V. Kůrková et al. (Eds.): ITAT 2014 with selected papers from Znalosti 2014, CEUR Workshop Proceedings Vol. 1214, pp. 93–96
http://ceur-ws.org/Vol-1214, Series ISSN 1613-0073, © 2014 O. Zamazal, V. Svátek

Towards Automation of Ontology Analysis Reporting

Ondřej Zamazal, Vojtěch Svátek

Department of Knowledge and Information Engineering
Faculty of Informatics and Statistics
University of Economics, Prague, Czech Republic
ondrej.zamazal@vse.cz, svatek@vse.cz

Abstract: Different kinds of ontologies are currently accessible either from different ontology catalogues or various ontology search engines. Heterogeneity of ontologies and ontology resources hinders ontology users in their work such as selection of an adequate ontology resource in which they could search for proper ontology to be used, reused or adapted with regard to their use case. Although there are many works which provided ontology analyses from diverse aspects, none of them enables straightforward access at any time. In this paper we present preliminary results of our ontology analysis and our plan to provide an automatic and generally available ontology analysis reporting service providing time snapshots of available ontologies. Further, we will present some available ontology catalogues, repositories, search engines and discuss how they could be included. In our work we focus on ontologies expressed in OWL and we address logical, structural, naming and annotation aspects of ontologies.

1 Introduction

Ontologies, as formal conceptual models typically describing a certain domain of discourse, are inherent part of the Semantic Web vision already since its inception in 2001 [3]. As the Semantic Web was evolving, typical semantic web ontologies were changing: from large ontologies carefully designed by AI experts and highly reused within the community (nineties), e.g. GALEN[1] via smaller web domain ontologies designed by many individuals rarely used or reused (since 2000) to small simple domain ontologies driven mostly by data modelling request, e.g. FOAF vocabulary.[2] Many of these different kinds of ontologies are currently accessible on the Semantic Web either via ontology catalogues or via ontology search engines. We jointly call them *ontology resources*.

Since there are many different ontology resources varying in typical ontologies they offer, ontology users face difficult situation of selecting an adequate ontology resource in which they will search for proper ontology to be used, reused or adapted with regard to their use case. To support ontology users, many works provided ontology analyses from various aspects, see Section 2. In a nutshell, these works differ in various aspects from which they analyse ontologies, but they all share their static nature, i.e. an analysis was done once at a certain point of time. Thus, ontology analysis research lacks its automation and continuous access. While an interpretation of statistics gathered during ontology analysis and its subsequent lessons learned can hardly be an automatic process, automatically providing summary overviews e.g. in tables with regard to diverse ontology aspects is obviously doable.

Our long-term goal is to provide an automatic ontology analysis reporting service in order to facilitate regular and up-to-date snapshots of ontology repositories. We include ontologies expressed in OWL[3] addressing four aspects:

- *logical* including types of entities, axioms, constructs and expressiveness,
- *structural* corresponding to the ontology graph based on asserted (rather than inferred) axioms and especially the taxonomy structure,
- *naming* dealing with entity naming. Naming in an ontology is mostly related to local parts of entity URIs and labels (values of rdfs:label property). Although ontologies are logical theories, entity naming is considered as an important part of an ontology [8].
- *annotation* which can contain important additional information written in textual or structural form.

The rest of the paper is structured as follows. Section 2 gives an overviews about ontology analysis related work. Section 3 presents some ontology repositories and ontology search engines considered so far. Next, we present envisioned ontology analysis process in Section 4. Section 5 provides preliminary results of some characteristics of ontology repositories and concisely overviews further interesting ontology analysis characteristics. Finally, Section 6 wraps up the paper and foresees some future work.

