Exploiting Interlinked Research Metadata

Shirin Ameri[1], Sahar Vahdati[1(✉)], and Christoph Lange[1,2]

[1] Smart Data Analytics (SDA), University of Bonn, Bonn, Germany
{ameri,vahdati,langec}@cs.uni-bonn.de
[2] Fraunhofer, Intelligent Analysis and Information Systems (IAIS), Sankt Augustin, Germany

Abstract. OpenAIRE, the Open Access Infrastructure for Research in Europe, aggregates metadata about research (projects, publications, people, organizations, etc.) into a central Information Space. OpenAIRE aims at increasing interoperability and reusability of this data collection by exposing it as Linked Open Data (LOD). By following the LOD principles, it is now possible to further increase interoperability and reusability by connecting the OpenAIRE LOD to other datasets about projects, publications, people and organizations. Doing so required us to identify link discovery tools that perform well, as well as candidate datasets that provide comprehensive scholarly communication metadata, and then to specify linking rules. We demonstrate the added value that interlinking provides for end users by implementing visual frontends for looking up publications to cite, and publication statistics, and evaluating their usability on top of interlinked vs. non-interlinked data.

Keywords: Interlinking · Linked open data · Research metadata · Scholarly communication · Semantic publishing

1 Introduction

Linked Open Data (LOD) is a popular approach for maximizing both legal and technical reusability of data, and enabling its connection with further datasets [2]. However, without further work, LOD datasets do not yet provide added value to end users, as they are only accessible for service and application developers familiar with Semantic Web technology and the datasets' vocabularies.

OpenAIRE (OA), the Open Access Infrastructure for Research in Europe [9], aggregates metadata about research (projects, publications, people, organizations, etc.) into a central Information Space. It so far covers more than 13 M publications, 12 M authors and scientific datasets. OA metadata has been exposed as LOD [14], aiming at maximizing its reusability and technical interoperability by:

- providing an infrastructure for data access, retrieval and citation (e.g., a SPARQL endpoint or a LOD API),
- interlinking with popular LOD datasets and services (DBLP, ACM, CiteSeer, DBpedia, etc.),

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 3–14, 2017.
DOI: 10.1007/978-3-319-67008-9_1

– enriching the OpenAIRE Information Space with further information from other LOD datasets.

This work focuses on enriching the OpenAIRE LOD by interlinking, and utilizing this interlinked data to provide added value to users in situations where they need scholarly communication metadata, e.g., when they are looking for a publication to cite, or for all publications of a given author.

2 Related Work

Rajabi has studied the exploitation of educational metadata using interlinking methods [8]. His work objectives closely related to ours; however its application domain is eLearning services and therefore he discusses the benefits of interlinking educational (meta)data in practice. Rajabi et al. provide a comparison of interlinking tools as well as interlinking rules [7] and a method for identification of duplicate links [6]. Hallo et al. follow the same objective as we do, i.e., publishing Open Access metadata as LOD [3]. Their work focuses on providing better search services on top of open journal datasets, but their data could be used as a candidate dataset for our interlinking. Recent work by Purohit et al. addresses the problem of scholarly resource discovery [5]. They also reviewed tools providing such services and present a framework for Resource Discovery for Extreme Scale Collaboration (RDESC)[1] which has common objectives with OA. However, they have not yet initiated interlinking of research metadata and the provision of a comprehensive knowledge graph.

3 Background: OpenAIRE LOD Services

The main motivation for exposing OA as LOD is to provide wider data access, and easier and broader metadata retrieval by enabling interlinking with relevant and popular LOD datasets [14]. Metadata about different types of entities – research results (publications and datasets), persons, projects and organizations – that the OA infrastructure aggregates is being exposed as LOD. OA LOD uses terms from existing vocabularies and, where necessary, defines new terms. Existing ontologies reused include SKOS, CERIF, DCMI Terms, FOAF [14,15]. Two prefixes/namespaces are OA specific: `oav:` http://lod.openaire.eu/vocab/ for the OA vocabulary, and `oad:` http://lod.openaire.eu/data/ for OA instance data.

