A Priori Ontology Modularisation in Ill-defined Domains

Dhavalkumar Thakker[1], Vania Dimitrova[1], Lydia Lau[1], Ronald Denaux[1], Stan Karanasios[2], Fan Yang-Turner[1]

[1]School of Computing, University of Leeds, LS2 9JT, Leeds, UK

[2]Leeds University Business School, Leeds, LS2 9JT, UK

{D.Thakker, V.G.Dimitrova, L.M.S.Lau, R.Denaux, F.Yang-Turner} @ leeds.ac.uk

{S.Karanasios} @ lubs.leeds.ac.uk

ABSTRACT

Modularisation is crucial to create re-usable and manageable ontologies. The modularisation is usually performed *a posteriori,* i.e. after the ontology is developed, and has been applied mainly to well-structured domains. With the increasing popularity of social media, Semantic web technologies are moving towards *ill-defined* domains that involve cognitively-complex processes carried out by humans and require tacit knowledge (e.g. decision-making, sensemaking, interpersonal communication, negotiating, motivating). In such domains, *a priori* modularisation can enable ontology creation to handle the complexity and the dynamic nature of knowledge. This paper outlines an *a priori* modularisation methodology for multi-layered development of ontologies in ill-defined domains, including an upper ontology layer, high-level and reusable domain layers, and case-specific layers. The methodology is being applied in several use cases in two EU projects – Dicode and ImREAL.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Knowledge Representation Formalisms and Methods

I.2.6 [**Learning]:** Knowledge acquisition

General Terms: Design

Keywords: Ontologies, Ontology Modularisation, Ontology Engineering, Semantic Web, Ill-defined domains.

1. INTRODUCTION

Modularisation is highly beneficial for designing large scale, re-usable ontologies. There are many successful implementations of large-scale ontologies in *well-structured* subject areas, e.g. bio-medical (GENE [1]) and medical (GALEN [2]) domains. Recently, the modularisation of such ontologies has been a primary focus of research efforts. The existing successes [3][4][5] offer *a posteriori* approaches and tools to introduce modularity in already developed ontologies that have not been designed in a modular way.

I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria. Copyright 2011 ACM 978-1-4503-0621-8 …$10.00.

There is a growing need for deploying semantic web technologies in *ill-defined* domains that involve cognitively-complex processes carried out by humans and require tacit knowledge, e.g. decision making, sensemaking, collaboration, interpersonal communication, advising, mentoring. With the recent proven success of semantic web and ontologies, the field is ready to take on the challenges offered by ill-defined domains, and to better understand the benefits semantics brings (e.g. reasoning, aggregation, automation) for applications in these domains (e.g. people engaging in collective decision making or sensemaking activities, or sharing stories from their work practice in communicating, negotiating or mentoring). To tackle the complexity of knowledge structuring in ill-defined domains, agile ontology engineering is required. Hence, modularisation from the outset, i.e. an *a priori*, is needed. However, there is a lack of methodologies that explain how to do *a priori* modularisation in ill-defined domains; the existing approaches are mainly *a posteriori* and have been applied mostly in well-structured domains.

We face challenges of ontological modelling in such ill-defined domains in the context of two European projects: ImREAL[1] and Dicode[2]. The ImREAL project deals with aggregating and understanding real world activities in the interpersonal communication domain for learning. Activity model ontology is being developed and used for semantic annotation, aggregation and browsing of user generated content from social media which provides intelligent access to examples with real world experiences that are used for augmenting simulated learning environments. The Dicode project deals with providing synergy between human and machine intelligence in decision making and collaboration within data-intensive environments. Decision making and collaboration ontology is being developed and used for Dicode service integration, as well as for semantic augmentation of medical diagnosis reports and user contributions to argumentative interactions.

To tackle the challenges of ontology engineering in ImREAL and Dicode, we have developed *a priori* modularisation methodology that enables dividing the domain ontology into several modules in order to handle the complexity and dynamicity of ontology modelling in ill-defined domains. The paper first outlines modelling challenges in ill-defined domains and clarifies why

[1] http://imreal-project.eu/

[2] http://dicode-project.eu/

modularisation is required (#2); then, presents *a priori* modularisation methodology for ill-defined domains (#3) and compares the approach with relevant work (#4); and finally, concludes by pointing at our contribution and future plans (#5).

