

An Evaluation Dataset for Linked Data Profiling

Andrejs Abele, John McCrae, and Paul Buitelaar

Insight Centre for Data Analytics, National University of Ireland, Galway,
IDA Business Park, Lower Dangan, Galway, Ireland
`andrejs.abele`, `john.mccrae`, `paul.buitelaar@insight-centre.org`

Abstract. Since the beginning of the Linked Open Data initiative, the number of published Linked Data datasets has gradually increased. However, the reuse of datasets is hindered by a lack of descriptive and reliable metadata about the nature of the data, such as their topic coverage. Manual curation of metadata is however costly and hard to maintain, because of which we advocate a Linked Data profiling approach that will be able to automatically extract topics from datasets as metadata. One of the main challenges in developing this is the lack of evaluation data, i.e. manually curated metadata (topics) for datasets. In this paper we describe such an evaluation dataset and the framework that enabled its creation.

Keywords: linked data, linked data profiling, topic extraction, metadata, evaluation dataset

1 Introduction

With the emergence of the Web of Data, in particular Linked Open Data (LOD), the amount of machine readable Linked Data (LD) datasets has rapidly increased. However, reuse of these datasets is hindered by a lack of descriptive and reliable metadata about the nature of the data, such as their topic coverage. Manual curation of metadata is however costly and hard to maintain, because of which we advocate a Linked Data profiling approach that will be able to automatically extract topics from datasets as metadata. There have been recent attempts to automatically create qualitative descriptions of LD datasets, such as by Fetahu et al. [8], but research in Linked Data profiling to date has largely been stymied by the unavailability of evaluation datasets.

Apart from the lack of metadata describing datasets, there is incomplete classification of existing metadata. Currently, the most widely used topic classification was introduced by the Linked Open Data cloud diagram [10]. This classification contains 9 topics (domains): Government, Publications, Life sciences, User-generated content, Cross-domain, Media, Geographic, Social web, Linguistics. In contrast, the “Linguistic Linked Open Data Cloud” [1] focuses only on linguistic data and provides two types of classification, i.e., by dataset type and by dataset license. Obviously, these classifications are very broad and do not, for example, support use cases where users are looking for more specific or similar datasets.

Our goal is, by extracting and analyzing topics from all datasets on the LOD cloud, to create an extended hierarchical topic classification for use in subsequent LD dataset classification. In this paper, we describe an important first step towards achieving this goal, consisting of the development of an evaluation dataset for topic extraction from LD datasets.

The rest of this paper is structured as follows: section 2 describes our approach in creating the evaluation dataset, whereas section 3 describes the dataset itself, followed by some conclusions of our work in section 4.

2 Topic Extraction from Linked Data

In this section we describe our approach for topic extraction from Linked Data, which is leveraged for the construction of the evaluation dataset. A graphical representation of the framework is provided in Figure 1. In this diagram, we show all steps of the process that are performed for each of the datasets (in the LOD cloud) in the plate labelled “D”. We also show the external datasets, especially DBpedia [5], which is a knowledge base that provides relations about entities and is used to generate the topic classification hierarchy. We also compile a term database that contains statistics about extracted n-grams (unigram, bigram, trigram) from LOD datasets, i.e., term frequency and inverse dataset term frequency. Before the extraction of n-grams, all literal text was transformed to lower-case, and all standard English stop words¹ and HTML keywords (e.g., http, https, string, font, family, style, size, width, html) were removed.