The data has been exposed in three ways: (1) small fragments of RDF, accessible by dereferencing the URI that identifies a particular entity, (2) a downloadable all-in-one dump[2], and (3) a SPARQL endpoint, i.e. a standardized query interface accessible over the Web[3].

It is envisaged to extend the OA LOD by enriching and interlinking it with the following types of data:

[1] https://tw.rpi.edu/web/project/RDESC.
[2] http://tinyurl.com/OALOD.
[3] http://lod.openaire.eu/sparql.

– data that has not (yet) been collected by OA's existing mechanisms, e.g., certain types of persistent identifiers of publications or people (e.g., ORCID),
– data that is expensive to collect and/or not included in the OA data model, e.g., data about scientific events, and
– data that is related to open research but out of the scope of the OA infrastructure itself and therefore not targeted to be ever collected, e.g., biographies of persons, or geodata about the locations of organizations.

The primary objectives are (1) providing added value to users, by enabling those who develop user-oriented applications and services to access a richer collection of relevant data than just OA's own, and (2) facilitating internal data management, e.g., by aiding the resolution of duplicates resulting from metadata being harvested from different repositories by linking to external reference points.

4 Interlinking

Interlinking the OA LOD with other LOD datasets required us to do the following preparatory work: (1) analyzing the OA metadata schema to find appropriate entity types and properties on which to interlink, (2) identifying candidate target datasets, and, (3) among existing link discovery tools, finding the one most appropriate for our purpose, before we could finally implement interlinking rules (Fig. 1).

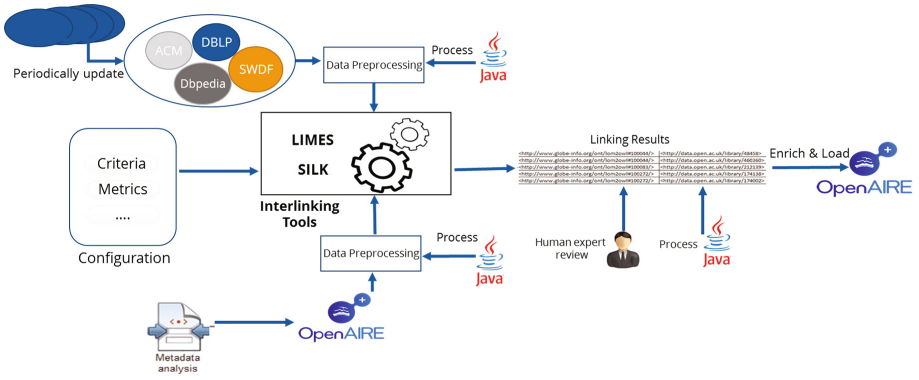

Fig. 1. Interlinking process

4.1 Identifying Properties Suitable for Interlinking

Not all properties of an OA entity are suitable for the purpose of interlinking to other entities, as Rajabi et al. have investigated in the related domain of metadata about educational resources [7]. Following their method, we analyzed

all OpenAIRE entities and their properties to discover linkable elements. We filtered out properties that potentially cannot be linked due to their specific values, for example Booleans (Yes/No), format values (PDF, JPEG), or language codes (en, de), and properties whose meaning is local to some source repository according to its policy, for example local identifiers or version numbers. This left us with properties such as 'publication title' and 'author name', 'published year', 'description', 'subject', etc., which have string or integer values. Where initial interlinking tests yielded subjectively satisfactory results, we chose the respective properties for interlinking – i.e. the following:

- **Title** and **Digital Object Identifier** of **Publication**,
- **Full name**, **First name** or **Last name** of **Person**s, and
- **Label** or **Homepage** of **Organization**s.

4.2 Investigating Existing Interlinking Tools

There exist a number of tools for creating semi-automatic links between datasets by running some matching techniques. These linking tools identify similarities between entities and generate links (e.g.owl:sameAs) that connect source and target entities. Rajabi et al. conducted a study that suggests that data publishers can trust interlinking tools to interlink their data to other datasets; accordingly, LIMES and Silk are the most promising frameworks [7]. Simperl et al. have compared various linking tools by addressing aspects such as required input, resulting output, considered domain and matching techniques used [11]. This allowed for a comparison from several perspectives: degree of automation (to what extent the tool needs human input) and human contribution (the way in which users are required to do the interlinking.

