# Structure-based Analysis and Modularization of Ontologies

Gökhan Coskun

Freie Universität Berlin
coskun@inf.fu-berlin.de
Supervisor: Prof. Dr.-Ing. Robert Tolksdorf
Phase of PhD: Second Phase

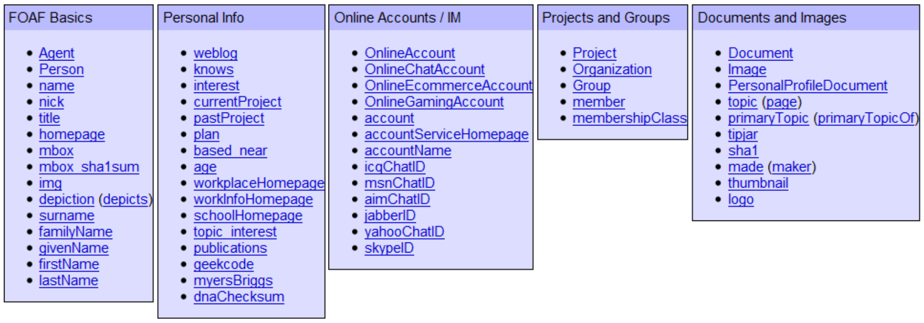
**Abstract.** Defined as a problem-relevant, explicit and formal specification of a shared conceptualization, ontologies became a new hype in the context of the Semantic Web. Being a shared knowledge its potential for information integration in the large World Wide Web is promising. But either the reuse of existing ontologies or the matching of different ontologies is unavoidable for this integration. Therefore means for analyzing ontologies as well as modularization techniques for partial reuse are very important and a key for the success of information integration based on ontologies. Considering ontologies as networks of concepts connected through properties, this work makes use of network analysis techniques and graph measures. It aims at gaining insight to which extent structure based techniques can be modified so they are paying attention to the semantics inherent in ontologies. The expected contribution is a method and tool support for ontology engineers to analyze and modularize ontologies in a (semi-) automatic way. The main goal is to improve the (re-)usability and maintainability by increasing the understandability and allowing ontology engineers to refactor and reuse existing ontologies easily.

**Key words:** Ontology Reuse, Ontology Modularization

## 1 Problem Statement

During the last two decades the interest in using ontologies has increased. According to the last few years this trend was mainly driven by the vision of the Semantic Web [4]. Defined as a problem-relevant, explicit and formal specification of a shared conceptualization, the importance of ontologies lays in the deep problem and domain analysis to create them. Because a good analysis clarifies the structure of the domain knowledge [8]. But a good analysis as only one part of the overall ontology creation process is a very cumbersome and time-consuming activity. In order to provide some structural guidance for the ontology creation process some ontology engineering methodologies have been proposed (e.g. Cyc Method [18], Uschold and Kings [26], Grüninger and Fox [12], KACTUS approach [3], Methontology [10], On-To-Knowledge [24], and NeOn [23]). They were followed by some machine learning approaches [7, 9, 21], which aimed at reducing the need for human intervention. The newest trend in ontology engineering is to build ontologies with a community in a collaborative manner (e.g. Holsapple et al. [14], DILIGENT [20], Dogma [16], HCOME [17], RapidOWL [2]). In most methodologies ontology reuse is recommended, because it is expected to reduce engineering costs by avoiding re-building already existing conceptual models. Apart form reducing costs, reusing existing ontologies increases interoperability from the viewpoint of the Semantic Web, where ontologies are primarily considered as shared knowledge [6, 5].

