OOSP: Ontological Benchmarks Made on the Fly

Ondřej Zamazal and Vojtěch Svátek

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

Abstract. The demo paper presents OOSP (Online Ontology Set Picker), a tool allowing to select, from major repositories, a set of ontologies that satisfy a user-defined sets of metrics. Its main purpose is allowing ontological tool designers to rapidly build custom benchmarks on which they could test different features. It could also serve for usage studies of different ontology language constructs and for pattern spotting. The web front-end allows to specify a broad range of metrics and delivers benchmarks along with their statistics of metrics, including a graph view.

1 Introduction and Motivation

The number of ontologies on the semantic web is steadily growing, and new tools for their management and exploitation are being built all the time. Their functionality needs to be tested on ontology collections allowing to balance between 1) sufficient coverage of different cases the tool might encounter, and, 2) presence of specific features crucial for a particular functionality. With respect to the latter, for instance, ontology repair tools [4, 1] can only be assessed on models with non-trivial concept expressions; similarly, thoroughly testing ontology visualization techniques [2] demands diverse ontology aspects such as large taxonomies, instances or various types of axioms; furthermore, reasoners often concentrate on certain OWL2 ontology profiles, such as EL [7]. Beyond the benchmarking scenario, usage analysis of different constructs can also be helpful when analyzing empirical modeling patterns [3] and devising best-practice patterns for the respective modeling problems. Since preferences regarding language constructs (and similar kinds of restrictions) are not met by existing ontology search/picking tools, such as Swoogle[1] or Watson,[2] we decided to develop a simple web-based tool called *Online Ontology Set Picker* (OOSP), leveraging on our previous work related to analysis of ontology repositories [8, 9] and offering a broad range of ontology selection criteria/metrics.

The demo paper is structured as follows. Section 2 describes the OOSP internals including the source repositories. Section 3 introduces its front-end. Section 4 provides comparison to related work. Finally, Section 5 wraps up the paper with conclusions and future work.

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

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

2 OOSP: Sources and Internals

At the source level, OOSP currently relies on two prominent ontology repositories, BioPortal and LOV. The content of the ontologies is processed using the OWL-API.[3]

BioPortal[4] is a library of well-curated biomedical ontologies. Currently (February 2015 snapshot) there are 420 ontologies in different formats, including some adapted from another repository, the OBO foundry.[5] Out of the 420 ontologies 36 were not available due to 'not found' error or private access. Further 9 ontologies pointed to zip archives, which we currently do not process. Finally, out of the 375 available BioPortal ontologies, our Feb. 2015 snapshot contains 317 (85%), since 36 ontologies were not processable due to unavailable imports and 22 due to parsing problems using OWL-API. To access the BioPortal ontologies we used RESTful services.

LOV[6] is a well-curated collection of linked open vocabularies used in the Linked Data Cloud. To date (Feb. 2015 snapshot) there are 475 ontologies covering diverse domains, e.g., publications, science, business or city. The ontologies/vocabularies are usually small and they are used within diverse linked open data applications. Out of the 475 ontologies 2 were not parseable by OWL-API and 12 ontologies were not processable due to unavailable imports. In all, our Feb. 2015 snapshot contains 461 LOV ontologies (97%). To access the LOV ontologies dump we used its SPARQL endpoint. This dump however does not contain all imported ontologies. In all, OOSP now enables an access to 778 ontologies.

The considered ontology metrics (82)[7] are divided into 7 groups covering the most important ontology aspects. *Entity* metrics (9) include numbers of entities (e.g., classes, instances); *axiom* metrics (27) include numbers of different axiom types (e.g., subsumption, equivalence); *class expression type* metrics (11) include expression types used for construction of anonymous classes (e.g., existential quantification); *taxonomy* metrics (9) include characteristics of taxonomy (e.g., the number of top classes, leaf classes, branching degree, maximum taxonomy depth); *OWL2 profiles and reasoning* metrics (7) include profile information along with information about consistency and number of unsatisfiable classes;[8] *annotation* metrics (6) include counts of selected annotation types (e.g., labels, comments) and of different languages involved in label annotations; finally, *detail* metrics (13) include some newly designed metrics related to domain/range (e.g., number of anonymous classes as domain definition).