In summary, these comparisons point out the two well-known open source interlinking frameworks that we also used: LIMES[4] (Link Discovery Framework for Metric Spaces) and Silk[5] (Link Discovery Framework for the Web of Data). In an evaluation of the two frameworks, the LIMES developers showed that LIMES considerably outperforms Silk in terms of running time, with a comparable quality of the output. Moreover, LIMES can be downloaded as a standalone tool for carrying out link discovery locally and consists of modules that can be extended easily to accommodate new or improved functionality.

Our comparative evaluation of Silk and LIMES, which finally made us choose LIMES based on the quality of the output, is presented in Sect. 6.1.

4.3 Identifying Interlinking Target Datasets

To identify appropriate target datasets to be interlinked with OA, we examined several datasets from the LOD Cloud, in the following steps:

[4] http://aksw.org/Projects/LIMES.html.
[5] http://silkframework.org/.

1. **Identifying publication-related datasets in DataHub:** our aim is to find datasets tagged with the same domain as that of OA or a related one. We therefore searched the DataHub portal[6] for datasets tagged with 'publication' or related domains. This search yielded more than 900 datasets.
2. **Checking data endpoint availability:** we filtered the datasets identified previously by checking their SPARQL endpoints' or RDF dumps' availability.
3. **Retrieving datasets specification:** of the remaining datasets (still more than 60), we next retrieved each dataset's specification (size, metadata schema, etc.). From an interlinking point of view, we considered data volume, frequent updates, and matches with the entity types and properties identified previously (Sect. 4.1) as the most important characteristics of a dataset. Moreover, we considered available links to other related datasets desirable.

Table 1 lists the ten most relevant datasets according to these criteria.

Table 1. List of candidate Datasets

Datasets	Size	Endpoint	Dump	Covered OA entity types
DBpedia	1 B	Available	NT	Person, Organization
DBLP	55 M	–	NT	Publication, Person
ACM	12 M	Available	RDF/XML	Publication, Person
CiteSeer	8 M	Available	RDF/XML	Publication, Person
BibBase	200 K	–	RDF/XML	Person, Publication, Organization
IEEE	200 K	Available	RDF/XML	Publication, Person
OpenCitations	3 M	Available	JSON-LD	Person, Publication, Organization
SWDF	242 K	–	RDF/XML	Person, Publication, Organization
BNB	109 M	–	NT, RDF/XML	Person, Publication
COLINDA	149 K	Available	RDF/XML	Publication
GeoNames	93 M	–	RDF/XML	Organization

4.4 Identifying String Matching Algorithms

One of the most important factors in discovering links effectively is choosing the right string matching algorithm. The results of our heuristic experiments shows that both tools supports string matching according to trigrams, Levenshtein [7], Jaro, Jaro-Winkler and cosine (all of them normalized); cf. Table 2. It shows detailed definition of the algorithms. In our initial experiments, Jaro and Levenshtein proved most reliable for identifying equivalent names and titles. Thus, we chose Levenshtein for long string values, i.e., publication titles, and Jaro for short string values, i.e., person names. An example of a metric definition in LIMES is shown below.

[6] https://datahub.io/.
[7] https://wikipedia.org/Levenshtein_distance.

Table 2. String matching algorithms

Metric	Description
Trigrams	uses the number of matching triples in both strings as $s = 2 \times \frac{m}{(a \times b)}$ where m is the number of matching trigrams, a is the number of trigrams in string 1, and b is the number of trigrams in string 2 [10]
Levenshtein	is based on the minimum number of insertion, deletion or replacement operations required to transform string 1 into string 2
Jaro	is a measure of characters in common, being no more than half the length of the longer string in distance, with consideration for transpositions; it is best suited for short strings such as person names [12]
Jaro-Winkler	is an optimized version of Jaro designed and best suited for short strings such as person names
Cosine	is the cosine of the angle between string vectors; for equal strings the angle between them will be 0 and the cosine will be 1 [10]

```
<METRIC>
  AND(Jaro(x.foaf:name, y.foaf:name)|0.8, Levenshtein(
  x.dcterms:creator/cerif:name, ^y.dblp:hasAuthor/dblp:title)|0.8)
</METRIC>
```

5 Use Cases

The main objective of OA LOD is to achieve maximum re-usability of OA data for developers of third-party applications and services [14]. Such applications and services may include statistical analyses beyond those in the scope of OA itself, efforts aggregating OpenAIRE and other data such as research data, or tools that support scientific writing and communication, e.g. online collaborative editors. To this end, we aimed to exploit the interlinked metadata of OA LOD in plugins for online collaborative editors to provide recommendations for authors of scientific papers. In the remainder of this section, two example scenarios are discussed in more detail to demonstrate our approach.

