Limani, Fidan; Latif, Atif; Tochtermann, Klaus

Conference Paper — Accepted Manuscript (Postprint)

Scholarly Resources Structuring: Use Cases for Digital Libraries

Suggested Citation: Limani, Fidan; Latif, Atif; Tochtermann, Klaus (2019) : Scholarly Resources Structuring: Use Cases for Digital Libraries, In: Garoufallou, Emmanouel Fallucchi, Francesca De Luca, Ernesto William (Ed.): Metadata and Semantic Research (MTSR 2019), Springer, Cham, pp. 248-259,
https://doi.org/10.1007/978-3-030-36599-8_22 ,
https://link.springer.com/chapter/10.1007%2F978-3-030-36599-8_22

This Version is available at:
http://hdl.handle.net/11108/426

Scholarly Resources Structuring: Use Cases for Digital Libraries*

Fidan Limani[1](✉), Atif Latif[1], and Klaus Tochtermann[1]

Leibniz Information Centre for Economics, Kiel, Germany
https://www.zbw.eu/
{f.limani, a.latif, k.tochtermann}@zbw.eu

Abstract. With the growing generation of links between scholarly resources, information infrastructures such as Digital Libraries (DL) are compelled to explore the potential of data (re)usability. Coupled with the need for increased (research) transparency and reproducibility, linked scholarly resources offer major convenience to researchers in their daily research work. In this way, it is easier for them to get the different research artifacts – be it publication, dataset, workflow, etc. – that form the complete research picture and (re)use any/all of its parts in their work. In this paper, we explore the potential from harnessing such links for a DL environment, model them based on an emerging standard, and represent and publish them via the Semantic Web technology stack. Moreover, to highlight our unique approach for realization of scholarly link collection, we present few use cases to illustrate the potential for a DL environment. Through this study we claim that by adoption of links as new resources, DLs can extend their collection and/or services for their users.

Keywords: Scholarly links · Digital Library · Semantic Web.

1 Introduction

Research data (RD) is shaping into one of the emerging scholarly research artifacts with considerable traction in the research community. The RD data generation and sharing potential [1] is putting it on par with traditional research artifacts, such as research publications. Moreover, it is becoming more common for scholarly-relevant organizations, such as government agencies, funding bodies, research events (conferences, journals, etc.), to accept and disseminate RD in addition to (or even independently, as is the case with data papers) research publications. Additionally, initiatives[1] that increase RD impact and enable the development of accompanying services and its (re)use are also becoming a common research practice.

Sharing RD is meant to support reproducibility and provide reusability [2]. However, to get the real impact of RD in sense of reuseability, it is critical to

* Leibniz Information Centre for Economics
[1] FAIR principles: https://www.force11.org/group/fairgroup/fairprinciples

accept and promote RD along with other notable complementary research artifacts. Research publications, as one such resource type, could provide the context required to reproduce and reuse RD, as well as use RD in ways not originally envisioned by the authors. This is but one example that demonstrates the benefits of providing relevant resource artifacts as a linked, research bundle; other research artifacts can similarly provide benefits when combined with relevant research resources.

In parallel to researchers as producers of RD, changing research practices are affecting the expectations that users (as consumers) have for information infrastructures such as Digital Libraries (DL). As new research artifacts become available, DLs have an opportunity to strive for a more comprehensive research picture that includes different aspects of a research besides publications, as well as explore new use cases and the benefits that this brings.

Due to the prominence of RD, having already explored scenarios of integrating heterogeneous resources with research publications and scientific blogs (see [3]), we now turn to publications-to-data links as complementary resources for DLs. Our aim in this paper is to structure links via suitable vocabularies and, as a follow up, explore potential use cases for DLs based on these link collections.

2 Research Motivation and Use cases

With the emergence of new research artifacts, users require a more holistic view of a research body of work. In this fashion, the general motivation for this work is to provide users of a DL with the experience of a single, complementary, research bundle, provided in a one-stop shop fashion. We focus on research publications and RD as key components of this research unit.

