Automated Exploration of Ontology Repositories

Ondřej Zamazal and Vojtěch Svátek

University of Economics, W. Churchill Sq.4, 130 67 Prague 3, Czech Republic
{ondrej.zamazal|svatek}@vse.cz

Motivation The choice of adequate ontology repository is an important prerequisite to finding an ontology to be reused or adapted for a concrete use case. As the repositories are mostly affiliated to particular communities within the semantic web, understanding the typical features of ontologies in each of them is also helpful for designers of ontology management tools.

Overall Process, Metrics and Results Our ontology exploration process includes ontology *collection*, *materialization* and then *metrics computation*; finally, the resulting metrics are *explored* using the *R language*[1] to automatically get a *summary report* in the form of tables. To automate the collection phase, we partly employed *Ontohub*,[2] which is an open ontology repository mirroring several other repositories. The materialization includes ontology storing (into the database) in order to decompose them into entities, names, relations, imported ontologies and head nouns. We use the OWL-API[3] to manipulate the ontologies.

We considered metrics related to four aspects of ontologies.[4] *Logical* and *structural* metrics include, e.g., the numbers of different types of entities and axioms. We also categorize the complexity of ontologies into bins (as in [2, 3]). The *naming* aspect reflects some basic information regarding the length of class name (local fragments of URI or labels), capitalization and usage of concatenation symbol/technique, i.e. a hyphen, underscore, camel-case or dot (as in [1]). For the *annotation* aspect we compute the proportions of RDFS annotations.

We explored ontologies from five prominent ontology repositories (Table 1 contains just a few selected metrics). Due to parsing problems, unavailability of ontologies or their imports we however did not collect all ontologies from the repository. *BioPortal*[5] is a web portal providing access to a library of well-curated biomedical ontologies via REST-ful services. It contains ontologies from another ontology repository, the OBO Foundry.[6] We collected ontologies using the Ontohub mirror where only ontologies with size below 5MB (thus only 342 of

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

[2] https://ontohub.org/

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

[4] Due to the space limitation full list of metrics and complete results are at the supplementary web page: http://owl.vse.cz:8080/MetricsExploration/

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

[6] http://obofoundry.org/

Metrics *(June 2014 snapshot)*		BioPortal	Dumontier	LOV	Protégé	TONES
Ontologies processed		254	70	353	41	183
Percentage of all		74%	95%	83%	44%	88%
Complex class using existential restr.	Avg	**57%**	28%	*7%*	14%	43%
Complex class as superclass	Avg	**74%**	*35%*	69%	57%	67%
Branching	Avg	**0.88**	*0.55*	0.48	0.61	0.79
	Max	**2.39**	*1.09*	1.77	1.43	1.78
Multiple inheritance	Avg	**32**	*0*	1	6	12
	Max	1877	*254*	321	497	**24800**
Annotation as label	Avg	38%	**51%**	32%	*13%*	38%
Annotation as comment	Avg	*1%*	37%	25%	**49%**	37%
Camel technique	Avg	*15%*	**61%**	39%	28%	29%
Underscore technique	Avg	**54%**	*0%*	*0%*	23%	36%

Table 1. Selected metrics. Average (Avg) is either mean or median, according to better representativeness. The min statistics is omitted since it is always zero. The max statistics is omitted for ratios because it is nearly always 100%. The larges value across all repositories is in bold.

the total) are available.[7] The *Dumontier* lab ontologies[8] are biological ontologies aimed at knowledge representation and reasoning. Their ontologies are quite interconnected (many mutual imports). *LOV*[9] is a well-curated collection of linked open vocabularies used in the Linked Data Cloud. The *Protégé* ontology library mostly contains ontologies developed within the Protégé editor. As there is no programmatic access to the library, we manually downloaded them. It turned up that out of 93 ontologies (except Dumontier ontologies on which there is also a link) 43% ontologies were not available. Finally, the *TONES* repository (using its Ontohub mirror of 207 ontologies - collected 88%) contains ontologies of various domains, many of them however designed for testing purposes.

Future Work We plan to run such an analysis repeatedly, include more repositories (preferably via Ontohub) and more metrics. We also want to keep the ontology exploration services available via a web interface[10] where the users could ask, on the one hand, for the latest summaries of particular repositories, and on the other hand for particular ontologies or ontologies meeting some criteria.

Ondřej Zamazal has been supported by the CSF grant no. 14-14076P and this research is also supported by UEP IGA project F4/34/2014 (IG407024).

References

1. Manaf N. A. A. M., Bechhofer S., Stevens R.: A Survey of Identifiers and Labels in OWL Ontologies. In: OWLED-2010.
2. Matentzoglu, N., Bail, S., Parsia, B.: A Snapshot of the OWL Web. In: ISWC 2013.
3. Wang T. D., Parsia B., Hendler J.: A Survey of the Web Ontology Landscape. In: ISWC-2006.

[7] To overcome 5MB limitation we gathered BioPortal ontologies directly by RESTful services. Corresponding ontology metrics are available via the supplementary web.

[8] `http://dumontierlab.com/?page=ontologies`

[9] `http://lov.okfn.org/dataset/lov/`

[10] A sample service, providing metrics for a given ontology, is already available from `http://owl.vse.cz:8080/MetricsExploration/`.