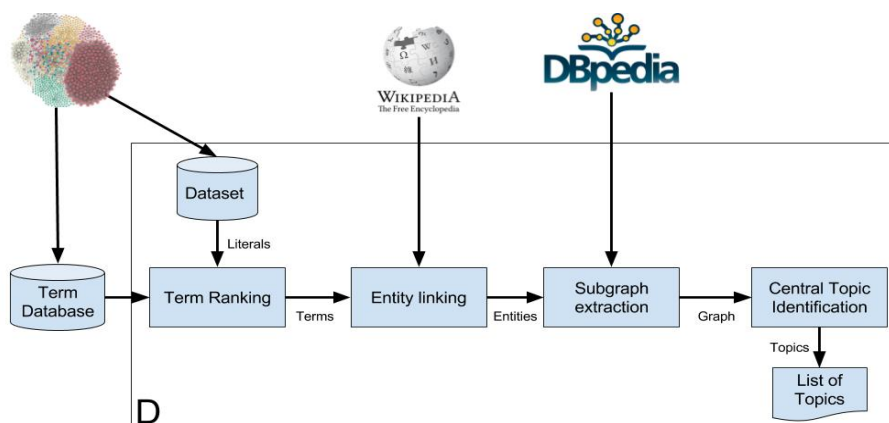


Fig. 1. Linked Data dataset extraction framework

¹ http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

2.1 Term Ranking

Our approach relies on the assumption that top ranked terms from a dataset provide enough information and it is not required to create statistics based on the whole textual context (all literals) of the dataset. For this reason we have implemented multiple ranking approaches, where the user can choose which ranking algorithm to use, or use aggregated top results from all ranking methods.

Term Frequency is a basic approach to identify term importance in a dataset, which provides good results on LD datasets as literals mostly contain nouns and if a noun often appears in the dataset, it is important.

Term Probability uses additional information that is collected while processing all datasets. To identify terms that frequently occur not only in the analysed dataset but also in other datasets, we aggregate the term occurrence over all processed datasets and calculate the probability that the term occurs in this dataset:

$$TP_d(x) = \frac{TF_d(x)}{\sum_{d'}^D TF_{d'}(x)} \quad (1)$$

Where d is a given dataset and D is the collection of all processed datasets. $TF_d(x)$ is the frequency of term, x , in dataset, d , and $TP_d(x)$ is the probability that term, x , will occur in dataset, d . The higher the probability, the more important the term is to the dataset.

Term Frequency–Inverse Document Frequency is the most popular approach and is commonly used in the field of information retrieval and text mining. To calculate TF-IDF[3] we use the following formula:

$$TFIDF_d(x) = TF_d(x) \times IDF(x) \quad (2)$$

Where $IDF(x)$ is inverse document frequency:

$$IDF(x) = \log \frac{1 + n_D}{1 + DF_D(x)} \quad (3)$$

Where n_D is the number of processed datasets, $DF_D(x)$ is number of datasets that contain term x .

Pointwise Mutual Information is a measure of association used in information theory and statistics, and we modified it to identify the most important terms in a dataset.

$$PMI(x) = \log \frac{P_{xd}}{PX(x)} \times n_D \quad (4)$$

$$P_{xd} = \frac{TF_d(x)}{n_d} \quad (5)$$

Where n_d is size of dataset d .

$$PX_x = \sum_d^D P_{xd} \quad (6)$$

2.2 Entity Linking

Once the top terms have been identified they are linked to DBpedia, for which we used an existing library: Yahoo Fast Entity Linker Core [6, 7]. This library performs query segmentation and entity linking to a target reference Knowledge Base (Wikipedia²). It is tailored towards query entity linking and meant for short text fragments in general, which is appropriate for our application as we also deal with short text fragments in the form of LD literals. DBpedia uses the same id's for entities as Wikipedia for articles, for this reason Wikipedia articles can be directly linked to DBpedia entities and their categories.

2.3 Subgraph Extraction

After the entities have been identified, we select the top most frequent and extract them together with their surrounding entities. We use a local DBpedia SPARQL endpoint and programmatically use two *SPARQL Queries*. First, the query searches for objects that are connected to the *entity* by the properties `skos:broader` or `dc:subject`, filtering out all objects that are not URIs. In the next step, we filter out those URIs that are not linked to DBpedia, as we cannot guarantee that external links will be resolvable.

```
"select distinct ?P ?O where {
VALUES ?P {<http://www.w3.org/2004/02/skos/core#broader>
           <http://purl.org/dc/terms/subject> }
<"+entity+"> ?P ?O
FILTER ( ISIRI(?O) ) }"
```