Even though most of these approaches mention the reuse of existing ontologies as possible starting point, none of them describe in detail how to discover and analyze candidate ontologies. This is very important, because reusing ontologies presumes availability of already existing ontologies and discovery of potential candidates for the particular use case. In this regard Ontolingua and OntoSelect libraries are available and search engines as Swoogle[1], Watson[2] and Ontosearch[3] has been already developed in the context of the Semantic Web. Although the problem of discovering potential candidate ontologies seems to be mainly solved, there is still an issue on selecting appropriate ontologies as well as understanding and analyzing them. Even though the Resource Description Framework (RDF) and the Web Ontology Language (OWL) files are based upon the Extensible Markup Language (XML) syntax, which is declared to be human readable, it takes some time to comprehend the content and the main structure and to understand the main idea and purpose of the model. Even the Friend of a Friend (foaf) vocabulary[4] which is rather small shows in its specification a grouping of the concepts as illustrated in Figure 1, in order to provide the reader an easier way to understand this vocabulary.



FOAF Basics

- Agent
- Person
- name
- nick
- title
- homepage
- mbox
- mbox_sha1sum
- img
- depiction (depicts)
- surname
- familyName
- givenName
- firstName
- lastName

Personal Info

- weblog
- knows
- interest
- currentProject
- pastProject
- plan
- based_near
- age
- workplaceHomepage
- workInfoHomepage
- schoolHomepage
- topic_interest
- publications
- geekcode
- myersBriggs
- dnaChecksum

Online Accounts / IM

- OnlineAccount
- OnlineChatAccount
- OnlineEcommerceAccount
- OnlineGamingAccount
- account
- accountServiceHomepage
- accountName
- icqChatID
- msnChatID
- aimChatID
- jabberID
- yahooChatID
- skypeID

Projects and Groups

- Project
- Organization
- Group
- member
- membershipClass

Documents and Images

- Document
- Image
- PersonalProfileDocument
- topic (page)
- primaryTopic (primaryTopicOf)
- tipjar
- sha1
- made (maker)
- thumbnail
- logo

**Fig. 1.** Concept groups of the FOAF vocabulary in the specification

[1] http://swoogle.umbc.edu

[2] http://watson.kmi.open.ac.uk/WatsonWUI

[3] http://www.ontosearch.org

[4] http://xmlns.com/foaf/spec/

In case of ontologies with hundreds and thousands of concepts (SUMO[5]: 965 concepts, DBPedia[6]: 934 concepts ) it is nearly impossible for the human mind to overview the whole model. But this is essential to decide if a candidate ontology is really useful and whether it needs some customization.

## 2 Main Questions of the Thesis

Ontologies are semantic models with different expressiveness levels which are represented in RDF and OWL. A structure-based approach to analyze and modularize these semantic models need to tackle some issues during the development process. This section should provide a brief overview about open research questions which need to be solved during the development process. The first part of the research questions are derived from the ontologies itself and their properties, while the second part focuses on the ontology engineers which are addressed as the users of this framework.

### 2.1 Ontologies as Graphs

RDF allows to create structured information as triples following the form (Subject, Predicate, Object). The graph syntax of RDF allows to represent triples as graphs where the subjects and the objects are nodes and the predicates are directed edges (from subject to object). At this level the inherent semantic of OWL ontologies are not taken into consideration. Furthermore, the nodes and edges have different types, which are reflected in the labels (namespace and localname), which is a problem for standard Social Network Analysis approaches [15]. Additionally, it is not possible to organize the edges and nodes into disjunct sets, because a resource which is a subject or an object in one statement might be a predicate in another statement. This problem can be avoided if in contrast to the RDF graph syntax every named entity of the ontology is represented as a node (even the predicate is a node, which is connected with the subject and the object). But as the number of properties which are used as predicates is much less than the number of resources used as subject and objects, this graph representation would lead to a very different structure in which the properties are very central nodes with high degree values.

Some predicates as “hasLabel” or “hasComment” have an impact on the structural values. That is, their centrality values might be very high. It is very important to filter such concepts, which have an impact on the structural analysis but are not necessary to understand the content of an ontology. Furthermore, it is important to take different namespaces into consideration. It might be useful, to consider concepts from one namespace as a class of nodes and to analyze the connectedness of nodes from namespace to nodes of different namespaces.