5.1 Look-Up Publications to Cite

The process of generating citations is too time consuming using state-of-the-art editors such as Fidus Writer[8]. Citations are created manually either by entering metadata such as author names, publication titles, etc., or copied from an existing BibTeX snippet. An application plugin to simplify the frustrating citing process can support researchers by instantly generating all required and possible citations.

[8] https://www.fiduswriter.org.

We implemented this plugin as a modal dialog window (jQuery/UI) [4]. Consider the following example scenario: *Suppose a researcher wants to* ***cite*** *a publication. He cannot remember the full information of that publication but just its partial title, which contains:* ***'opencourseware observatory'.*** Our implementation supports this in the following steps: (1) the user can select the desired type of research output (Publication or Dataset) from a drop-down menu (Fig. 2A), (2) the user can perform a search based on different attributes, e.g., publication title, author name or publication year (Fig. 2B), (3) the user specifies the selection of the corresponding text, i.e., here, 'opencourseware observatory', in the search field (Fig. 2C). (4) From the results suggested, the user selects the desired one to insert into the text (Fig. 2D).

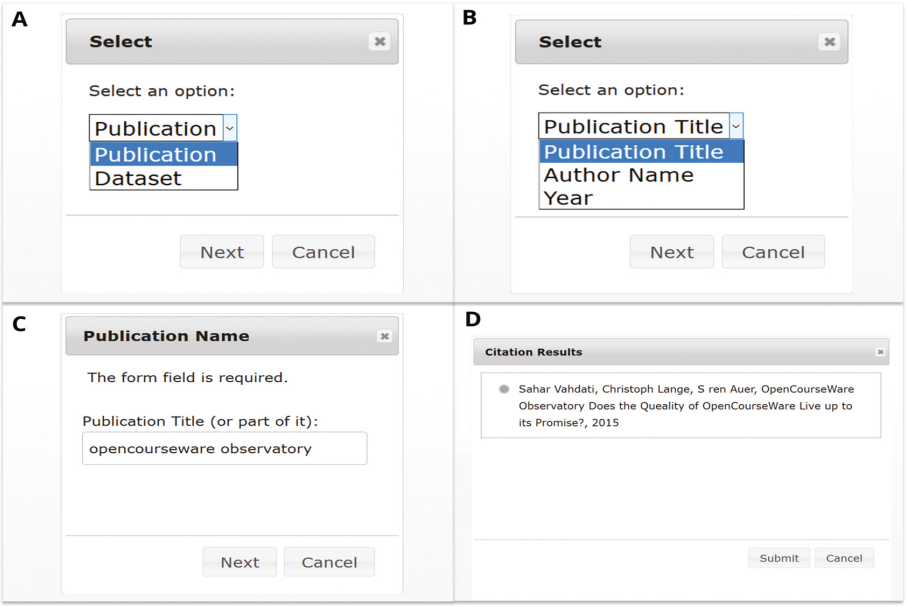

Fig. 2. Looking up Publications to Cite

5.2 Look-Up Author and Statistics

For researchers and publishers, it is important to find publications, authors, journals or conferences related to their research area. However, most of the time it is difficult to find this information in the enormous amount of data on the Web [13]. When users run multiple queries over the most popular data sources for their research fields, the results will not be connected with each other. Thus, our motivation is to develop a plugin that not only retrieves and visualizes data from the OA dataset, but also finds and displays related objects that may be of interest to the user, obtained from interlinking with information from various other online

resources, such as DBLP. Furthermore, we explore possibilities for presenting related data in a useful manner (e.g., using statistical analysis); cf. [13]. This plugin provides the following features:

- Perform a search based on author name
- Retrieve and visualize the author's information obtained from OA dataset
- Find further information by following links from a search result to other datasets, e.g., DBLP
- Display statistics for a certain type of information, e.g., an author's number of publications per year, or co-author relationships.