2 Related Work

Ontology analysis has already been performed many times, from different perspectives and in different ranges. Ding and Finin [5] evaluated 1.7M *RDF documents*[4] in order to better understand the status quo of the semantic web to date. They employed a number of metrics and usage

[1] http://www.opengalen.org/

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

[3] http://www.w3.org/TR/owl2-primer/

[4] http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

patterns, such as aggregation over URL domains and individual websites, or number of triples used to define a term. Wang et al. [15] analysed a collection of *ontologies* (688 OWL ontologies and 587 RDF schemas) from the logical and structural viewpoint: the shape of the class hierarchy (lists, trees or multitrees), proportion of certain OWL language constructs or logical expressiveness. The statistics counted for diverse metrics allowed them to characterize the semantic web from the semantic documents/terms perspective. Matentzoglu et al. [7] gathered crawl-based OWL corpus (about 4500 ontologies) and compared it with 4 ontology repositories or samples from ontology search engines (the BioPortal,[5] the Oxford,[6] the Swoogle[7] and the TONES[8]) regarding basic characteristics such as number of different kinds of entities, number of various axiom types, distribution of OWL profiles etc. They concluded that crawl-based OWL corpus is close to curated repositories in terms of ontology size and expressivity. Their process of gathering ontologies includes a careful filtering procedure to ensure collecting real single OWL ontologies rather than arbitrary OWL files.

Vocabularies as lightweight ontologies have been inspected in many surveys. Suominen and Hyvönen [11] validated *SKOS vocabularies*[9] against (SKOS-specific) quality measures and a tool (Skosify) was provided to correct some reported errors. The landscape of SKOS vocabularies was also inspected by Manaf et al. [6], where the focus was on high-level structural properties such as the number of hierarchy levels or in- and outgoing links to other entities. Large number of vocabularies (almost three thousands) have been analyzed by Cheng [4], specifically focusing on the mutual relatedness of web vocabularies from the semantic, lexical, expressiveness and distribution perspective. The high number of vocabularies involved was however achieved by gathering them in a bottom-up manner, via extracting new vocabularies from RDF documents. Entities from diverse RDF documents (almost 16 million, from the *Falcons search engine*[10]) were grouped based on their common namespace. In order to measure the vocabularies' relatedness, their instances were also taken into account.

Besides ontologies and vocabularies, *linked datasets* are also being inspected. A tool for extensive analysis of linked data sets, *LODStats*, by Auer et al. gathers comprehensive statistics about RDF datasets. Statistics are available either from web LODStats web-page[11] or they can be accessed using SPARQL endpoint.[12]

Other projects directly connected their ontology analysis with practical applications. While RDFS schemas have been analyzed by Theoharis et al. [13] in order to create a benchmark for semantic web tools, e.g. query language interpreters, Rosoiu et al. [9] concentrated on an analysis of OWL ontologies in order to generate a suitable benchmark for ontology alignment. Next, Tempich and Volz [12] analyzed ontologies from the DAML ontology library in order to tune parameters for generation of synthetic ontologies suitable for performance evaluation of semantic web reasoners. They used a clustering approach for discovering structurally similar ontologies. Each ontology type is then represented as a synthetic ontology.

Last but not least, our work is strongly related to Ontology Evaluation, which focuses on assessing the quality of a single ontology. According to Vrandecic [14], an ontology can be evaluated from several aspects. The vocabulary aspect is dealing with evaluating names used in the ontology. The syntax aspect includes quality issues related to ontology serialization in its surface syntax, where many trivial best practices should be fulfilled (e.g., terminological axioms should precede facts), along with syntax validation. The structure aspect deals with the surface structure of axioms and their constituent constructs. For the last aspect the number of proposed and implemented metrics is highest, since it can be effectively measured through common graph metrics and returns easily understandable numbers. The semantics aspect evaluates an ontology considering the inferential semantics of OWL. Thus, semantic metrics measure the models (and their entailments) that are described by the structure.

Our ongoing work on ontology analysis reporting is distinguished from ontology analyses works, among other, by: 1) inclusion of the naming and annotation aspect of ontologies; 2) continuous provision of fresh results (given the dynamic state of the subject analysed) in large scale and 3) user access to all data and time snapshots of OWL ontologies via web interface.

