View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Dagstuhl Research Online Publication Server

Interlinking SciGraph and DBpedia Datasets Using Link Discovery and Named Entity Recognition Techniques

Beyza Yaman
Institute of Applied Informatics, Leipzig, Germany
yaman@infai.org

Michele Pasin
Springer Nature, London, UK
michele.pasin@springernature.com

Markus Freudenberg
Leipzig University, Leipzig, Germany
markus.freudenberg@eccenca.com

Abstract

In recent years we have seen a proliferation of Linked Open Data (LOD) compliant datasets becoming available on the web, leading to an increased number of opportunities for data consumers to build smarter applications which integrate data coming from disparate sources. However, often the integration is not easily achievable since it requires discovering and expressing associations across heterogeneous data sets. The goal of this work is to increase the discoverability and reusability of the scholarly data by integrating them to highly interlinked datasets in the LOD cloud. In order to do so we applied techniques that a) improve the identity resolution across these two sources using Link Discovery for the structured data (i.e. by annotating Springer Nature (SN) SciGraph entities with links to DBpedia entities), and b) enriching SN SciGraph unstructured text content (document abstracts) with links to DBpedia entities using Named Entity Recognition (NER). We published the results of this work using standard vocabularies and provided an interactive exploration tool which presents the discovered links w.r.t. the breadth and depth of the DBpedia classes.

2012 ACM Subject Classification Information systems → Semantic web description languages; Computing methodologies → Natural language processing; Information systems → Entity resolution

Keywords and phrases Linked Data, Named Entity Recognition, Link Discovery, Interlinking

Digital Object Identifier 10.4230/OASIcs.LDK.2019.15

Category Short Paper

1 Introduction

Scientists often search for the articles related to their research areas, however, often they fail to find the relevant publications on the search engines due to lack of semantics on document oriented search results. Thus, creating meaningful links and relations over various data sets is required to discover relevant results for the given user queries. In this paper, we describe how Linked Data technologies are applied to a publications metadata dataset from Springer Nature, such that, it is enriched with bi-directional relations to DBpedia concepts. Consequently, automatically generated semantic relations permit to construct more interesting discovery tools and contribute to the emergence of a more deeply interlinked web of data.

Springer Nature is one of the leading publishers for the educational sources and publishes large amount of articles online each year providing top-level studies to the service of the researchers but discoverability of the content is the common issue among all data sets. Thus,


© Beyza Yaman, Michele Pasin, and Markus Freudenberg;
licensed under Creative Commons License CC-BY
2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 15; pp. 15:1–15:8

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

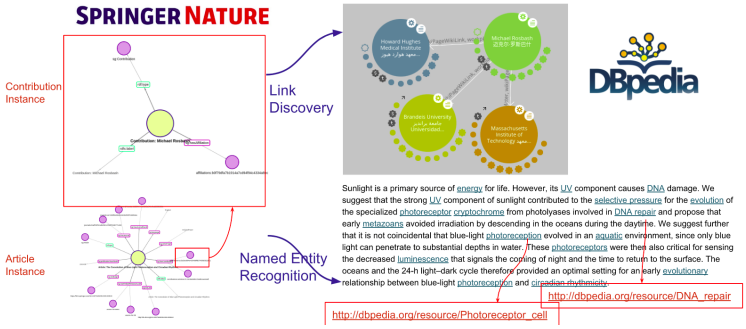


Figure 1 Two Main Approaches for interlinking SciGraph and DBpedia.

Springer Nature introduced SN SciGraph which is a Linked Open Data platform of Springer Nature Publishing and its key partners offering content from the scholarly domain. SN publishes documents and data where users can search and find the entities related to science and the scholarly domain. Platform provides around 1.5 to 2 billion triples across the research landscape dating from 1839 to 2018, e.g., funders, research projects, conferences, affiliations and publications. The model, ontology and the data sets are published under public licences providing its services to the users to explore the SciGraph data landscape in an interactive manner using SN Scigraph Explorer[1]. Moreover, data specialists can retrieve rich data descriptions for SciGraph objects by using the Linked Data API.

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web [1]. DBpedia data is available as Linked Data revolutionizing the way applications interact with the Web and the data sets which can serve many purposes, e.g., natural language processing, knowledge exploration, query answering. DBpedia dataset was chosen to link with SciGraph, since, it is one of the most connected and referenced data hubs on the Linked Data cloud. Not only being a data hub but also having a good categorization and type hierarchy structure convinced us that DBpedia is the most suitable data set for our use-case.