Other open research questions derived from the graph representation of ontologies based on RDF and OWL are:

[5] http://www.ontologyportal.org/translations/SUMO.owl

[6] http://dbpedia.org/ontology

1. RDF specification allows to create blank nodes, which have an influence on the structure of the graph. How should these blank nodes be handled?
2. Ontologies allow reasoning which leads to a change in the structure of the ontology. Should these changes taken into account. That means, should a reasoning process executed before the structure-analysis process starts?
3. Ontologies might be expressed in different expressiveness levels (OWL Lite, OWL DL, OWL FULL). What is the impact of the expressiveness on the structure of an ontology?
4. Besides the schema ontologies represented in OWL may include instances. Is it important to take them into account? In which cases do they have to be considered and which cases not?

### 2.2 A Framework for Ontology Engineering

During the last years there is an increasing interest in the usability aspect of software products. For the success of a framework it is very important to take the addressed users' needs into account during the design process. Because this works addresses ontology engineers, their expectations of an ontology analysis and modularization framework have to be obtained and used as guidelines for the design process. This issue needs further investigation. At this point following questions have been identified:

1. How important are realtime and interactivity for ontology engineers?
2. What are the needs of ontology engineers, that have to be taken into account by developing an ontology analysis and modularization framework?
3. Ontology Engineering methodologies are mostly heavyweight. How can the framework be used in different methodologies?

## 3 General Approach

This work is based on the belief that the utilization of semantic models would improve the quality of information systems and would enable interoperability in distributed open systems as the Web. But creating ontologies from scratch as well as analyzing, reusing and maintaining existing ontologies are complex tasks which are hindering broad acceptance and application of ontologies. This is identified as the main problem, which this work tries to solve by developing and implementing methods as well as techniques to analyze and modularize ontologies. Regarding the definition "The design science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts" [13] the design science paradigm is apparently the most suitable research methodology for this work. According to [25] the process of design science research is a design cycle which comprise the following five subprocesses, which are used as a guidance for this research and to clarify the structure of this document.

1. Awareness of problem: Section 1 described the lack of appropriate techniques to analyze and modularize large ontologies, in order to simplify the reuse process. Efficient and flexible reusability in turn is seen as a key for the success of information integration based on ontologies. This problem is the main motivation for this research.
2. Suggestion: This work suggests to use structural information about ontologies to support the ontology analysis and modularization process in order to simplify ontology reusage. The problems of realizing this approach and the research question which have to be responded were presented in Section 2 while the proposed solution is discussed in Section 4 while .
3. Development: The current state of the design and implementation of an ontology analysis and modularization framework is presented in the second part of Section 4.
4. Evaluation. At this stage the evaluation process did not start yet. Therefore there will be a short presentation of the first ideas about this work's evaluation in Section 5.
5. Conclusion. Finally, Section 6 provides the conclusion and as this work is still in progress it describes the next steps.

## 4 Proposed Solution

Considering ontologies as networks of concepts connected through properties, network analysis techniques and using network measures (e.g. node centrality, betweenness, density, similarity) are a promising approach to analyze and modularize ontologies. As a very well established discipline in science there are a lot of sophisticated methods and tools for network analysis available. We believe that these methods can be modified, extended (in order to take the semantics into consideration) and applied to ontologies, so that the ontology structure can be used to analyze the content and to identify regions, which can be seen as network "communities" and can be extracted as modules. Furthermore, we are convinced that structure analysis enables a first evaluation of the usability by allowing different views, so that existing ontologies can be easier comprehended by ontology engineers. This is very important because refactoring and reusing of existing models assume that these models are understood.

The foundation of this work is the hypothesis, that analyzing and modularization of an ontology can be done in an efficient manner, by using structural information about the ontology. Some previously done related work have shown that this approach is promising. Structural analysis in [11] is motivated by the idea to measure the importance of a node in an RDF graph, without distinguishing between schema and data. For ranking the nodes the closeness centrality values are used. AKTiveRank [1] is a system which is motivated to facilitate reusing existing ontologies. It aims at improving ontology search engines by ranking ontologies based on structural properties of the search terms within the whole ontology. Four different measures are defined, which are calculated separately by ignoring the instances and the resulting values are merged.