We implemented a modal search dialog, which enables users to run keyword searches (Fig. 3A). By forming the query with a part of an author name and selecting the desired person (Fig. 3B), our plugin yields the following results (Fig. 3C):

- list of publications and year of publication for each author
- list of co-authors
- statistical graphs based on the above results

Moreover, we utilize links to external datasets such as DBLP, SWDF, and enrich our result with information from those datasets (Fig. 3D).[9]

6 Evaluation

6.1 Evaluation of Interlinking Tools

To find the common and individual links created by selected interlinking tools, we wrote a script [1, Appendix C], which compares the contents of results obtained by two tools and returns the number of common links and also the number of links found by one tool but not by the other. In an experiment with considering publications of OA data and publications of DBLP data LIMES was able to match 432 entities, i.e. more than Silk. The number of common records discovered by both Silk and LIMES is 358. 74 links were found by LIMES but not by Silk, and 3 links were found by Silk but not by LIMES.

In addition to the number of discovered links, reliability of the obtained links is also important. Thus, to evaluate the quality and reliability of the links obtained via each tool, we created a reference linkset (gold standard) consisting of 100 publication resource selected from OA and by manual research found 38 links to SWDF. We then ran Silk and LIMES to find only links from these 100 selected OA resources to SWDF and then compared their output to the gold standard. We computed precision, recall and F-measure to check completeness and correctness of the links found; Table 3 shows the results. Precision is the ratio

[9] Note that the encoding problem ('Sören Auer' in OA instead of 'Sören Auer' in DBLP) stems from the OA data.

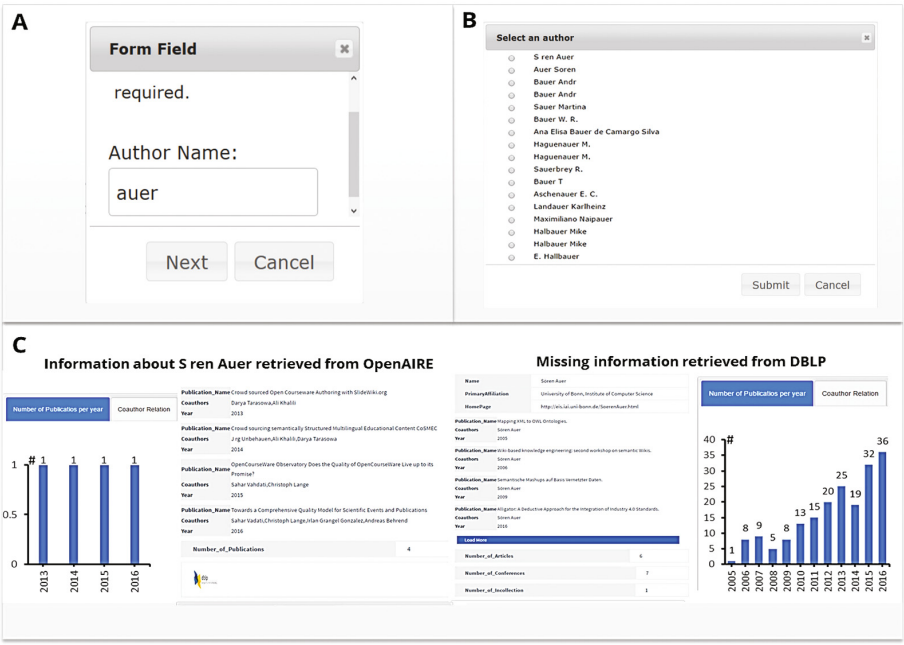

Fig. 3. Author lookup feature's process

of the number of relevant items to the number of retrieved items, i.e.: Precision $= \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$. In our case, this means

$$\text{Precision} = \frac{(\text{Number of created links} - \text{Number of incorrect links})}{\text{Number of created links}}$$ and indicates the correctness of links discovered. Recall is the ratio of the number of retrieved relevant items to the number of relevant items, i.e.:

$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$. In our case, this means

$$\text{Recall} = \frac{(\text{Number of created links} - \text{Number of incorrect links})}{(\text{Number of correct links} + \text{Number of missing links})}$$ and indicates the completeness of links discovered. F-measure is a combined measure of accuracy defined as the harmonic mean of precision and recall, i.e. $\text{F1} = \frac{2 * precision * recall}{precision + recall}$.