Considering these two large data sets, our main objective was to investigate application methods to enrich and improve Scigraph by employing bi-directional relations of Linked Data technologies. Thus, the contribution of this paper is three-fold: *i)* discovering links for metadata enrichment on SN articles to increase discoverability of the articles *ii)* increasing the impact of SciGraph in LOD cloud by identifying links in the existing datasets *iii)* exploring scholarly publications using DBpedia concepts. We present the applied methodologies in the following example.

Example. Fig. 1 shows above mentioned approaches illustrating on a Springer-Nature article from the Nobel Prize winner Michael Rosbash. The article's bibliographic metadata is represented as Linked Data within SN SciGraph via `sg:Article` class. This object contains information about the article's authors via `sg:Contribution` class which is used to trigger a Link Discovery algorithm and to find Michael Rosbash URI in DBpedia. This result in turns allows connecting two data sets with further links (see upper part of the Fig. 1). On the other hand, the text abstract of the article contains a wide range of keywords. They are useful to a human reader in order to have an idea about the topics mentioned in the article, however,

[1] `https://scigraph.springernature.com/explorer`

they lack formal semantics and hence cannot be interpreted effectively by machines. In order to increase the machine readability of the abstract, the text is enriched by discovering and linking these keywords to DBpedia resources via a Named Entity Recognition algorithm. (see at the bottom of Fig. 1).

In the rest of this paper, we will provide more details about the tools and methodologies adopted for these two approaches. We will discuss the obtained results and faced challenges. The remaining part of the paper proceeds as follows: Second section describes the applied methodologies in this study and the produced data sets with appropriate metadata. Third section discusses the prototype to explore publications using DBpedia concepts. The fourth section of this paper presents our conclusions, and finally, fifth section examines the possible future research directions.

2 Approach

In this section, we describe the employed techniques to interlink two data sets with relevant background and implementation details, as well as, outlining the principal results produced from the tasks.

2.1 Link Discovery

Link discovery (LD), which is considered as entity reconciliation in relational databases, is the process of automatically discovering the overlapping parts of heterogeneous data sets, and linking individual records from these data sets by exploiting their specific properties. Link Discovery is described along these lines [4]: Given two sets S (source) and T (target) of instances, a (complex) similarity measure θ over the properties of s $\in$ S and t $\in$ T, and a similarity threshold $\theta \in [0, 1]$, the goal of LD is to compute the set of pairs of instances $(s, t) \in S \times T$ such that $\gamma(s, t) \geq \theta$.

Considering this definition, we investigated some of the implemented tools specialized for Linked Data, namely, LogMap[3], KnoFuss [7], Silk[11], LIMES[6], RiMOM[10] and RuleMiner[8] to select the most convenient tool for our project. We used two frameworks to test our data set: Silk and Limes due to their advantages among other tools [5]. These advantages are high range of input types (RDF, SPARQL, CSV), various similarity measures (e.g., string, geospatial, date), ability to produce user defined links (e.g. `skos:closerMatch` while other tools only support `owl:sameAs` links), open source usage, graphical user interface, manual (rule-based description language) and semi-automatic (supervised methods) configuration possibility which allows generating links based on the similarity functions expressed in XML link specification. In the next section, we present the implementation details using this link specification configuration file.

2.1.1 Implementation Details

The interlinking process is performed by running an interlinking script with above mentioned interlinking tools between two overlapping web data sets: SciGraph *Contribution* class and DBpedia *Person* class. We produced a link specification configuration to find interlinks between instances and algorithm of the configuration which can be found in our GitHub repository. However, we have seen that Silk has a wider range of operations and transformation functions which are applied to the properties (tokenizations, lowercase etc) than Limes. Therefore, although we used Limes to test the tool and to contribute to its development, we exploited only Silk to produce links from the actual data set.

While extending the configuration file iteratively to find the best configuration, we also extended the data set with more distinctive properties. Therefore, common links between SciGraph and DBpedia data sets are increased by enriching both of them with additional properties: i) SciGraph data set is extended with properties from Orcid data set which provides a unique ID for each researcher. ii) DBpedia links are extended with unique ids from Grid data set. Thus, the links between *Affiliation* class from SciGraph and *Organization* class from DBpedia are increased by adding *Affiliation* information to the configuration. However, instead of link discovery method, these links are discovered by using link traversal methods by creating direct links between Grid organizations[2] and DBpedia Organizations discovering 30.426 links between those data sets.

2.1.2 Results

We have executed the configuration on the *Contribution* instances of the 2017 abstract articles and DBpedia *Person* instances. Since the data sizes are very large, we have limited the properties in the data sets, including only the ones used in the configuration file. Even though the framework executed for 30 days, only 11.6% of the tasks were completed, thus, we had to interrupt the execution but 47.913 links have been found in this period.