3 Ontology Resources

Six prominent *ontology resources*, on which we base our research, are as follows.

The *BioPortal* [16] is a web portal providing access to a library of well-curated biomedical ontologies via RESTful services. Currently, there are 417 ontologies in different formats. The BioPortal contains ontologies from another ontology repository, the OBO foundry.[13] The primary format of the OBO foundry is OBO format but the ontologies are also available in OWL. Ontologies in BioPortal vary a lot in terms of number of entities (from couple of entities to tens of thousands entities) or complexity. We can find there ontologies of complexity lower than OWL-Lite as well as ontologies with complexity of OWL 2.

LOV[14] is a well-curated collection of linked open vocabularies used in the Linked Data Cloud. To date there

[5] http://bioportal.bioontology.org/

[6] http://www.cs.ox.ac.uk/isg/ontologies/

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

[8] http://owl.cs.manchester.ac.uk/repository/

[9] http://www.w3.org/2009/08/skos-reference/skos.html

[10] http://iws.seu.edu.cn/services/falcons/

[11] http://stats.lod2.eu/stats

[12] http://stats.lod2.eu/sparql

[13] http://obofoundry.org/

[14] http://lov.okfn.org/dataset/lov/

are 409 ontologies covering diverse domains, e.g., publications, science, business or city. Most ontologies are structurally simple, i.e. they often have complexity lower than OWL-Lite, and there are usually small; yet, they are used within diverse linked open data applications. Aside a list of available ontologies, there is also a SPARQL endpoint for accessing the ontologies' metadata.

The *Protégé*[15] ontology library mostly contains ontologies developed within the Protégé editor. As there is no programmatic access to the library nor a concise list of available ontologies, links to OWL files must be extracted using some tailored wrapper. Currently, it has 98 ontologies which also includes well-known test ontologies, e.g. Pizza ontology. This repository has rather small ontologies (up to hundreds of entities) and their complexity mostly correspond to OWL DL.

The *TONES* repository contains ontologies of various domains, many of them however designed for testing purposes. Similarly as Protégé library, it has no direct programmatic access nor a list of available ontologies except the HTML page.[16] Currently, it has 174 ontologies including OBO ontologies (already present in BioPortal). Some of the ontologies are large (over 1000 entities) and most have the complexity of OWL-Lite or OWL-DL.

Besides ontology repositories there are also search engines. The *Swoogle* semantic web search engine extracts metadata for documents of specific filetypes (rdf or owl) and computes the relations among them. Nowadays, Swoogle indexes almost 4 million semantic documents and allows to search for ontologies and their instances over this index. This engine does not provide a public API.

Watson [2] is a semantic web search engine[17] for ontologies and semantic documents. There are about 20,000 cached ontologies. Watson provides keyword search and a number of methods for manipulating with ontologies, e.g., basic metrics as number of classes etc.

In all, BioPortal, LOV and Watson provide programmatic access to ontologies; processing Swoogle output is restricted by the service [7]; ontologies in Protégé and TONES can be accessed using a tailored wrapper.

4 Ontology Analysis Reporting

We plan to make *ontology analysis reporting service* (see Figure 1) available via web interface where on the one side web users could ask for the latest summaries (automatic reports) of particular ontology repositories ("(a) retrieve summary") and on the other side they could ask for particular ontologies or ontologies meeting certain criteria ("(b) retrieve ontologies").

