additional problem in many Big Data applications is that terminology used at one time for one set of data might have a different meaning than what appears to be the same terminology used at a different time for another set of data. For ontologies to deal with this effectively, they must not only to evolve over time but also map the previous meanings to the new ones.

2.1.5 Just Do It Yourself

It might be easier to design a new ontology for an application rather than spend time to find possible ontologies for reuse and then to understand them sufficiently well enough to determine whether or not they satisfy the user's requirements. If this is indeed the case, then it will be important to create new ontology development environments that better support design for reuse.

Ontology Design Patterns are an approach that can be used to directly incorporate reuse into the ontology

development methodology. By explicitly capturing the reusable aspects of an ontology, a design pattern allows the designer to more effectively specify the commonalities among otherwise disparate components.

2.2 Recommendations: Towards Best Practices

- Wise reuse possibilities follow from knowing the project requirements. Competency Questions should be used to structure the requirements, as part of an agile approach. The questions help frame areas of potential content reuse.
- Be tactical in formalization. Take what content you need and represent it in a way that directly serves your objective.
- Small ontology design patterns provide more possibilities for reuse because they have low barriers for creation and potential applicability, and offer greater focus and cohesiveness. They are likely less dependent on the original context in which they were developed.
- Better metadata and documentation of and for ontologies and schemas is needed to facilitate reuse. Some work in this area, such as Linked Open Vocabulary, is underway and should be supported.
- Better ontology and schema management is needed. Governance needs a process and that process needs to be enforced. The process should include open consideration, comment, revision and acceptance of revisions by a community.
- The explicit specification of ontology fragments should be incorporated into development methodologies in the ontology lifecycle.

3. Using Ontologies in Big Data and the Semantic Web

The Web of Data provides great opportunities for ontology-based services, but also poses challenges for tools for editing and using ontologies, and to techniques for ontological reasoning and ontology engineering.

- How is the semantic content being used/shared/reused?
- What is the role of ontologies in these applications?

3.1 What Ontologies are Needed?

Ontologies can tackle variety in Big Data by aiding the annotation of data and metadata. Data sets will differ in completeness of metadata, granularity and terms used. Ontologies can reduce some of this variety by normalizing terms and filling in absent metadata.

A more recent use of ontologies for data analytics that has potential for high impact is for managing data provenance, including any transformations, analyses and interpretations of the data that have been performed. Currently, most Big Data projects handle provenance in an ad hoc rather than systematic manner. Ontologies for describing data provenance do exist, such as the PROVO ontology. Developing standard ontologies for commonly used, but informal, process models such as the OODA loop and JDL/DFIG fusion models could have a significant impact on data analytics. The KIDS framework is an

example of such a formalization. Standard statistical reasoning ontologies are another area that has the potential for having a high impact.

At the global level, there are too many domains to have very deep semantics common to them all. Nevertheless, Schema.org has been tackling the formidable problem of developing a generally accepted vocabulary that is now being used by over 5 million domains, and gradually introducing deeper semantics. Incorporation of ontologies into the Schema.org framework is challenging but has the potential of significant benefits.

It is unlikely that we will be able to make Web-wide ontological commitments. Where projects such as Watson (IBM) limit themselves to a few simple taxonomies, other large collaboration efforts may agree on a limited subset of ontologies, such as parts of some molecular biology ontologies, the Gene Ontology and other OBO Foundry ontologies. It is possible to create ontologies from big data, but it is difficult. Manually building ontologies is labour intensive, mining data for reusable information suffers from the potential inconsistency, incompleteness and irrelevance of data “out there”, and machine learning may require further research for being applied to learning ontologies from big data.

3.2 Expressiveness

The notion of expressiveness refers to the logical properties of an ontology representation language. The Ontology Spectrum characterizes the range of different languages from RDF, OWL, and RIF through to Common Logic and modal logics. A critical question for both ontology users and developers is the selection of the appropriate language. In fact, many of the earlier debates about the nature of ontologies (i.e. what is an ontology?) have their roots in the different expectations that users have for the expressiveness of the underlying ontology representation language.

The expressiveness of an ontology representation language is closely related to the requirements for any ontology that is intended for a particular application. RDF, the native language of linked data, goes a long way in big data settings, because of the low ontological commitment it enforces, while still allowing to link to complex descriptions. On the other hand, many traditional applications of ontologies, such as semantic integration and decision support, have required more expressive languages such as RIF and Common Logic.

Building lightweight ontologies – one often speaks of vocabularies in such cases – requires new, agile engineering techniques. The recent Linked Open Terms (LOT) approach starts with reuse, taking advantage of the great number of vocabularies that already exist on the Web. Where the terms needed to describe the data at hand cannot be found in existing vocabularies, the knowledge engineer will have to create new ones, but is encouraged to link them to existing ones.

The Watson developers did not build a formal ontology of the World, with which they would try to unify formal logical representations of the questions. Instead, they locally learned ontologies on demand, drawing on formal as well as informal sources, using different reasoning techniques. First, hypotheses are generated. Secondly, evidence is retrieved for them; approaches include keyword matching against as-is natural language text sources. The challenge is to disambiguate types (e.g. “person” vs. “place”) of entities and predicates. This can be partly solved using existing taxonomies such as YAGO.