In [15] Semantic Network Analysis (SemNA) is introduced to analyze ontologies for the purpose of reuse and re-engineering. Different notions of node centrality are used, namely degree centrality, betweenness centrality and eigenvector centrality. Analyzing the network structure of an ontology as a basis for partitioning the class hierarchy into disjoint and covering set of concepts is presented in [22]. Its main goal is to support distributed maintenance, selective reuse and efficient reasoning.

Therefore this work investigates on the application of network analysis techniques and network measures (e.g. node centrality, betweenness, density, similarity) to ontologies and aims at gaining insight to which extent structure based techniques can be modified so they are paying attention to the semantics inherent in ontologies. The expected contribution is a method and tool support for ontology engineers to analyze and modularize ontologies in a (semi-) automatic way. The main goal is to improve the usability and maintainability by increasing the understandability and allowing ontology engineers to refactor and reuse existing ontologies easily.

### 4.1 Current State of the Artifact

The current development of the artifact is at a very early state. As a very well known integrated development environment Eclipse allows to implement functional extensions through plugins. In this regard we have identified functional components which can be implemented as Eclipse plugins so an Ontology Modularization and Integration framework can be realized. Figure 2 illustrates the architecture of this framework.

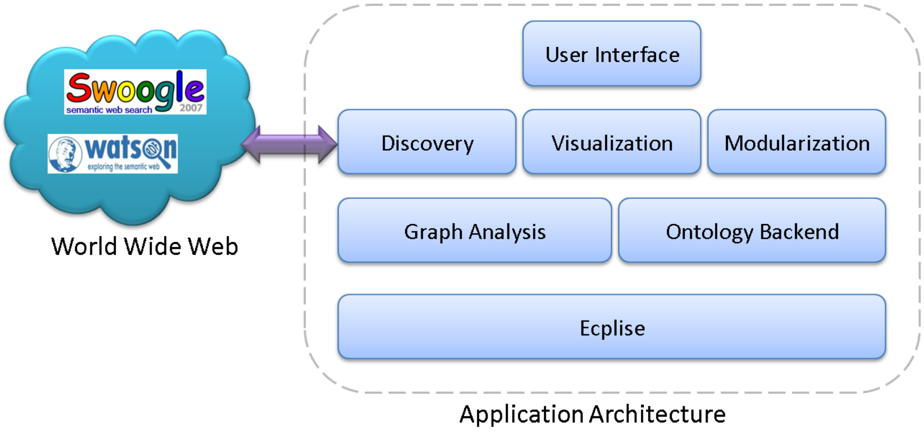
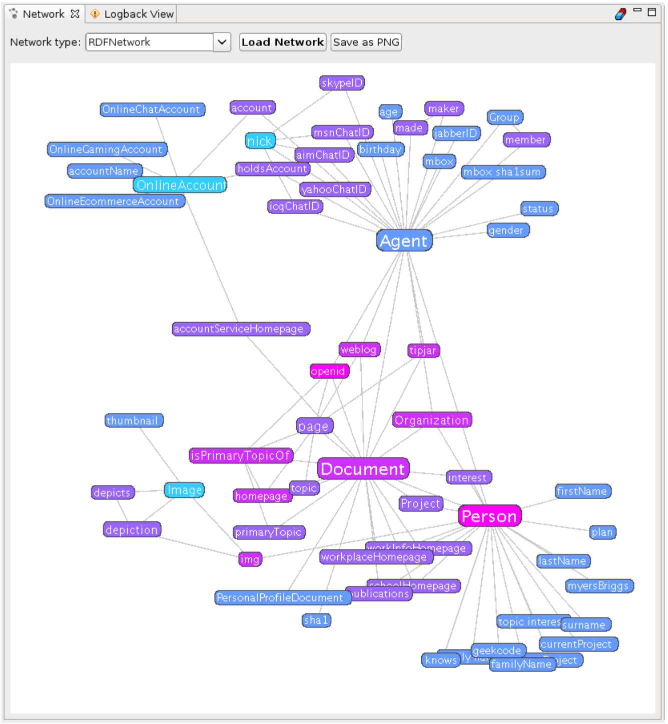