In order to provide such services independently on availability of ontology resources or ontologies, we materialize all ontologies into a database ("(1) materialization") as a central point of the software architecture design. The database is populated with ontologies from ontology resources using their programmable access or tailored wrapper. Imported ontologies are also stored into the database [7]. The materialization of ontologies in the database processes and decomposes ontologies into their parts: entities, names (local fragment of entity URIs), relations (axioms), imported ontologies etc. Next, various ontology metrics are computed and their results stored into the database as well ("(2) metrics computation"). Since ontology resources can include the same ontologies, deduplication process follows ("(3) deduplication"). Deduplication process could be based on entity to entity comparison. However, this would be computational very demanding. Therefore, we first search for duplicates candidates based on computed metrics such as number of classes, object properties etc.[18] and then we can apply detail (e.g. entity to entity) comparison on duplicates candidates. This approach tends to be highly precise, since we do not exclude false duplicates. We think that ontology versions, being reflected as slight variants according to computed metrics, should be kept and analyzed as different ontologies. Summarizing results ("(4) summarization") of ontology metrics uses *R language for statistical computing*.[19]

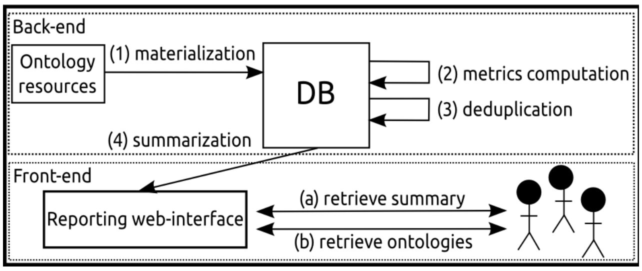

Figure 1: Ontology Analysis Reporting Architecture

We implement our ontology analysis workflow (so far partly back-end components from Figure 1) via Java programs.[20] We manipulate the ontologies via OWL-API[21] and decompose them into a MySQL database.

5 Ontology Analysis Characteristics

In our work we consider ontology metrics related to four aspects of ontologies inspired by related work in Section 2. Logical metrics represent basic characteristics in terms of a number of classes, complex classes (defined by anonymous expressions), properties, instances, axioms and annotations. We provide these characteristics[22] in Table 1 where average, median and maximum values are

[15] http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library

[16] The link to download all ontologies does not work. [June 2014]

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

[18] Apropriate set of metrics to be used for deduplication will be tuned based on further testing.

[19] http://www.r-project.org/

[20] We plan to make all programs freely available.

[21] http://owlapi.sourceforge.net/

[22] As March 2014 snapshot.

Metrics		BioPortal	LOV	Protégé	TONES	Watson
Classes	Avg	270	30	126	153	79
	Med	223	14	34	72	25
	Max	994	509	717	948	986
Complex classes	Avg	191	22	163	126	76
	Med	50	4	27	28	7
	Max	5194	659	603	2752	1042
Object properties	Avg	34	23	46	21	14
	Med	8	12	14	5	8
	Max	1337	288	313	414	291
Data properties	Avg	10	11	12	17	9
	Med	0	3	7	0	2
	Max	488	217	74	708	159
Instances	Avg	88	17	355	100	54
	Med	0	2	18	0	2
	Max	5819	702	2872	3542	2961
Axioms	Avg	2634	543	3103	1588	941
	Med	1635	216	442	318	231
	Max	38056	23839	21087	44764	19361
Annotations	Avg	1213	228	330	204	446
	Med	617	92	8	3	42
	Max	16058	2900	2937	4260	13764

Table 1: Characteristics of ontology resources

given. Due to various issues (e.g. non-parseable ontologies by OWL-API or unaccessible imports), we could not automatically retrieve all ontologies from the ontology resources. Thus, we analyzed 171 ontologies from BioPortal, 302 from LOV, 20 from Protégé and 122 ontologies from TONES. For this first run of ontology analysis we restricted the process to ontologies with more than 0 and less than 1001 classes. From Table 1 we can see that on average BioPortal has large ontologies in terms of number of classes, axioms and annotations. On the other side, LOV has, on average, very small ontologies in terms of all listed metrics in the table except annotations and axioms. While Protégé, TONES and Watson are on average comparable in terms of number of classes and properties, Protégé has typically ontologies with many more instances than are in any other ontology resource.