2.1 Research Artifact Links

Providing links between research resources as a way to enable reproducibility and credit researchers for sharing their data (see [4]), to mention but a few, is already recognized and supported in the research community. This enables the development of services based on the links between different research resource types (publications, data, software, etc.). Figure 1 illustrates an instance of a link model that we adopt to represent and base our use cases on. The link instance can bring different aspects of a research into (a single) view; model extensions as new research artifacts become available is also possible.

2.2 Research Motivation

Motivated by data reuse or reproducibility, in a typical scenario, given a publication, the user wants to also access the supplementing RD. Note that the reverse discovery path could also work: given a RD, she would try to identify research or data paper(s) that help contextualize the dataset for a better understanding of its applicability, limitations, etc. Moreover, in order to apply the existing or test

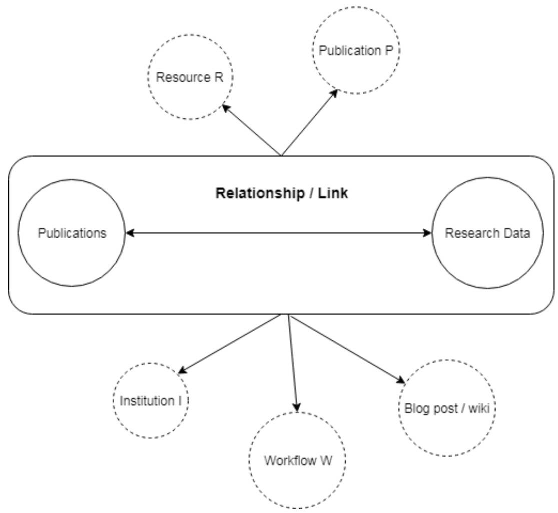

Fig. 1. A link model bringing multiple research artifacts together

a new approach to a wider RD collection, she might be interested in additional RD collections.

An environment that handles different research artifacts could rely on the links between artifacts to provide a complementary research picture to its users. Such complementarity is what we want to enable by providing a semantic structure to the links.

The research motivation, then, requires that we support (meta)data integration of heterogeneous collections (of research artifacts), including publication, RD, and links between them, and represent them in a common model with a query capacity to support the use case scenarios, part of which are demonstrated in section 6.

2.3 Motivating Technology

Linked Data (LD)[2] provides a conceptual and technological fit for our research undertaking – typifying the links between scholarly resources. Based on established vocabularies, links are structured to provide precise meaning. Moreover, being represented in RDF, the model is easily extendable, as new attributes that describe this research context evolve, or new RDF vocabularies become available. Lastly, since it is to be expected that research resources be represented differently across projects, LD is especially handy for (meta)data integration of heterogeneous data sources. In the modeling effort of the link collections, we will consider a variety of vocabularies that reflect the metadata requirements, and are well-established, open, well maintained and documented.

[2] http://linkeddata.org/

3 Related Work

The availability of different research artifacts is presenting new opportunities to scholarly research infrastructures and many initiatives are already under way to provide a more complete (and complementary) research picture by bringing these deliverables together. The research resources linking is just one approach to this effect.

Mayernik et al. [5] report on the challenges and opportunities for linking resources across institutional repositories. Burton et al. [6] present the Scholix initiative: a framework that supports linking resources between providers (hubs) of scholarly literature. In another work, Hoekstra et al. (see [7]) explore linking from FigShare[3] articles to external resources, such as DBpedia, DBLP, etc., and publish the links as Linked Open Data (LOD). At a more general level, projects like Research Object[4] and RMap[5] bring research deliverables of different types in a common unit that parties can act upon. In this way, they recognize and handle a broader research perspective via its corresponding artifacts, from workflow to software to presentation slides, etc., as means to provide a richer scientific context for users. Moreover, being extensible enables them to accommodate new artifacts, depending on the requirements (see [8] for such an example).