Table 1 Found links by Link Discovery approach.

Task	#SciGraph Instances	#DBpedia Instances	#Found Links
Contribution-Person	1.412.018	1.396.811	47.913 links

2.2 Named Entity Recognition

Named entity recognition (NER) is the automatic extraction process of name identification in the unstructured text. This process involves identification of proper names in texts, and classification into a set of predefined categories of interest with the possibility of connecting them to a knowledge base (DBpedia, Wikidata) to enrich the data semantically and allow to extract new connections based on created links. This structured information has the potential of deducing new inferences and arriving to the new conclusions with much more meaningful solutions, as well as, more relevant answers to the posed queries.

In the scope of this work, we first analysed the different NER tools, namely, DBpedia Spotlight[2], Stanford NER[3], AlchemyAPI [4], ANNIE [5], Open Calais [6]. Among them all, DBpedia Spotlight, which is a tool enabling automatic annotation of text documents with DBpedia URIs, is selected to conduct our experiments. DBpedia Spotlight is chosen due to its public licence, its optimal results with preliminary abstract tests and its wide range of linking possibility to the DBpedia resources chosen among more than 380 cross-domain types existing in DBpedia (e.g., people such as Obama, chemical compounds such as alkali salt or more general concepts such as humanitarian aid). The tool uses spotting, candidate selection, disambiguation and filtering respectively to discover the name entities in the text content and produces either candidate links or named entity links with requested data format, e.g. NIF, XML, JSON.

[2] https://www.grid.ac/downloads
[3] https://nlp.stanford.edu/software/CRF-NER.html
[4] https://www.ibm.com/watson/alchemy-api.html
[5] http://services.gate.ac.uk/annie/
[6] http://www.opencalais.com/

2.2.1 Implementation Details

DBpedia Spotlight provides a flexible configuration to the users according to their specific needs via DBpedia Ontology type filters, resource prominence (support) and disambiguation confidence of the spotted phrase. Type filter annotates only resources of a certain type or set of types, however, filter usage is avoided because of the interdisciplinary nature of the abstracts which might result with very restrictive outcomes. Support parameter defines the minimum number of inlinks a DBpedia resource has to have in order to be annotated where high support selects the more famous links. We configured this parameter to be low (20) to avoid the filtering of more relevant links. Moreover, we set higher confidence (0.85) for the actual data set to avoid noises after test evaluations on the abstract texts with 0.45 and 0.55 confidence.

We implemented a tool to produce the interlinks between data sets automatically for the given configuration which is openly provided to the community usage[7]. Although the tool is employed for the Springer Nature abstracts, it can be configured for any type of text to produce named entities. This tool allows analyzing the abstracts according to their topic and language, producing the links between articles and DBpedia resources. The tool has been assessed by processing the test data for analysis purposes of the abstracts with several adjustments to find optimized configuration for best results.

2.2.2 Results

In the scope of this work, 2017 article abstracts and 2017 book chapter abstracts are used from overall SN data sets: i) Articles data set is assessed by the given configuration and as a result, 187.107 abstracts are processed to identify the named entities in the content. The statistics on found entities are presented in Table 2. ii) Book chapters data set is assessed by the given configuration, thus, at the end 4880 abstracts are assessed where the statistics for the book chapters can be seen in Table 2. It is apparent from both articles and book chapters table that increase of the confidence value causes a decline on the produced number of the entities respectively. However, having more accurate links also comes with a side affect decreasing the number of the correct links as well. These data sets can be found on the GitHub repository of the project [8].

Table 2 NER Results for Articles and Book Chapters.

Data Set	Confidence	#abstracts entities	#distinct entities	# found entities	Average link per abstract	Execution time
Articles	0,85	187.107	54.077	776.424	4,14	~ 483 ms
Articles	0,55	187.107	89.138	2.841.682	8,7	~ 537 ms
Articles	0,45	187.107	274.204	3.967.124	10,26	~ 580 ms
BookChapters	0,85	4880	7.538	24.127	4,94	~ 332 ms
BookChapters	0,55	4880	12.227	45.205	9,26	~ 380 ms
BookChapters	0,45	4880	14.911	61.013	12,5	~ 434 ms

Metadata. NIF dataset is produced for each article and book chapter abstract with the prefix of the article for the named entities using NIF core ontology[9]. Moreover, provenance links are provided from the phrases to the article to reference the source of the phrase back

[7] http://hacks2019.michelepasin.org/dbpedialinks
[8] https://github.com/dbpedia/sci-graph-links
[9] http://persistence.uni-leipzig.org/nlp2rdf/

to the article as it can be seen in Listing 1. This triple shows the origin of the phrase by using `prov:hadPrimarySource` property. Such that, it allows us to traverse the phrase back to its origin source article or it would be possible to find named entities for a given article. This piece of information is included in the data set folder as well.