A swing back to lightweight approaches has also occurred in the field of web services. Generally, a service

consumer finds a web service that a service provider has registered in a central registry, and then communicates with the web service in order to execute it. Semantic web service descriptions, in addition to the basic syntactic WSDL description, is required for finding and comparing service providers, for negotiating and contracting services, for composing, enacting and monitoring them, and for mediating heterogeneous data formats, protocols and processes. Traditionally, the semantics of web services would have been described using heavyweight ontologies such as WSMO or OWL-S based on expressive ontology languages, and these services would have been assumed to communicate by heavyweight XML messages according to SOAP. As the semantics-first modeling approach promoted by WSMO or OWL-S was not taken up in practice, the more recent linked services initiative now promotes a bottom-up annotation and interlinking approach with more lightweight RDF(S) based ontologies for service description, and it faces the reality that the majority of web services is implemented using lightweight REST interfaces.

3.3 Scalability

One aspect in which both Big Data and Semantic Web applications differ from other applications of ontologies is in the scale of the problems which are being addressed. Together with performance constraints, scalability has a profound impact in how the required ontologies are represented and used. The joint demands of volume and velocity lead to tradeoffs between expressiveness of the ontology language and the efficiency of reasoners for that language. The development of large scale reasoning techniques will hopefully alleviate some of these concerns. Another approach is to use hybrid methods which incorporate the semantic content of an ontology without requiring an explicit axiomatization of the ontology to be used with a reasoning engine.

3.4 Questions

- What combination of ontology engineering and reasoning techniques will be used for big data problems?

- Should one even try to represent big amounts of knowledge using ontologies? Do even light-weight ontologies scale to big data? Or would it rather suffice, as use cases in biology suggest, to use ontologies for annotating big data with terms?

4. Barriers and Bottlenecks

Sometimes there are barriers and bottlenecks to the use of ontologies, both in terms of reuse of existing content or in developing new content. These barriers and bottlenecks can be due to the cost of development and deployment of ontologies, the timeliness of being able to deliver solutions, incomplete knowledge about or skills in ontological engineering on the part of the ontology developers, a mismatch between the application requirements and the intended domain coverage and reasoning requirements of the ontologies, the use of inadequate tools at different stages of the ontology development lifecycle, or sociological, cultural, and motivational issues involving the stakeholders, application developers, domain experts, and ontologists.

Realistically, all of the above factor into the cost of development and deployment of ontologies, and so reuse of existing semantic content is the potential cost-saver. However, non-ontological solutions are often done faster and cheaper as one-offs using other technology, because the value proposition of ontology reuse (vastly cheaper development and maintenance costs amortized over multiple ontology application lifecycles) is not communicated to and thus not understood by the supporting community.

The benefits of using ontologies are paradoxically also sometimes a barrier. Because ontologies more accurately reflect the real world domain semantics than other data models, and because the underlying technology is so flexible that it enables very quick proof-of-concept and/or testing, ontology development and reuse also allows throwing together almost anything and then fixing it up after the fact.

4.1 Questions

Among the questions that the Ontology Summit brought forward concerning the barriers and bottlenecks to the use of ontologies are the following:

- How do challenges related to cultural and motivational issues relate to technical issues, e.g., tool support?
- How to get community buy-in?
- What are the tradeoffs between expressiveness vs. pragmatics?
- Who will develop all the ontologies we would ideally need?
- What is the role of crowd-sourcing?
- What is the state-of-the art with respect to quality control?
- How is the industry addressing ontology engineering bottlenecks and what are the technological solutions available on the market today?
- How much (deep) semantics do customers really need?
- Which ontology tools are needed and when are they needed?
- Can ontology acquisition, development, integration, and reuse be automated more?

a) What are the barriers to use/sharing/reuse of existing semantic content?
b) What are the barriers to developing the semantic content that is required?

5. Conclusions and Recommendations

5.1 Recommendations

5.2 Challenge Problems

We can also pose a number of challenge problems which will hopefully serve to focus and guide future collaboration among the three communities of Applied Ontology, Semantic Web and Big Data.

- What ontologies are required by Semantic Web and Big Data applications?

- Is scalability the fundamental challenge for using ontologies on the Web?

- Is the design and application of ontologies on the Web fundamentally different than existing techniques?

- Are we encountering new ontology engineering bottlenecks in Semantic Web and Big Data applications?

- Can the variety problem in Big Data applications be addressed using existing techniques for semantic integration, such as ontology mappings?

- What benchmark data sets can be used to guide future work in the integration of ontologies?

The following emerging common themes and issues have arisen in several of the tracks. The Champions are encouraged to address any of the themes that they find relevant to their Syntheses. The editors will coordinate the contributions and ensure that these ideas get covered within the Communique:

Ontology Reuse
Automated Ontology Gap-Filling (Gaps in Ontologies)
Evolution: Dynamic Ontologies and Adaptation
Crowdsourcing Curation
Building Ontologies from Small Modules
Working with Existing Datatypes
Employing Multiple Languages
Data/Metadata Annotation and Semantic Tagging
Ontology Mapping
Adaptation to Existing Workflows of Domain Experts
Machine-learning Algorithms
Tool Incompatibility
Ontology Design Patterns
Large-scale Reasoning
Time-consuming KR Processes
Education and Buy-in
Variety, Heterogeneity, and Hybrid Methods