Kramer et al. (see [9]) focus on relating semantified (RDF represented) datasets to relevant resources (publications, organizations, studies, people, etc.) in the domain of social sciences, and describe 5 use cases that benefit from this undertaking. Moreover, Wiljes et al. (see [10]) apply Linked Open Data principles to represent the research data artifacts of an institutional repository. This provided an effective approach to handling RD heterogeneity, RD contextualization (considering available institutional publications), and enrichment capabilities to external collections (such as DBpedia). Relying on the (Semantic) Web technologies, Kauppinen et al. [11] introduce a vocabulary (the Linked Science Core Vocabulary) that enables structuring research resources (data, publications, workflows, processes, etc.) to be better (semantically) represented and used (accessed, referenced, linked, etc.). Finally, Fathalla et al. [12] in their work bring fine-grained access to the constituting parts of survey articles, as one of the research output in scholarly communication. Via an ontology designed for this purpose, they show the benefits for researchers during the literature search on a certain topic.

4 Dataset Selection

In this section, we present the link collection from the 2 sources selected for this paper; we further describe the resources in terms of their metadata features required to denote these resources in a common representation model. In selecting the dataset we were guided by 2 aspects: 1) dataset that supports use cases

[3] https://figshare.com/
[4] http://www.researchobject.org/
[5] https://rmap-hub.org/

important to DL and 2) one that provides a large collection to support various scenarios. In following, we briefly present both link collections:

1. Intra-institutional link collection: Modest in size, consists of resources from a single domain (economics) – originally not linked with each other – part of the same DL ecosystem and governed by a single institution. This selection enables us to explore use cases that especially exploit the benefit of linked resources in a DL. In this case, we rely on a link collection that we create for the purpose of this paper; it involves publications and datasets as resources for the link.
2. Public link collection: A richer collection (cross-domain, large in size, etc.) is especially important to study research aspects or benefits that generalize over different DL settings. Such a link collection that spans multiple research domains and enables us to implement interesting use case scenarios – cross-disciplinary ones included, typically out of scope for individual DLs, is what we rely on for this dataset selection.

Following is a brief description of both collections together with the rationale for using them in this work.

a) DL publication-to-data links
The ZBW[6] has two subject portals that deal with publications and research data, respectively. Although part of the same institution, there are no established links between resources from the two collections. Researchers are encouraged to submit their publications and data in these repositories, but there is no (explicit) linking of the two required or provided. We apply a simple approach to establish links between these resources automatically, and use them to demonstrate one of our use cases in this paper. Lets see a short description of each subject portal, and the linking approach for these two collections:

– EconBiz[7]: a subject portal focusing on publications from the domain of economics and business administration. Its collection consists of many types of publications, such as conference or journal papers, books chapters, master and PhD thesis and working papers, etc. Currently the collection stores more than 10 million publications across participating databases, with a minimal collection of datasets (not considered in our experiment), accompanied with a set of services for discovery and recommendation to support researchers.
– JDA[8]: targets RD from journals in the domain of economics and management, including different formats: PDF, text, tabular, scripts or implementation code, etc. As a service, it supports journals as a platform for storage, dissemination, and access control for their datasets. At the time of harvest the JDA collection contains 106 datasets from 6 different journals.
– Publications-to-Dataset linking: Our aim is to link publication and dataset that stem from the same research work. In case an author has a publication

[6] https://www.zbw.eu
[7] https://www.econbiz.de/
[8] http://www.journaldata.zbw.eu/

in EconBiz and a dataset in JDA, we do the matching based on the degree of overlap between the publication and dataset title (often, a replication data/code/script", etc., string is added to the dataset title). The result from our simple matching processes resulted with 70 JDA entries to EconBiz publications, thus this represents 70 data-to-publication links to use.

b) Public link collection: The Scholexplorer Service
For the purpose of this work we use the link collection from OpenAIRE's Data Literature Interlinking service, Scholexplorer[9]. This service currently interlinks more than 1.3 million publications and 8.2 million datasets, all via more than 56 million bi-directional links, spanning multiple disciplines. It is worth noting that there are two types of links stored in this collection: dataset-to-publications and dataset-to-dataset links, from different data providers, all modelled according to common link metadata schema, which we introduce later in the paper.