The evaluation revealed 9 missing links and one incorrectly discovered link in Silk and 1 missing in LIMES. This corresponded to a Precision of 1, a Recall of 0.97 and an F-measure of 0.98 for LIMES and a Precision of 0.96, a Recall of 0.76 and an F-measure of 0.84 for Silk. The main advantage for LIMES within this small evaluation is the execution time. However we consider the best practices so far which showed that LIMES outperforms Silk dealing with big data. Therefore, due to the fact that we got more relevant, reliable and accurate results from

Table 3. Evaluation of interlinking tools result against a gold standard

Tool	Number of created links	Number of missing links	Number of incorrect discovered links	Precision	Recall	F-measure
LIMES	37	1	0	1	0.97	0.98
Silk	29	9	1	0.96	0.76	0.85

LIMES compared to Silk, we chose LIMES for further interlinking OpenAIRE with other datasets.

6.2 Evaluation of Interlinking Results

We configured LIMES to generate owl:sameAs links between resources with a similarity of above 95%. However, the question is to what extent resources linked in this way are actually the same. Given the size of the linkset, manually assessing and analyzing each link would have been too time-consuming. We therefore picked a number of sample links from each linkset based on its size, aiming at feasibility of a manual inspection (150 samples of publication links, 200 samples of person links and 25 samples of organization links). We then manually verified the correctness of each link and computed precision as 'number of correct links'/'number of sample links'. In the absence of a gold standard, we did not compute recall.

Table 4. Number of inter-links and precision values obtained between OA and DBLP, SWDF, ACM and DBpedia for publications, persons and organizations.

Links between	Target dataset	Target instances	Generated links	Sample of generated links	Verified links	Precision
Publication	DBLP	164890	2276	150	147	0.98
Publication	SWDF	5009	432	150	150	1.0
Publication	ACM	10378	1082	150	136	0.9
Person	SWDF	11184	2000	200	180	0.9
Person	DBLP	932000	6852	200	111	0.55
Person	DBpedia	23373	1088	200	80	0.40
Organization	SWDF	3212	866	30	30	1.0
Organization	DBpedia	3472	38	30	30	1.0

The number of links obtained between OA and DBLP, SWDF, ACM and DBpedia for publications, persons and organizations is displayed in Table 4 along with the precision for each linkset. We obtained high precision in Publication and Organization interlinking, but not in Person interlinking. This is because

initially we carried out Person interlinking by just comparing the names, which was not sufficient, as different persons may have the same name. In future work, we should improve the linking rule for persons taking into account not only their names but also the titles of their publications.

6.3 Usability and Usefulness of Services

We used a custom survey to measure the usability and usefulness of the implemented services discussed in subsection 5.1 (for full details see [1]). 8 participants were first introduced to the idea and the services. We asked them to use the services and figure out the answer of 10 pre-defined questions. Finally, two questionnaires, one for usability and the other for usefulness (10 questions each) were handed out to be filled by them. The questionnaires were designed using System Usability Scale[10]. The results show that most of the participants agreed that our applications are very useful in terms of supporting authors and publishers as well as easy to use an easy to learn. Two of them indicated they needed to learn a bit in the beginning on how the system works. Half of the participants were confident using system and they found it easy to explore. They also mentioned, they would recommend it to experts and use it frequently. Overall, usability of the services is scored as 76.56%.

Satisfaction of the users on usefulness was much higher. Author look up and citation services are selected as a highly useful feature to assist researchers. Three participants were experts of SPARQL queries, however the rest asked for a bit more use-friendly interface both for querying and result representation. 5 of the participants scored the system as easy to use.