Similarly as in [5] we computed the overlap between the collections, considering two ontologies as similar if their overlap, as the ratio of their signature intersection and signature union, is at least 90%. Based on this, the BioPortal

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

[4] `http://bioportal.bioontology.org/`

[5] `http://obofoundry.org/`

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

[7] We state a number of metrics in brackets.

[8] We applied the HermiT reasoner: `http://hermit-reasoner.com/`.

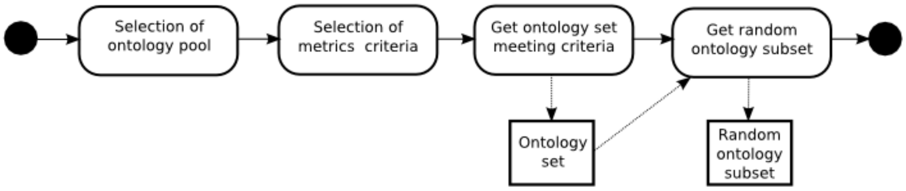
Selection of ontology pool
Selection of metrics criteria
Get ontology set meeting criteria
Get random ontology subset
Ontology set
Random ontology subset

Fig. 1. OOSP workflow. Rounded-shaped rectangles depict steps of the workflow and squares depict output of the steps.

and LOV snapshots only share 5 ontologies (e.g., the *biotop* ontology). There are 4 overlapping ontologies within LOV. These overlapping ontologies differ in annotation properties which are not currently considered in our overlap computation. Due to very large ontologies in BioPortal we were not able to compute its overlap.

3 OOSP: Front-end

OOSP, available at `http://owl.vse.cz:8080/OOSP/`, is a web-based application implemented using Java Servlet Pages, JavaScript and OWL-API. Ontologies with their imports are stored on disk and ontology metric criteria values and statistics are stored in a MySQL database.

OOSP follows a four-step workflow depicted in Figure 1. First, the initial *ontology pool* is selected; for now, it can be BioPortal, LOV, or their union. Second, the user can browse through the seven ontology *metric* types and specify values (*max* and/or *min*, except nominal values such as OWL profiles) for individual metrics. Currently, the combination of all selected restrictions has the semantics of *conjunction* (we plan to add more flexibility in future). To make the restriction setting more informed, six statistics are provided: the ratio of ontologies having at least one occurrence of the object aggregated (via count, average or max) by the metric;[9] the ratio of ontologies for which respective metrics is unknown (*N/A*);[10] and descriptive statistics (median, average, standard deviation and maximum) of the metric over all ontologies.[11] Third, the user obtains the *ontology set* meeting the provided restrictions, and fourth, OOSP can *randomly select* a subset of it, with required cardinality. For both, restricted ontology set and its randomly selected subset, OOSP provides a table containing all metrics values for all selected ontologies. An ontology from the set can be downloaded in three ways: one *separate* ontology as OWL file, one ontology with *all ontologies from its import closure* as zip archive, or *ontology merged with its import closure* as one OWL file. There are further three ontology-set-wise download options: only the table (in CSV); ontology set summary descriptive statistics (also in

[9] For binary metrics such as OWL profiles, it is simply the ratio of positive values.

[10] E.g., the reasoner could not process some ontologies due to unsupported datatypes.

[11] We omit minimum since it is usually zero.

CSV); and actual ontologies as OWL files (ZIP archive). Finally, for selected eight metrics (classes/instances counts, axiom types, DL constructs, OWL 2 profiles, annotations, domain/range definition types) OOSP also offers graphs of ontology set statistics.

4 Related Work

Ontology repositories provide various search options to access their ontologies. The Watson search engine allows to search ontologies using keywords. Via its Java API Watson also provides a SPARQL endpoint along with some precomputed metrics: concept coverage, DL expressivity, representation language (e.g., RDFS), numbers of classes, properties, individuals and statements. BioPortal provides a term-based search for classes and properties in ontologies, where one can further restrict the ontology category (e.g., anatomy). BioPortal RESTful services offer several count-based metrics per ontology e.g. number of classes, properties LOV also provides a RESTful service for a term-based search over ontologies or terms, and a SPARQL endpoint. Other ontology repositories solely provide collection of ontologies without rich metadata (e.g., the Oxford Ontology Library, Protege Ontology Library, or Ontohub).