5 Domain Modeling: Publications, Datasets and Links

Many initiatives that model research resources linking are emerging (see Section 3), and based on our resource features we choose Scholix[13] as a representation framework. In this section we briefly present the metadata requirements of our 2 datasets, and proceed to represent the resulting model via existing (machine-readable) structured vocabularies for an even wider access and distribution.

5.1 Publications and Datasets: Metadata Requirements

In terms of metadata, research resources span from having minimal to extensive metadata descriptions. As a result, when modeling the resources there is a need to balance between use case requirements (based on available metadata) and model capability. At times we struggle with providing the minimum-required metadata, and at times we have to leave certain elements out of the model in order to reach this balance. Next, we describe the decisions which we took for our datasets modeling and links collection.

a) Publications and Datasets: Metadata features
This selection contains the common descriptive metadata, such as title, creators, identifier, publisher, publication date, license information, etc. For practical reasons – no immediate support for current use cases foreseen, for example – we leave few elements that can be found in both collections out of our metadata model. While important to the respective communities, these elements represent fine-grained descriptions and fall outside our current research scope. The metadata features that we consider from both collections determine the use case scenarios that are able to be implemented; as more metadata become available for both resource types, the number of possible use cases will increase correspondingly. In addition to the properties in Scholix, we also include:

[9] http://scholexplorer.openaire.eu/index.html

– Subject term: denotes the subject of a resource; it presents a terminology linking capacity for our datasets – a nice feat to explore use cases that involve both research publications and RD.
– Number of files (applies to RD only): designates the number of files that constitute a dataset. We identify use case scenarios where such aspect is important to a user.

When determining the final set of metadata to consider, we balanced between getting as close as possible to a standardized model, as well as selecting metadata specific for the disciplines and communities. The former provides modeling breadth, especially when considering future dataset extensions, whereas the latter provides modeling depth to the use cases we implement.

b) Links: Metadata description approach
We model the links based on the **link information model** from Scholix[10]. The model captures common attributes for research resources (publications and datasets) and links between them, which makes it relatively easy for communities to apply (see Table 1). The properties in bold are mandatory, whereas the rest are optional – a good (minimalist) take for available metadata that would adopt this model.

Table 1. Link and Resource properties from Scholix model

Link	Resource (source and target)
Link Publication Date (1)	**Object Identifier (1)**
Link Provider (1..N)	**Object Type (1)**
Relationship Type (1)	Object Title (0..1)
License URL (0..1)	Object Publisher (0..1)
	Object Creator (0..N)
	Object Publication Date (0..1)

Let's briefly treat the **link** attributes of this model, which is different from the resources it links (publications and datasets): date of link publication and its provider(s) (there can be more than 1 provider for a link) are self-explanatory; relationship type of the link specifies the nature of the resources being linked (does one derive from, cite, is part of, etc., the other resource?); license URL provides license information for the link (excluding the resources being linked). The link attributes could provide different cases for users, such as data provenance and information quality (depending who the provider is), certain relationship of linked resources, licensing arrangements, etc.

The links in the model are one-directional, and we rely on the Relationship Type property for that. The values for these properties are adopted from Dat-

[10] http://www.scholix.org

aCite[11]s controlled vocabulary for its `relationType` sub-property. Before Scholix v3, there was an Inverse Relationship Type property included in the schema to denote a bi-directional link, but an ontology that describes relationships could easily infer and support such a feature for a link if necessary.

We largely adopt Scholix's information package, with a slight extension based on the requirements discussed in a), which pertains to resources attributes. We next represent the resources from our collections into something more semantical.