7 Conclusion and Future Work

We have presented an approach for interlinking the OpenAIRE research metadata with related Linked Open Datasets, and tools that exploit these new connections to the benefit of end users. After identifying appropriate elements for interlinking, selecting candidate datasets and comparing interlinking tools, we applied the LIMES tool to interlink OpenAIRE concepts to four datasets providing related information (DBLP, DBpedia, ACM, SWDF) and evaluated the precision of the results. We achieved high precision for publications and organizations, whereas the interlinking of persons requires further improvement. Aiming at enhancing the reusability of the interlinked OpenAIRE LOD, we implemented two plugins to assist researchers: a citation lookup service and a tool that looks up statistics about authors. Our usability evaluation suggests that these plugins are easy to use, consistent, adequate for frequent use, and well integrated.

Interlinking OA dataset with other relevant datasets is an ongoing task for the OA LOD team. Deployment of OA interlinking with already examined datasets in the infrastructure of OA is a future work. Based on the current observations,

[10] https://wikipedia.org/wiki/System_Usability_Scale.

we also plan to enhance the interlinking results between OA and other candidate datasets related to other fields such as biology and astronomy and provide a more advanced evaluation. We plan to adopt the implemented services into the infrastructure of the OA and have them publicly available with a better design.

Acknowledgments. This work has been partially funded by European Commission grant 643410 (OpenAIRE) and by DFG grant AU 340/9-1. This work has been partially funded by the European Commission with a grant for the H2020 project OpenAIRE2020 (GA no. 643410) and OpenBudgets.eu (GA no. 645833).

References

1. Ameri, S.: Exploiting interlinked research metadata to provide recommendations for authors of scientific papers. MA thesis. University of Bonn (2017). http://eis-bonn.github.io/Theses/2017/Shirin_Ameri/thesis.pdf
2. Bauer, F., Kaltenböck, M.: Linked Open Data: The Essentials (2011)
3. Hallo, M., Luján-Mora, S., Cháez, C.: An approach to publish scientific data of open-access journals using linked open data technologies. In: EDULEARN (2014)
4. JQuery UI 1.12 API Documentation. https://jqueryui.com/dialog/
5. Purohit, S., et al.: Effective tooling for linked data publishing in scientific research. In: International Conference on Semantic Computing (2016)
6. Rajabi, E., Sicilia, M.-A., Sanchez-Alonso, S.: Discovering duplicate and related resources using an interlinking approach: the case of educational datasets. J. Inf. Sci. **41**(3), 329–341 (2015)
7. Rajabi, E., Sicilia, M.-A., Sanchez-Alonso, S.: Interlinking educational resources to web of data through IEEE LOM. Comput. Sci. Inf. Syst. **12**(1), 233–255 (2015)
8. Rajabi, E., et al.: Interlinking educational data to web of data. In: Big Data Optimization: Recent Developments and Challenges (2015)
9. Rettberg, N., Schmidt, B.: OpenAIRE supporting a european open access mandate. Coll. Res. Libr. News **76**(6), 306–310 (2015)
10. Rodichevski, A.: Approximate String Matching Algorithms. http://www.morfoedro.it/doc.php?n=223&lang=en
11. Simperl, E., et al.: Combining human and computation intelligence: the case of data interlinking tools. Int. J. Metadata Semant. Ontol. **7**(2), 77–92 (2012)
12. Stackoverflow: Difference between Jaro-Winkler and Levenshtein distance. http://stackoverflow.com/questions/25540581/difference-betweenjaro-winkler-and-levenshtein-distance
13. Uyar, E., Brehmer, S., Athamnah, M.: Accessing, Analyzing and Linking Data from DBLP with other Internet Resources (2013)
14. Vahdati, S., et al.: LOD Services. Deliverable D8.2. OpenAIRE2020 (2015)
15. Vahdati, S., Karim, F., Huang, J.-Y., Lange, C.: Mapping large scale research metadata to linked data: a performance comparison of HBase, CSV and XML. In: Garoufallou, E., Hartley, R.J., Gaitanou, P. (eds.) MTSR 2015. CCIS, vol. 544, pp. 261–273. Springer, Cham (2015). doi:10.1007/978-3-319-24129-6_23. arXiv:1506.04006 [cs.DB]