Listing 1 Phrase provenance link to its article.

```
<http://scigraph.springernature.com/things/articles/d8a8cee79015eecf1ff48e2edd4c27a3#
    offset_684_696> <http://www.w3.org/ns/prov#hadPrimarySource> <http://scigraph.
    springernature.com/things/articles/d8a8cee79015eecf1ff48e2edd4c27a3>
```

Parallel to the creation of the NIF dataset, also Backlinks data set is created which includes direct links from SciGraph article phrases to the DBpedia resources in the quadruple format as it is presented in Listing 2. This quadruple connects the phrase with DBpedia via `schema:mentions` property with additional information of its article, the tool it is produced by and the confidence of the tool.

Listing 2 Phrase backlink to DBpedia with confidence value.

```
<http://scigraph.springernature.com/things/articles/d8a8cee79015eecf1ff48e2edd4c27a3#
    offset_684_696> <http://schema.org/mentions> <http://dbpedia.org/resource/
    Transfection> <http://scigraph.springernature.com/things/articles/
    d8a8cee79015eecf1ff48e2edd4c27a3#nlptool=spotlight&confidence=1.0>
```

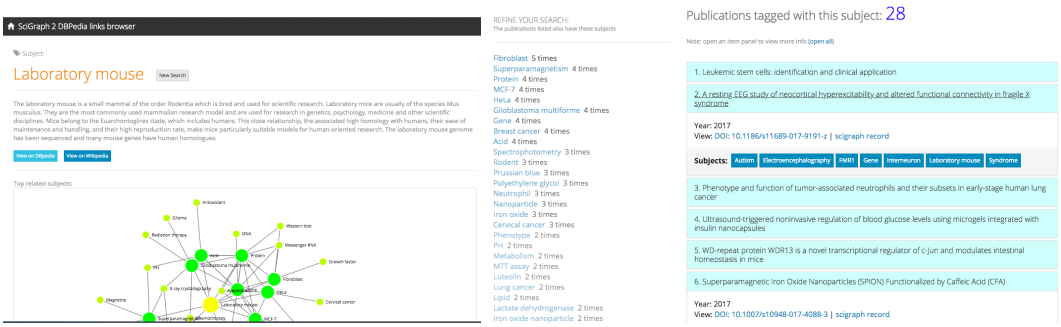
3 Application: Discovering publications using DBpedia concepts

In order to assess the relevance and usefulness of the extracted links using named entity recognition approach (see Section 2.2), a web application tool is developed that allows discovering SN publications using the DBpedia concepts they have been tagged with.

The application, which is freely available online[10], allows users to explore a subset of the data presented in this paper (87k publications tagged using 54k DBpedia concepts). An exploration journey can be initiated either by searching for a specific DBpedia concept using keywords or by listing out all of them alphabetically. Once a concept of interest has been selected, a "topic" page for that concept is presented to users, which provides a description of the concept (dynamically retrieved from DBpedia) and a list of publications tagged with that concept (Fig 2). In order to make the browsing experience more interesting and allow for a more serendipitous discovery of related content, the application presents to users other relevant concepts employing various mechanisms: first, an interactive network visualization representing the most frequent co-occurring concepts (Fig. 2.a); second, a text list of all co-occurring concepts with counts (akin to a facet search); finally, the full list of concepts related to each single publication can be displayed on-demand via a simple open/close panel widget (Fig. 2.b).

The goal of this exploration interface was to assess the relevance of DBpedia concepts via face-to-face user testing sessions involving domain experts; furthermore, it helped us shed some light on whether the kind and range of concepts available are appropriate for this kind of publication-discovery tasks. Finally, it also let us review these results with the Springer Nature ontology managers who are responsible for the (mostly manual) ongoing tagging of new content with keywords and ontology concepts. Historically, this task has been particularly time-consuming and difficult to manage, since it relies on a subject taxonomy developed in-house[11]and on the help of internal editors and domain experts.

[10] `http://hacks2019.michelepasin.org/dbpedialinks/`
[11] `https://scigraph.springernature.com/explorer/taxonomies/`



(a) Topic view of the articles with categories. **(b)** Open/close panel mechanism of the platform.