**Fig. 2.** Architecture of the Ontology Analysis and Modularization Framework

Based on the developed architecture, the decision was made to reuse the SONIVIS:Tool[7] to realize the targeted system. The SONIVIS:Tool is a network analysis software which is based upon Eclipse and allows easy extension through the Eclipse Plugin system. It provides already the Graph Analysis and the Visualization components and makes use of the Eclipse User Interface. Figure 3 illustrates the foaf vocabulary where the node size depends on the node degree.



Network type: RDFNetwork

**Fig. 3.** Structure visualization of the foaf vocabulary with node size depending on the node degree

The biggest nodes in the visualization are "Agent", "Document", "Person", "OnlineAccount", and "Organization". If these concepts are compared with the concept groups (especially the group names) from the specification as illustrated in Figure 1 it obvious that there is a similarity. The group names "Personal Info, "Documents and Images" and "Online Accounts"contain some of the concepts which have a high centrality in the structure. This first insight is an indication for the applicability and usability of the chosen approach and justifies further investigation.

[7] http://www.sonivis.org

## 5 Evaluation

Research activities always have to be validated in order to measure its quality. Design science can make use of different evaluation approaches to evaluate the outcome. The most popular approaches are case study, professional review, goal-free evaluation, and goal-based evaluation.

At this stage of this work it has not been clarified in detail how the outcomes can be evaluated. The first ideas are to evaluate the ontology analysis aspect through professional reviews of different ontology engineers. The important question is whether these ontology engineers gain new insight about their ontologies when they are using this framework. As their personal opinion cannot really be quantified and objectively compared it is still an open question, in which degree this is really applicable. For the modularization of ontologies it is intended to apply different ontology evaluation methods as [19] on the produced ontology modules to check their quality. For this approach case studies are necessary which are not found at this stage.

The goal-free evaluation is mainly a comparison activity of different solutions for the same problem based on some pre-defined criteria. Based on an in depth literature work about the state-of-the-art these criteria needs to be defined. In contrary, the goal-based evaluation focuses on the designed artifact itself. It gives a qualified view on the achievements of the proposed solution. The requirements which have been identified by the developer during the problem analysis process are used to evaluate to which extend they have been truly achieved. For this approach the problem to be solved have to be analyzed deeply and the requirements which have to be fulfilled by the artifact have to be formulated concretely.

## 6 Future Work

The vision of the Semantic Web brought new attention to ontologies by underlining its knowledge sharing aspect. Ontologies are considered as the most important means for information integration in the highly distributed and open Web. But either the reuse of existing ontologies or the matching of different ontologies is unavoidable for this integration. Therefore means for analyzing ontologies as well as modularization techniques for partial reuse are very important and a key for the success of information integration based on ontologies. Following the design science research methodology this work is grounded on the hypothesis, that analyzing and modularization of an ontology can be done in an efficient manner, by using structural information about the ontology.

As this work is still in progress the design and development process is ongoing and there are open research questions (see Section 2) which need further investigation. It is also expected that new questions will arise. Additionally, as mentioned in Section 5 it is still an open issue how this work is going to be evaluated. Use cases as well as criteria for goal-based and goal-free evaluation needs to found and defined.

**Acknowledgements** This work has been partially supported by the "InnoProfile-Corporate Semantic Web" project funded by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder - Entrepreneurial Regions.