2. CHALLENGES IN ILL-DEFINED DOMAINS AND MODULARISATION

The ill-defined domains we deal with impose several challenges to the construction of a domain model in the form of an ontology:

Domain complexity. Understanding and defining these domains is hard and involves heterogeneous sources of knowledge, i.e. expertise from multiple disciplines, such as psychology, social science, communication and computer science.

Expanding knowledge scope. Due to the evident complexity of the domain, the conceptualisation process is generally dynamic and evolves with user activities. This means that certain concepts and relationships are unidentified in the beginning. Hence, it is not possible to build an all-encompassing ontology in the very first instance. As more knowledge constructs become articulated, it is likely that the domain covered by the ontology may be expanded to cover other domains. The issue of expanding knowledge scope also brings forth problems of reusability and scalability.

Reusability. As the knowledge scope expands, while new knowledge emerges, it is more appropriate to reuse and build on existing structure instead of starting from scratch. Another aspect of reusability is relevant to a global context – we aim at publishing components from the ontology to be reused on the WWW in order, which can enable usage by a wider community and can be helpful for interdisciplinary research [6].

Scalability. The scalability of semantic infrastructure is another potential problem while dealing with expanding knowledge scope. For instance, a tailored ontology built for applications related to ill-defined activities in a medical domain (e.g. decision making in clinical trial diagnosis or interpersonal communication in doctor-patients interviews) can be expanded to reuse large medicine ontology, such as MeSH[3] that currently includes over 250,000 terms corresponding to medical subjects. The rising number of instances in this case can be a serious problem to handle within a single knowledge repository.

Modularisation (which enables creating smaller but connected ontologies or replacing oversized ontologies with smaller ones [7]) can be highly beneficial for handling the key challenges of ontology modelling in ill-defined domains, because:

- Modularity facilitates development of smaller, re-usable pieces of ontologies hence allowing managing *complexity* and promoting understandability of the ontologies across heterogeneous and complex domains.
- Modular approach also allows managing future expansion of the ontologies hence *absorbing expanding knowledge* in a way that either minimum restructuring is required or the restructuring is manageable.
- Modular approach encourages and facilitates *reusability* as it is easier to re-use part of ontologies.
- Modularisation also allows handling *scalability* problems by offering design choices where it is possible to distribute ontology/knowledge bases in different repositories/graphs based on the modules they represent. Hence, queries and reasoning can be done against a subset of the ontology, instead of one large monolithic ontology.

We outline next a modularisation methodology that encourages development of domain ontologies in multiple layers.

3. *A PRIORI* ONTOLOGY MODULARISATION

The methodology outlined in this section is applicable to a class of problems that involve cognitively-complex processes carried out by humans and require tacit knowledge. The understanding of such domains involves inter-disciplinary domain experts who often utilise a theoretical framework to guide the understanding of such domains. Ontology development begins with some *theoretical framework* and arrives to *case specific* domain ontologies. We follow a 3-layered development of domain ontologies consisting of an upper ontology layer, high-level and reusable domain layer, and case specific layer. Each layer can be developed as one or more ontology modules, see Figure 1.

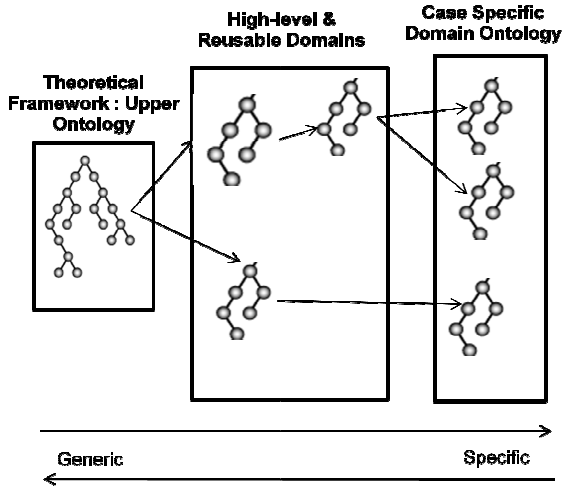

Figure 1. Multi-layered ontology development to conceptualise ill-defined domains.