5.2 Getting Semantical: RDF Modeling

As described in Section 4, metadata attributes of the resources we deal with publications, datasets, and links are generally of descriptive nature, which supports functionalities such as discovery, (resource) identification, etc. Since we focus on (semantic) structuring of links, our goal was to reuse established vocabularies instead of creating ones from scratch. Due to the availability and maturity of vocabularies that describe datasets and publications, it was straightforward to select and structure (in RDF) these resources; for the specifics of link structuring, starting with the link itself, i.e., denoting that a resource is of type link, we combined few vocabularies to support the link metadata attributes proposed by the Scholix framework.

Regardless of resource type, the common, general attributes are captured well by the Dublin Core Metadata Initiative (DCMI)[12] and the Bibliographic Framework Initiative[13]. The former provides for usual descriptive metadata properties such as title, publisher, creator, subject, date, and size duration; whereas the latter enables to specify the type of the resource – publication or dataset. Note that the size and duration property from DCMI terms supports the description of the dataset in terms of size, as introduced as a requirement in Section 5.1 a).

Some of the resources in our collection contain structured identification attributes, including the scheme, identifier, and the role it plays in an identification scenario (primary vs alternative resource identification, for example). DataCite ontology [14], developed based on the metadata standard of the same name, addresses the identification aspects for all the resources of our model.

In addition to the descriptive vocabularies listed above, the Citation Typing Ontology[14] is used to represent the "link" part of the model. Its properties enable us to define the citing and the cited entities, as well as the relationship that these entitites are linked by, such as "cited/cites as data source" or "cited as related", etc. Its additional attributes for publication date and license are based on DCMI, whereas the Europeana Data Model [15] supports the link attribute required to define the entity that provides the link.

There is always the option to extend existing vocabularies/ontologies or develop a custom one for the problem at hand. However, amid multiple competing/different link models available, a standardized version has yet to come. Thus,

[11] https://schema.datacite.org/
[12] http://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[13] https://www.loc.gov/bibframe/
[14] https://sparontologies.github.io/cito/current/cito.html

we rely on existing vocabularies that not necessarily are conceptualized with linking research resources as a key driver, but nonetheless are a good fit to link model requirements.

6 Links harvesting and storage: The Workflow

This section describes the workflow which we made use of in this study. It is a three stage approach which starts with resources harvesting, modeling and conversion to RDF, and ends with storage of resulting links (see Figure 2). In order to share the technical environment and processing of our datasets, following is a brief description of the activities that constitute all the parts of the workflow.

- **Link collections**: We establish the intra-institutional link collection (EconBiz-JDA) via the REST API of the corresponding repositories. For the Scholexplorer collection we use the (JSON) data dump available in the Zenodo research repository [16] (the Scholexplorer REST API was unavailable at the time of harvest). Due to its large size, the collection is organized in several (compressed) files that contain link batches of approximately 2.5 GB each, of which we include only a subset sufficient for this work.
- **Link harvesting**: For the generation of first link collection, we search the EconBiz collection for a match (mainly based on the title) with every dataset available in JDA. With this approach, we were able to identify 70 links. Scholexplorer, on the other hand provides the bulk of the links and is more straightforward to harvest. The subset that we use consists of over 2.3 Million links.
- **Link modeling and storage**: We convert link collections to RDF based on the vocabularies presented in 4.2. For the RDFizing process we rely on Apache Jena[15]. We store the datasets in separate named graphs, as this provides easier future management of the collection, such as update, maintenance, but also provision of a more granular access.

7 Use cases: Explored scenarios

In general terms, the use cases revolve around (fine-grained) search and, as applicable, involve different metadata elements that describe links, publications and research data, such as: publication date, resource provider, persistent identifier, resource type of the resource and dataset size, etc.

From publications to data Find RD that support or are relevant to a publication. This can be a data paper or data that revolves around the subject of the paper to certain extent (a narrower, related, etc., subject term).