## References

1. Harith Alani and Christopher Brewster. Metrics for ranking ontologies. In Denny Vrandečić, Mari del Carmen Suárez-Figueroa, Aldo Gangemi, and York Sure, editors, *Proceedings of the 4th International Workshop on Evaluation of Ontologies for the Web (EON2006) at the 15th International World Wide Web Conference (WWW 2006)*, pages 24–30, Edinburgh, Scotland, May 2006.
2. Sören Auer. The rapidowl methodology–towards agile knowledge engineering. In *WETICE*, pages 352–357. IEEE Computer Society, 2006.
3. A. Bernaras, I. Laresgoiti, and J. Correra. Building and Reusing Ontologies for Electrical Network Applications. In *ECAI96. 12th European conference on Artificial Intelligence*, pages 298–302. John Wiley & Sons, Ltd., 1996.
4. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, May 2001.
5. Elena Paslaru Bontas and Malgorzata Mochol. Towards a reuse-oriented methodology for ontology engineering. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering TKE 2005*, 2005.
6. Elena Paslaru Bontas, Malgorzata Mochol, and Robert Tolksdorf. Case studies on ontology reuse. In *Proceedings of the 5th International Conference on Knowledge Management*, 2005.
7. Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology Learning from Text: An Overview*, volume 123. IOS Press, 7 2005.
8. B. Chandrasekaran, John R. Josephson, and V. Richard Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14:20–26, 1999.
9. Philipp Cimiano, Aleksander Pivk, Lars Schmidt-Thieme, and Steffen Staab. Learning taxonomic relations from heterogeneous sources of evidence. In *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence. IOS Press, 2005.
10. Mariano Fernandez, Asuncion Gomez-Perez, and Natalia Juristo. Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering*, pages 33–40, Stanford, USA, March 1997.
11. Alvaro Graves, Sibel Adali, and Jim Hendler. A method to rank nodes in an rdf graph. In Christian Bizer and Anupam Joshi, editors, *International Semantic Web Conference (Posters & Demos)*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.
12. M. Grüninger and M. S. Fox. Methodology for the design and evaluation of ontologies. In *International Joint Conference on Artificial Inteligence (IJCAI95), Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.
13. A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–106, 2004.
14. Clyde W. Holsapple and K. D. Joshi. A collaborative approach to ontology design. *Commun. ACM*, 45(2):42–47, 2002.

15. Bettina Hoser, Andreas Hotho, Robert Jschke, Christoph Schmitz, and Gerd Stumme. Semantic network analysis of ontologies. In *Proceedings of the 3rd European Semantic Web Conference*, volume 4011 of *LNCS*, pages 514–529, Budva, Montenegro, June 2006. Springer.
16. Mustafa Jarrar and Robert Meersman. Formal ontology engineering in the dogma approach. In Robert Meersman and Zahir Tari, editors, *CoopIS/DOA/ODBASE*, volume 2519 of *Lecture Notes in Computer Science*, pages 1238–1254. Springer, 2002.
17. Konstantinos Kotis and A. Vouros. Human-centered ontology engineering: The hcome methodology. *Knowledge and Information Systems*, 10(1):109–131, July 2006.
18. D.B. Lenat and R.V. Guha. Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. 1990.
19. HUANG Ning and DIAO Shihan. Structure-based ontology evaluation. In *IEEE International Conference on e-Business Engineering, 2006. ICEBE '06*, pages 132–137, 2006.
20. Helena Sofia Pinto, Steffen Staab, and Cristoph Tempich. Diligent: Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, Valencia, Spain, 2004.
21. Francesco Sclano and Paola Velardi. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*, Funchal (Madeira Island), Portugal, March 2007.
22. Heiner Struckenschmidt. Network analysis as a basis for partitioning class hierarchies. In *Workshop on Semantic Network Analysis, ISWC*, 2006.
23. Mari Carmen Suarez-Figueroa and Asuncion Gomez-Perez. Neon methodology for building ontology networks: a scenario-based methodology. In *Proceedings of the International Conference on SOFTWARE, SERVICES & SEMANTIC TECHNOLOGIES*, 2009.
24. York Sure, Steffen Staab, and Rudi Studer. On-to-knowledge methodology (otkm). In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies: International Handbook on Information Systems*, pages 117–132. Springer, 2004.
25. H. Takeda, P. Veerkamp, T. Tomiyama, and H. Yoshikawam. Modeling design processes. *AI Magazine*, 11(4):37–48, 1990.
26. M. Uschold and M. King. Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, Montreal, Canada, 1995.