The second query retrieves objects (entities) that are connected to our original *entity* and filters out all objects that are not URIs and are not linked to DBpedia.

```
"select distinct ?O0 ?P ?O where {
VALUES ?P {<http://www.w3.org/2004/02/skos/core#broader>
           <http://purl.org/dc/terms/subject> }
<"+entity+"> <http://purl.org/dc/terms/subject> ?O0 .
?O0 ?P ?O FILTER ( ISIRI(?O) ) }"
```

To reduce the amount of irrelevant data, we remove certain types of entities e.g.:

- <http://dbpedia.org/resource/Template>
- <http://dbpedia.org/ontology/wikiPageWikiLink>
- <http://dbpedia.org/ontology/language>

Once all the entities from the queries are connected they form a small graph (DBpedia subgraph) that covers all topics in the given dataset.

² <https://en.wikipedia.org>

2.4 Central Topic Identification

Similarly to Term Ranking, we implemented multiple graph centrality algorithms to identify central topics:

Betweenness[2] is a centrality measure of a node within a graph. A node is important if it facilitates the flow of information between other nodes in the graph. The betweenness centrality is strongly biased towards nodes with high degree or nodes that are central in large local groups of nodes.

Degree centrality represents the number of edges that a node has.

EigenVector is a centrality measure that represents a node’s influence in a network. It assigns relative scores to all nodes in the network based on the idea that connections to high-scoring nodes provide higher score.

PageRank[9] is a measure similar to EigenVector that calculates influence of a node based on its connection to other influential nodes.

Closeness is a centrality measure that represents the average length between the node and all other nodes in the graph.

3 Evaluation Dataset

The requirements for a Linked Data topic extraction evaluation dataset are as follows: consist of accessible LD datasets that contain textual data (literals) and for each dataset provide a list of relevant topics associated with it. An example of the evaluation dataset is shown in Table 1

Table 1. Evaluation Dataset Example

Dataset name	Topic list
education-data-gov-uk	Education,School,United_Kingdom
olia	Linguistics, Ontology_(information_science)
nobelprizes	Awards, Nobel_Prize, Royal_Swedish_Academy_of_Sciences

3.1 Dataset Collection

We wanted to use known datasets, so annotators could easier annotate them. For this reason we downloaded all 570 datasets that are present in the LOD Cloud Diagram 2014[11]. When we attempted to download these using information present at DataHub³, only 245 were reachable by using our automated crawler.⁴

³ <https://datahub.io/>

⁴ On 07.06.2016

3.2 Dataset Filtering

For annotators to be able to annotate the dataset, they require background information about the dataset, for which reason we excluded 70 datasets that did not contain any metadata (not even a link to the data publisher). Furthermore, for any topic extraction system to be able to process a dataset, the dataset has to contain textual data (literals). 34 datasets did not contain any literals. After excluding all datasets that did not comply with our requirements we obtained a collection of 141 datasets for our evaluation dataset construction.

3.3 Candidate Topic Extraction

To provide annotators with a reasonable amount of topics to select from for every dataset, we processed the datasets using the approach described in section 2. As mentioned in section 2.1, our approach provides multiple term ranking options. For this reason and to remove any bias towards one of the ranking methods, we selected the top results from all methods. When we performed Central Topic Identification, we collected top results using each centrality measure to remove bias towards one of the algorithms, as described in section 2.4.

After collecting all possible topics for each dataset, a domain expert who is knowledgeable in LOD domains, reviewed the data to further narrow down the topic options for the annotators. The domain expert, after reviewing the output of the framework, excluded multiple datasets for the following reasons: 1) datasets with text literals in a language other than English (our approach only supports the analysis of English text) 2) datasets that cover multiple domains (making it challenging for annotators to come to an agreement). The domain expert further limited the number of possible topic options for each dataset, where the number can be from nine to fourteen topics per dataset. In Table 2 we show 3 of the processed datasets and their description. As shown, these datasets cover diverse domains.