We start with the scenario where the user retrieves the RD that directly support (as a primary source of data) the research paper at hand. If the user

[15] https://jena.apache.org/

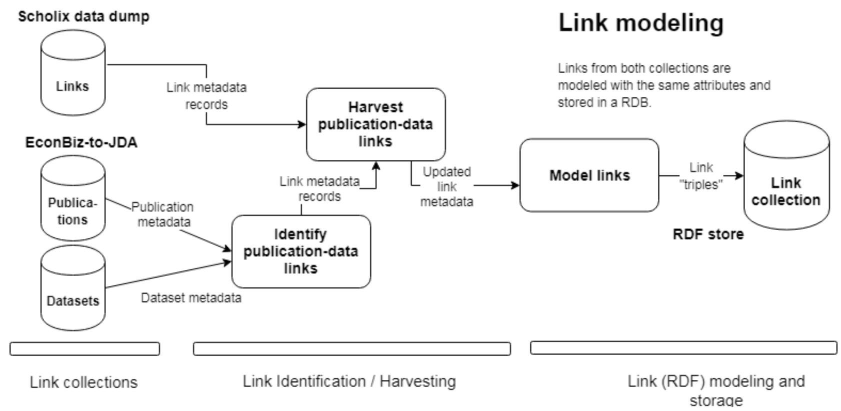

Fig. 2. From datasets to central link collection: a workflow

wants to further specify the result, she can refine the query to include the most re-used RD in the collection (based on the number of times it has been cited). On the other hand, if the results are scarce or the user wants to broaden the search, she could also retrieve all the RD by the same author of the paper. In another scenario, a user can retrieve the "trending" RD (RD being cited the most in a more recent time frame) – and their corresponding publications – for a quick impression of what her community is currently working on. In a final scenario for this part, the user can rely on the "subject terms" to search for a field of interest across link providers for a more interdisciplinary search scenario (search for a fish type to see its fishing quotas, market fluctuations in a certain period, as well as the impact from climate conditions on its habitat).

Importantly, the DL collection we are working with primarily supports publications, thus in presented scenarios we assume the user first selects a publication and then proceeds to find RD. This aspect can easily be reversed (start with RD of interest, and find publications and/or RD) and practical to the DL especially if its collection reaches a critical mass of RD.

From data to (even more) data: In this set of scenarios, the user starts with a RD and wants to find relevant/related RD within and across research domains.

Having collected (a set of) RD, a researcher wants to identify related RD from the same subject (or field) that she could re-use (for reproducibility) or combine it with her own or additional RD for new research purposes. In this case, the "subject term" metadata element plays an important role as it enables us to search across collections from different institutions that provided their links (and consequently, publications and RD). As mentioned, the available metadata in our collection provide many filtering capabilities for this scenario such as:

restricting links based on certain time frame, the type of resources they linked to and the institution that publishes them, etc.

Moreover, users want to see which are the related disciplines of the "trending" RD. Usually every linked resource has a metadata about the entity that cover certain research fields. In this way, a researcher who wants to search for RD of interest from few research fields, such as economics and social sciences, can explore what "trending" RD from one discipline could be matched with others in the other discipline.

Generic use case scenarios: This set of scenarios provides more general information, which, although not the first use case of choice, could turn useful to the researchers; examples include:

- List resources that are linked by the same publisher, publication date, domain, and other relevant metadata.
- Based on links that cite my research artifacts (publications or RD), who is using my RD? In what scenarios and context (information you get after reading a citing paper or RD, for example)? This question would apply both to individual researchers and institutions.
- Show me the potential of relevant resources based on a certain criteria such as: classification terms for the subject of coverage, resources type, number of files that constitute a resource, etc.

We used Apache Jena framework (and its Fuseki server) to realize the mentioned use cases (either programmatically, or via its SPARQL editor); for the reproducibility of use case scenarios, we provide RDF dump to interested parties.