In our ongoing work, we will also consider ontologies from the structural viewpoint, e.g., the number of top classes and leaf classes, or the maximum number of superclasses/subclasses. Next, we plan to provide ontology analysis for the naming aspect, i.e. the use of concatenation symbols, capitalization and complex analysis aimed at naming patterns [10]. Finally, we plan to inspect ontologies with regard to annotations, i.e. which types of annotations dominate in each ontology resource.

6 Conclusions and Future Work

Our ongoing work aims at an ontology analysis reporting service. We described the ontology repositories to be involved, provided a sketch of ontology analysis reporting architecture, and presented the preliminary results of logical characteristics for five ontology resources, along with mentioning ontology metrics to be further considered.

In future we will implement all the ontology metrics mentioned and apply them on, at least, the six mentioned ontology resources. We also plan to provide a reporting service as a *web interface* enabling access to all ontologies and their respective characteristics as well as characteristics of each resource and across all ontologies in future. We also plan to proceed from elementary features to semi-automatic discovery of (intentional and implicit) *patterns*.

Acknowledgement Ondřej Zamazal has been supported by the CSF grant no. 14-14076P, “COSOL – Categorization of Ontologies in Support of Ontology Life Cycle”.

References

[1] Auer S., Demter J., Martin M., Lehmann J.: LODStats - An Extensible Framework for High-performance Dataset Analytics. In: Proc. of the EKAW 2012. 2012.

[2] d’Aquin M., Baldassarre C., Gridinoc L., Angeletou S., Sabou M., Motta E.: Characterizing Knowledge on the Semantic Web with Watson. In: EON Workshop at ISWC’07, Busan, Korea. 2007

[3] Berners-Lee, T., Hendler J., Lassila O.: The semantic web. *Scientific american* 284.5 (2001): 28-37.

[4] Cheng G.: Relatedness between Vocabularies on the Web of Data: A Taxonomy and an Empirical Study. *Journal of Web Semantics*. 2013.

[5] Ding L., Finin, T.: Characterizing the Semantic Web on the Web. In: ISWC 2006.

[6] Manaf N. A. A., Bechhofer, S., Stevens, R.: The Current State of SKOS Vocabularies on the Web. In: ESWC 2012. 2012.

[7] Matentzoglu, N., Bail, S., Parsia, B.: A Snapshot of the OWL Web. In: Proc. ISWC 2013, Springer, LNCS, 2013.

[8] Nirenburg S., Wilks Y.: Whats in a symbol: Ontology and the surface of language. *Journal of Experimental and Theoretical AI*, 2001.

[9] Rosoiu M. E., Trojahn C., Euzenat J.: Ontology Matching Benchmarks: Generation and Evaluation. In: Ontology Matching Workshop 2011.

[10] Šváb-Zamazal O., Svátek V.: Analysing Ontological Structures through Name Pattern Tracking. In: EKAW-2008, Acitrezza, Italy, 2008.

[11] Suominen O., Hyvönen E.: Improving the Quality of SKOS Vocabularies with Skosify. In:EKAW 2012.

[12] Tempich C., Volz R.: Towards a Benchmark for Semantic Web Reasoners - An Analysis of the DAML Ontology Library. In: Evaluation of Ontology-based Tools (EON). 2003.

[13] Theoharis Y., Tzitzikas Y., Kotzinos D., Christophides V.: On Graph Features of Semantic Web Schemas. Knowledge and Data Engineering, IEEE Transactions on, 20(5), 692-702. 2008.

[14] Vrandecic, D.: Ontology Evaluation. Ph.D. Thesis. Karlsruhe. 2010.

[15] Wang T. D., Parsia B., Hendler J.: A Survey of the Web Ontology Landscape. In: ISWC-2006.

[16] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., Musen, M. A.: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research, 39(suppl 2), W541-W545.