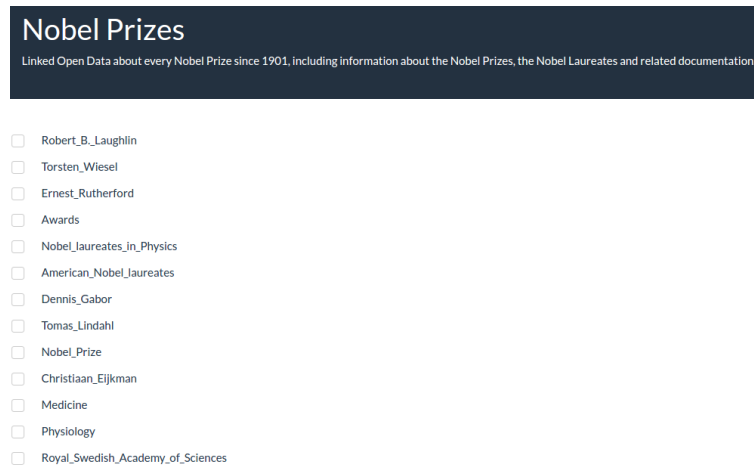
Table 2. List of datasets

Dataset	Description
clean-energy-data-reegle	Linked Clean Energy Data (reegle.info) Comprehensive set of linked clean energy data
nobelprizes	Nobel Prizes Linked Open Data about every Nobel Prize since 1901, including information about the Nobel Prizes, the Nobel Laureates and related documentation.
statistics-data-gov-uk	Statistics.data.gov.uk Linked data about administrative areas used within UK government official statistics.

3.4 Survey creation

For each annotator to be able to perform their task, they require a list of topics for each dataset, as well as a description of the dataset. For some of the datasets there is sufficient description in Datahub.io (e.g., nobelprizes), but there are many that do not have any descriptions (e.g.,dws-group). For these datasets we went to the publisher’s homepage and manually extracted the description for the dataset.

To publish and manage the survey, we used LimeSurvey⁵, an open source tool. Annotators could select up to 10 topics for each dataset, and they can see only one dataset at a time. In Fig.2 it is shown how a question looks like in the LimeSurvey system.



The image shows a LimeSurvey question form. At the top, there is a dark blue header with the text "Nobel Prizes" in white. Below the header, a smaller line of text reads: "Linked Open Data about every Nobel Prize since 1901, including information about the Nobel Prizes, the Nobel Laureates and related documentation." Below this, there is a list of 14 topics, each preceded by an unchecked checkbox:

- Robert_B_Laughlin
- Torsten_Wiesel
- Ernest_Rutherford
- Awards
- Nobel_Laureates_in_Physics
- American_Nobel_Laureates
- Dennis_Gabor
- Tomas_Lindahl
- Nobel_Prize
- Christiaan_Eijkman
- Medicine
- Physiology
- Royal_Swedish_Academy_of_Sciences

Fig. 2. LimeSurvey Question form example

3.5 Dataset Description

After evaluating the results, we kept only those topics where from 10 annotators at least 6 agreed, which was chosen to balance the number of topics extracted with the overall accuracy of the annotation. In Fig 3 we can see that only for 13 topics all annotators came to a full consensus, and for 91 topics they agreed that the topics do not belong to the dataset. Fig 4 shows the number of topics in each dataset, where 6 or more annotators reached a consensus about the topic. As can be seen, there is 1 dataset (dataset 27) that has 0 topics. This dataset is about “Ontos News Portal”. From 13 possible topics, the highest consensus was reached for the topic ”News_media”, but it was only by 5 annotators and it is below our threshold. The full evaluation dataset in TSV format can be found