8 Conclusion

In this work we showcased the links between scholarly resources as value drivers to information infrastructures such DLs. We harvested more than 2.7 million links from 2 different link collections, and structured them in a common representation model via Semantic Web technologies. As a result, we were able to explore many use case scenarios that fit in a DL environment.

With the initial results in, we plan to test our workflow with the complete link collection from Scholexplorer, as well as other available collections. Another follow up includes enrichment of links and resources being linked for a richer research/knowledge context for users. An issues that we identified during our work was the metadata inconsistencies for the different resources (identification schemes, metadata variety, etc.), which we need to consider especially when expanding our scope (and RDF link collection).

In our future work, in addition to publications and datasets, we plan to package resources via linking. Moreover, being that we use a graph representation for the harvested links, we would like to experiment with alternative graph representation strategies, such as the Label Property Graph (LPG), and explore more analysis-driven scenarios over the resulting link collection. These analyses are especially important as the link collection grows and includes new resource types. Finally, during the RDF modeling part of the work, at times, it felt like there is a lack of an ontology that would represent the link model adopted in the study, and we see this as a beneficial follow up to this work.

References

1. Christine L Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078, 2012.
2. John Kratz and Carly Strasser. Data publication consensus and controversies. *F1000Research*, 3, 2014.
3. Fidan Limani, Atif Latif, and Klaus Tochtermann. Bringing scientific blogs to digital libraries. In *WEBIST*, pages 284–290, 2017.
4. Adrian Burton, Hylke Koers, Paolo Manghi, Sandro La Bruzzo, Amir Aryani, Michael Diepenbroek, and Uwe Schindler. The data-literature interlinking service: Towards a common infrastructure for sharing data-article links. *Program*, 51(1):75–100, 2017.
5. Matthew S Mayernik, Jennifer Phillips, and Eric Nienhouse. Linking publications and data: Challenges, trends, and opportunities. *D-Lib Magazine*, 22(5/6):11, 2016.
6. Adrian Burton, Hylke Koers, Paolo Manghi, Markus Stocker, Martin Fenner, Amir Aryani, Sandro La Bruzzo, Michael Diepenbroek, Uwe Schindler, and C Authr. The scholix framework for interoperability in data-literature information exchange. *D-Lib Magazine*, 23(1/2), 2017.
7. Rinke Hoekstra, Paul Groth, and Marat Charlaganov. Linkitup: Semantic publishing of research data. In *Semantic Web Evaluation Challenge*, pages 95–100. Springer, 2014.
8. Markus Stocker. From data to machine readable information aggregated in research objects. *D-Lib Magazine*, 23(1):1, 2017.
9. Stefan Kramer, Amber Leahey, Humphrey Southall, Johanna Vompras, and Joachim Wackerow. Using rdf to describe and link social science data to related resources on the web. 2012.
10. Cord Wiljes, Najko Jahn, Florian Lier, Thilo Paul-Stueve, Johanna Vompras, Christian Pietsch, and Philipp Cimiano. Towards linked research data: An institutional approach. (994), 2013.
11. Tomi Kauppinen, Alkyoni Baglatzi, and Carsten Keßler. Linked science: interconnecting scientific assets. pages 383–400, 2016.
12. Said Fathalla, Sahar Vahdati, Sören Auer, and Christoph Lange. Towards a knowledge graph representing research findings by semantifying survey articles. In *International Conference on Theory and Practice of Digital Libraries*, pages 315–327. Springer, 2017.
13. RDA/WDS Scholarly Link Exchange Working Group. Scholix metadata schema for exchange of scholarly communication links (version 3). *D-Lib Magazine*, 2017.
14. Silvio Peroni, David Shotton, Jan Ashton, Amy Barton, Egbert Gramsbergen, and Marie-christine Jacquemot. Datacite2rdf: Mapping datacite metadata schema 3.1 terms to rdf, Feb 2016.
15. Antoine Isaac et al. Europeana data model primer. 2013.
16. Sandro La Bruzzo and Paolo Manghi. Openaire scholexplorer service: Scholix json dump, March 2018.