Upper ontology layer. The upper ontology layer covers base concepts derived from the chosen theoretical framework which facilitates the understanding of the cognitive processes performed in the domain. For example, in ImREAL, we use the Activity Theory [8] to get an understanding of interpersonal communication activities, including different applications such as job interviews, student mentoring, doctor-patient interviews. The Activity Theory provides a framework to identify the key concepts describing an activity, such as SUBJECT, OBJECT, TOOLS, MOTIVATION, OUTCOME, COMMUNITY, DIVISION OF LABOUR, RULE, NORM, ACTION, OPERATION, and the main axioms linking them (e.g. every ACTIVITY has at least one SUBJECT, and is defined by an ACTIVITY VIEWPOINT that is viewpoint of the SUBJECT).

Similarly, when developing a sensemaking ontology in Dicode, we start with a cognitive framework for intelligence analysis [9] which describes the processes people undertake when deriving novel intelligence from massive data (e.g. conducting scientific experiments or clinical trials). Based on this, we identify the key concepts describing intelligence analysis, such as ACTOR, SENSEMAKING OPERATION, INFORMATION SOURCE, INFORMATION FORAGING LOOP, SENSEMAKING LOOP, SUPPORT SYSTEM, OUTCOME, and the main axioms linking them (e.g. an ACTOR (sensemaker)

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

has EXPERTISE, utilises INFORMATION SOURCES to conduct SENSEMAKING OPERATIONS, and produces an OUTCOME.)

Conceptualisation at this level is conceived and developed *independently from its usage context* and avoids defining any concepts that are tied to a particular use case. This upper ontology is then expanded to specialise the base conceptualisation while moving into the direction of a case specific ontology.

High-level reusable domain layer. A middle layer, which *evolves organically through use*, is used to make the connection between the upper ontology layer and the case specific ontology layer. It specifies the upper level concepts in more detail by linking them to more specific concepts that are used across multiple case specific domain ontology. This layer provides more specific detail of the domain, which is still re-usable and high-level. The middle layer can include a number of layers depending on the commonalities between case specific layers. We term these middle layers as high-level reusable domains, since it is not always clear or required to have these set of layers. Only thinking in terms of reusability makes it possible to see requirement for such middle layer. Since we deal with several use cases across the same domain, the middle ontology layer plays a crucial role in connecting the various use cases and ensuring re-usability.

The middle layer in the activity model ontology in ImREAL includes concepts that expand the key activity concepts from the upper layer. For example, interpersonal communication TOOLS are expanded to include MENTAL TOOLS (e.g. VERBAL COMMUNICATION, NON VERBAL COMMUNICATION, BODY LANGUAGE) and PHYSICAL TOOLS (e.g. CV, CLOTHING). The middle layer in the sensemaking ontology in Dicode includes concepts that expand the key information analysis concepts from the upper layer. For example, SENSEMAKING OPERATIONS are expanded with operations relevant to the Dicode use cases (e.g. ABSTRACTING, CLASSIFYING, COMPARING, FILTERING, SEARCHING, VISUALISING); and the information sources are specified (e.g. STRUCTURED DATA, UNSTRUCTURED DATA, QUALITATIVE DATA, QUANTITATIVE DATA).

Case specific layer. The case specific layer defines the concepts that are specific to each use case (i.e. closer to the content and usage). The case specific layers in the ImREAL and Dicode ontologies are derived based on several knowledge sources: interviews with stakeholders from each use case, relevant documentation, user stories. The knowledge sources are analysed by domain experts following the theoretical framework and building a knowledge glossary. The concepts from the glossary are then encoded in an ontology using an intuitive ontology authoring tool ROO, developed in our group [10], which enables active involvement of domain experts. At this stage, it is often discovered that there are commonalities in the use cases, ontological statements are moved to the middle layer. This may include expanding a module from the middle layer or even starting a completely new middle layer module. For examples of concepts in this layer, the reader is directed to the ImREAL and Dicode project web sites.