Figure 2 SciGraph Exploration Tool.

In general, despite the preliminary and informal character of these testing sessions, we still were able to gather some key findings:

- All users appreciated the breadth and depth of the concepts used to tag publications, often recognizing that it would be extremely costly to reproduce it at scale by using human annotators. Springer Nature publications simply covers too many subject areas for a manual approach to be sustainable.
- Although we used a rather high threshold for the Spotlight extraction algorithm (confidence of 0,85), we still encountered several instances of DBpedia concepts which are completely irrelevant (eg., "A roads in Zone 3 of the Great Britain numbering scheme" `http://hacks2019.michelepasin.org/dbpedialinks/entities/80611`, or "A Deeper Understanding" `http://hacks2019.michelepasin.org/dbpedialinks/entities/80649`). It's hard to speculate as to what percentage of data is wrongly annotated without a more systematic analysis. However, as a solution to this problem, it seems reasonable to assume that a mechanism to filter out extracted concepts based on the broader topic of a publication (e.g. "chemistry" or "physics") would be beneficial.
- The navigation mechanisms based on co-occurring concepts proved to be a powerful mean to explore the data set via relevant yet non-trivial pathways. In other words, they seemed to allow for a more serendipitous discovery mechanism compared to more static, taxonomy or ontology driven semantic relationships.
- Ontology managers particularly appreciated the fact that concept definitions are extracted from DBpedia automatically. Normally ontology managers spend a lot of time trying to get such definitions from subject matter experts, so they thought that using a Wikipedia definition as a starting point (or fall back) could be very valuable.
- Similarly, despite the wrongly tagged publications, ontology managers thought that often the DBpedia concepts could serve to identify under-represented areas in the corpus. Hence they could be used as candidate concepts for the official in-house subject taxonomy used at Springer Nature.

4 Conclusions and Future Work

In this paper, we have presented two approaches to increase the discoverability and reusability of the Springer Nature SciGraph scholarly data by integrating them to DBpedia, a highly interlinked data set in the LOD cloud. In order to achieve this goal, we applied techniques that a) improve the identity resolution across these two sources using Link Discovery for the

structured data and b) enrich SN SciGraph unstructured text content with links to DBpedia entities using NER. The educational publications are presented through topical navigation with specific links to DBpedia and Wikipedia to provide additional information from the open source knowledge. Overall, we strongly believe that the better connected scholar content can be highly useful for the researchers and end-users benefit from the created content.

Automated data will never be entirely accurate so mechanisms are in place for registered users to correct data when it is found to be wrong [9]. Thus, as future work, we aim at:

- evaluating the quality of the produced data sets employing crowd-sourced user feedback to produce higher quality contents.
- using these preliminary results in order to set up a more robust user evaluation study, which aims are reviewing larger sections of the concepts extracted.
- devising and testing more intelligent mechanisms to improve the accuracy of the DBpedia concepts associate to a publication: e.g. by clustering them based on general fields of studies so to be able to score them against the broader topic of a publication (which is available via journal or book level product tags).

References

1 Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

2 Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM, 2013.

3 Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. Logmap: Logic-based and scalable ontology matching. In *International Semantic Web Conference*, pages 273–288. Springer, 2011.

4 Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. A survey of current link discovery frameworks. *Semantic Web*, 8(3):419–436, 2017.

5 Axel-Cyrille Ngonga Ngomo. On link discovery using a hybrid approach. *Journal on Data Semantics*, 1(4):203–217, 2012.

6 Axel-Cyrille Ngonga Ngomo and Sören Auer. Limes-a time-efficient approach for large-scale link discovery on the web of data. In *IJCAI*, pages 2312–2317, 2011.

7 Andriy Nikolov, Victoria Uren, and Enrico Motta. KnoFuss: A comprehensive architecture for knowledge fusion. In *Proceedings of the 4th international conference on Knowledge capture*, pages 185–186. ACM, 2007.

8 Xing Niu, Shu Rong, Haofen Wang, and Yong Yu. An effective rule miner for instance matching in a web of data. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1085–1094. ACM, 2012.

9 Yves Raimond, Michael Smethurst, Andrew McParland, and Christopher Lowis. Using the past to explain the present: interlinking current affairs with archives via the semantic web. In *International Semantic Web Conference*, pages 146–161. Springer, 2013.

10 Jie Tang, Bang-Yong Liang, Juanzi Li, and Kehong Wang. Risk minimization based ontology mapping. In *Content Computing*, pages 469–480. Springer, 2004.

11 Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Silk-a link discovery framework for the web of data. *LDOW*, 538, 2009.