⁵ <https://www.limesurvey.org>

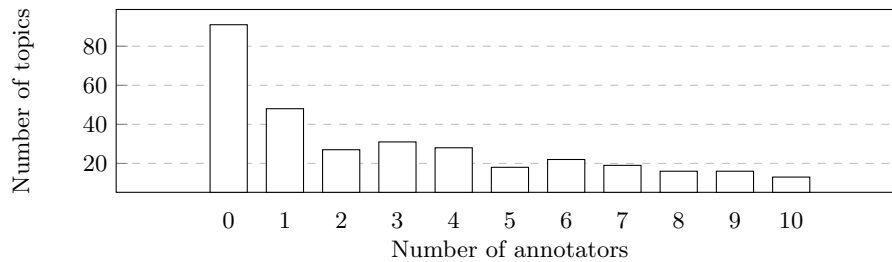


Fig. 3. Number of topics selected by certain number of annotators

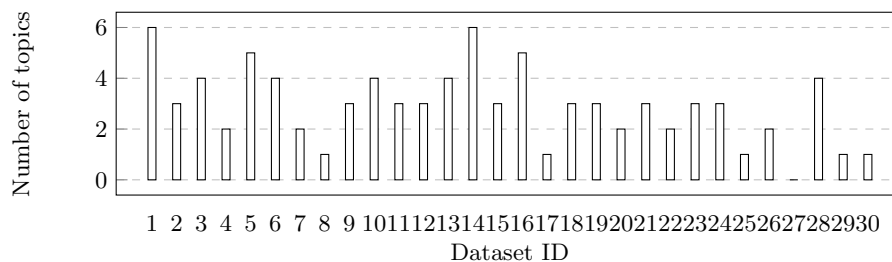


Fig. 4. Number of topics for each dataset in the evaluated dataset

here⁶. The full survey data in TSV format can be found here⁷.

4 Conclusion

We have constructed an evaluation dataset for Linked Data profiling by applying a topic extraction approach that links extracted topics to categories drawn from the DBpedia hierarchy. The assignment of topics to datasets had a high degree of agreement among human annotators, which makes it a good benchmark for automatic Linked Data profiling approaches. The evaluation dataset will enable us also to define a more fine-grained topic classification for Linked Data, which will be important for further uptake and automatic use by data consumers. Current categorizations, such as used for the LOD cloud diagram [10], are very high-level and are not organized hierarchically.

Acknowledgements

This work was supported by Science Foundation Ireland under grant number SFI/12/RC/2289 (Insight) and by the European Union under grant number H2020-644632 (MixedEmotions).

⁶ <https://nuig.insight-centre.org/unlp/evaluationdataset-tsv/>

⁷ <http://nuig.insight-centre.org/unlp/datasetsurveydata-csv/>

References

1. Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff. Linking linguistic resources: Examples from the Open Linguistics Working Group, In: Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics. Representing Language Data and Metadata*, Springer, Heidelberg, p. 201-216., 2012.
2. Freeman, Linton C.: "A set of measures of centrality based on betweenness." *Sociometry* (1977): 35-41.
3. Salton, Gerard, and Michael J. McGill. "Introduction to modern information retrieval" McGraw-Hill, ISBN 0-07-054484-0. (1983): 402-403.
4. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
5. Auer, Sren, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. "Dbpedia: A nucleus for a web of open data." In *The semantic web*, pp. 722-735. Springer Berlin Heidelberg, 2007.
6. Blanco, Roi, Giuseppe Ottaviano, and Edgar Meij. "Fast and space-efficient entity linking for queries." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 179-188. ACM, 2015.
7. Pappu, Aasish, Roi Blanco, Yashar Mehdad, Amanda Stent, and Kapil Thadani. "Lightweight Multilingual Entity Extraction and Linking." In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 365-374. ACM, 2017.
8. Fetahu, Besnik, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl: "A scalable approach for efficiently generating structured dataset topic profiles." In *European Semantic Web Conference*, pp. 519-534. Springer International Publishing, 2014.
9. Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd: "The PageRank citation ranking: bringing order to the web." (1999).
10. Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak: "Linked Open Data cloud diagram 2017, <http://lod-cloud.net/>"
11. Schmachtenberg, Max, Christian Bizer, and Heiko Paulheim: "State of the LOD Cloud 2014." University of Mannheim, Data and Web Science Group [en ligne] 30 (2014).
12. Auer, Soren, Jan Demter, Michael Martin, and Jens Lehmann: "LODStatsan extensible framework for high-performance dataset analytics." In *International Conference on Knowledge Engineering and Knowledge Management*, pp. 353-362. Springer Berlin Heidelberg, 2012.