One of the questions to be answered while developing a methodology is how to identify what modules go into each layer. Standard strategies for module identification are described in [7]. From these, we have followed two: (a) *semantic-driven modularisation* - in both projects, domain experts and knowledge engineers' are jointly deciding on the modular representation of the domain by following the theoretical framework and identifying the generality and specificity of use cases; (b) *reusability-driven modularisation* – a knowledge engineer further inspected the ontology and identified concepts that can be re-used across use cases, and moved them to the middle ontology layer.

The methodology presented here facilitates iterative ontology development. For example, after the first iteration of development for use cases A and B, if it is discovered that the theoretical foundation is also applicable to new application domain – use case C. Then, following the modularisation methodology, the related domain ontology module can be built by extending one of the modules from the middle layer and, if necessary, can update the dependent parent modules. In the case of revision and further knowledge availability, it is possible to pin down one or more modules and revisit these smaller modules, instead of having to deal with one large ontology which will smooth the process of restructuring.

4. RELATED WORK

In the current state of Semantic Web movement, modularisation of ontologies is considered best practice. The existing approaches develop techniques and tools for *extracting ontology modules* or *partitioning ontologies*. [11][12][13] propose an approach to modularisation using *module extraction*, i.e. takes an input ontology and extracts modules that match a set of requirements. In principle, these module extraction approaches can be considered contributing to *a priori* modularisation, i.e. aiding a knowledge engineer in considering and finding existing modules while creating new ontologies. However, these approaches fall short of outlining guidelines a knowledge engineer needs to consider while designing modular ontologies from outset, for example how to decide possible modules for the domain or how to derive generic or specific conceptualisation aspects.

[3][4][5] propose another approach to modularisation using *ontology partitioning*, i.e., dividing large monolithic ontologies into a number of partitions. For example, [3] proposes automatic partitioning of large ontologies into smaller partitions based on the structure of the class hierarchy and by taking into account the internal coherence of the concepts. Since the partitioning is applied after an ontology is created, this approach can be classified under *a posteriori* modularisation methodology.

Limited research considers *a priori* modularisation. The NeOn[4] toolkit is notable for its tool-level support for creating modular ontologies from the beginning. It supports knowledge engineers in both module extraction and ontology partitioning. However, there is no work in applying modularisation to handle the challenges of ontology modelling in ill-defined domains. Most importantly, there are no systematic attempts to apply *a priori* modularisation to optimise ontology modelling and deployment across several use cases within the same domain, which is the focus, and the key contribution, of the work presented here.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have outlined an *a priori* modularisation methodology which is aimed at addressing the challenges posed by ill-defined domains that require (a) an understanding of cognitive-complex processes and (b) knowledge sources, tacit or otherwise, from multiple disciplines. The paper contributes to research and practice in semantic technologies by presenting an approach for *a priori* modularisation that can be followed for addressing key challenges of ontology engineering in ill-defined domains (which are becoming popular with the ubiquity of social media and user generated content).

[4] http://neon-toolkit.org/wiki/Main_Page

This work is ongoing. The next phase includes parallel activities in both ImREAL and Dicode extending the work presented here. To examine the appropriateness of the proposed methodology for ill-defined domains we apply it in several use cases in both projects which fit the characteristics of ill-defined domains.

In ImREAL, we have developed the ontology for the job interview use case (including the three layers), and have used it in an intelligent content assembly workbench which integrates and semantically augments relevant content from YouTube [14] and user job interview stories. Following the Activity Theory framework, we have also conceptualised the second use case developing the knowledge glossary. We are currently encoding the additional concepts brought by the second use case, which leads to expanding the middle layer (e.g. to include cultural norms) and defining a new module in the use case layer. This will allow us to utilise the developed semantic augmentation services to also assemble content related to cultural awareness in mentoring. The third use case involves doctor-patient interviews. It will be conceptualised using the same methodology, which will lead to expanding the middle layer (e.g. with emotional signals) and adding another module in the use case specific layer.