The most relevant is the work by Matentzoglu et al. [6], where the Manchester OWL repository[12] is presented. It contains a crawl-based created Manchester OWL Corpus (MOWLCorp) presented in [5], a snapshot of BioPortal and Oxford Ontology Library. The goal of this repository, similarly to ours, is to create and share ontology datasets. It provides access to five pre-constructed datasets and an experimental REST-based web service that should allow users to create a custom dataset. Authors in [6] also mentioned an experimental data set creator allowing users to create custom datasets based on a wide range of metrics. However, on the respective web-page[13] there is only available offline generation of custom datasets where a user can specify his/her requirements: ontology pool, import handling, OWL2 profiles and special wishes specified in a HTML form, while the custom dataset is to be generated offline by the portal maintainers.

In comparison, our work focuses on the web-based front-end allowing to build an experimental ontology set useful as benchmark for ontology tool developers and ontology experimenters. Therefore, we do not precompile any ontology set collection but we rather provide a broad range of metrics that can work as on-the-fly restrictions. Next, besides BioPortal repository we also considered LOV repository since it is also a well-curated source and it also contains ontologies broadly used. To cover potentially many various use cases we also provide extra metrics types such as taxonomy, annotation and detail metrics. Further, we put more emphasis on different types of additional downloads: besides actual ontologies (and optionally their imports) it is also possible to download a table with ontology metric values and summary statistics, plus associated graphs. We cannot provide more practical comparison between the end-usage of OOSP and

[12] `http://mowlrepo.cs.manchester.ac.uk/`

[13] `http://mowlrepo.cs.manchester.ac.uk/generate-custom-dataset/`

the Manchester OWL repository, since the latter does not have a front-end for on-the-fly ontology dataset generation available.

5 Conclusions and Future Work

This paper presents OOSP, a web-based tool allowing ontology developers and experimenters to create a benchmark ontology set based on selected metrics, from two curated ontology repositories. Different ontology metric types can be useful for various use cases (e.g., benchmarking ontology repair tools, ontology visualization tools, reasoners, or tracking frequent patterns in ontologies). OOSP provides detailed metrics values for each ontology from the selected set as well as overall summary statistics, including a graph form.

The current version of OOSP allows to reproduce each ontology set selection (by setting the same initial pool and metrics restrictions) by other users; however, in the future we also want to enable permanent storage of the ontology set analysis results, so as to corroborate exchange of information between different experimenters. We also plan to include more ontology repositories (e.g., Oxford Ontology Library), involve more graphs and more detailed ontology metrics. Finally, we plan to evaluate the usability of benchmark construction by OOSP within a concrete domain, via a user-study with potential ontology tool developers, e.g., for ontology visualization tools.

Acknowledgement Ondřej Zamazal has been supported by the CSF grant no. 14-14076P. This research is also supported by UEP IGA project no. F4/90/2015 and long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

1. Kalyanpur A., Parsia B., Sirin E., Cuenca Grau B. (2006). Repairing unsatisfiable concepts in OWL ontologies. In: 3rd European Semantic Web Conference 2006.
2. Katifori A. et al.: Ontology visualization methods - a survey. In: *ACM Computing Surveys (CSUR).* 39(4), 10 pages, 2007, ACM.
3. Khan, M. T., Blomqvist, E.: Ontology design pattern detection - initial method and usage scenarios. In: 4th Int. Conference on Advances in Semantic Processing 2010.
4. Lehmann J., Bühmann L. (2010). ORE - a Tool for Repairing and Enriching Knowledge Bases. In: 9th International Semantic Web Conference 2010.
5. Matentzoglu, N., Bail, S., Parsia, B.: A Snapshot of the OWL Web. In: 12th International Semantic Web Conference 2013.
6. Matentzoglu, N., Tang, D., Parsia, B., Sattler, U. The Manchester OWL Repository: System. In: 13th International Semantic Web Conference 2014 at poster session.
7. Noessner J., Niepert M.: ELOG: A Probabilistic Reasoner for OWL EL. In: 5th Conf. Web Reasoning and Rule Systems (RR 2011), Galway.
8. Zamazal, O., Svátek, V.: Towards Automation of Ontology Analysis Reporting. In: 14th Conference on Information Technologies Applications and Theory 2014.
9. Zamazal, O., Svátek, V.: Automated Exploration of Ontology Repositories. In: 11th International Workshop on OWL: Experiences and Directions (OWLED 2014).