We are also applying the modularisation methodology for building data intensive decision making and collaboration support ontology in Dicode. There are three use case partners that have diverse need for intelligence analysis in data intensive environments [15]. For example an use case from breast cancer scientific research involve intuitive decision making such as judgement on sources for analysis or perceiving gaps in analysis while other use case on medical diagnosis of Rheumatoid Arthritis involves decision making tasks such as identifying change in condition or identifying tendency in wrists. We are currently encoding the Dicode ontology for the medical diagnosis use case. Following the intelligence analysis sensemaking framework we have developed the upper layer, and have used the interviews with stakeholders to built the knowledge glossary and identify concepts for the middle and use case specific layer. We will apply the ontology to semantically augment and link medical diagnosis reports to facilitate the sense makers. The ontology will also enable us to describe the services developed by Dicode to support intelligence analysis.

The *a priori* modularisation methodology presented here, which is based on theoretical frameworks and use case specific information, allows us to compare different use cases and enables reusability of ontology querying and semantic augmentation services across similar use cases. The two EU projects - ImREAL and Dicode - provide practical testbeds for validating the methodology, helping us to examine its benefit in future work.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT 257831 (ImREAL project) and grant agreement no ICT 257184 (DICODE project). The authors are grateful to project partners for stimulating ideas, and specifically to the domain experts involved in ontology authoring.

7. REFERENCES

[1] Harris, MA. and Clark JI, I.A. 2006. The Gene Ontology (GO) project in 2006. *Nucleic Acids Research* 34, 322-326.

[2] Rector, A. and Rogers, J. 1999. Ontological Issues in using a Description Logic to Represent Medical Concepts: Experience from GALEN. *IMIA WG6 Workshop: Terminology and Natural Language in Medicine* (Phoenix Arixona).

[3] Stuckenschmidt, H. and Klein, M. 2004. Structure-based partitioning of large concept hierarchies. In *Proc. 3rd International Semantic Web Conference (ISWC2004)* (Hiroshima, Japan, November 7-11, 2004).

[4] Schlicht, A. and Stuckenschmidt, H. 2008. A Flexible Partitioning Tool for Large Ontologies. In *Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Sydney, Australia, December 9-12, 2008). IEEE Computer Society 482-488.

[5] Kusnierczyk, W. 2008. Taxonomy-based partitioning of the Gene Ontology. *Journal of Biomedical Informatics* 41, 2 (2008), 282-292.

[6] Simperl, E. 2009. Reusing ontologies on the Semantic Web: A feasibility study. *Data Knowledge Engineering*. 68, 10 (2009), 905-925.

[7] Parent, C. and Spaccapietra, S. 2009. Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization. In *Modular Ontologies*. Springer-Verlag, Berlin, Heidelberg, 2009, 5-23

[8] Leontiev, A.N. 1978. Activity, consciousness, and personality. Prentice-Hall, Englewood Cliffs, NJ.

[9] Pirolli, P. and Card, S. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis* (McLean, VA, USA, May 2-6, 2005).

[10] Denaux, R., Dolbear, C., Hart, G., Dimitrova, V. and Cohn, A.G. 2011. Supporting Domain Experts to Construct Conceptual Ontologies: A Holistic Approach. *Journal of Web Semantics* (in press).

[11] Doran, P., Tamma, V., and Iannone, L. 2007. Ontology module extraction for ontology reuse: an ontology engineering perspective. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (Lisbon, Portugal, November 6-10, 2007), ACM, 61-70.

[12] Seidenberg, J. and Rector, A. 2006. Web Ontology Segmentation: Analysis, Classification and Use. In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland, May 23-26, 2006). ACM, New York, 13-22.

[13] Grau, C., Horrocks, B., Kazakov, I., and Sattler, Y. U. 2008. Modular reuse of ontologies: Theory and practice. *J. of Artificial Intelligence Research*, 31, 273-318.

[14] Ammari, A, Dimitrova, V, and Despotakis, D. 2011, Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models. In *Proceedings of Workshop on Augmented User Models at the 19th International Conference on User Modeling, Adaptation and Personalisation* (Gerona, Spain, July 11-15, 2011).

[15] Ammari, A, Dimitrova, V, Lau, L, Tzagarakis, M, and Karacapilidis, N. Augmented Collaborative Spaces for Collective Sense Making: The Dicode Approach. *Proceedings of on Adaptive Support for Team Collaboration, at the 19th International Conference on User Modeling, Adaptation, and Personalization Personalisation* (Gerona, Spain, July 11-15, 2011).