Limani, Fidan; Latif, Atif; Tochtermann, Klaus

Article — Accepted Manuscript (Postprint)

Links between research artefacts: use cases for digital libraries

International Journal of Metadata, Semantics and Ontologies (IJMSO)

Suggested Citation: Limani, Fidan; Latif, Atif; Tochtermann, Klaus (2021) : Links between research artefacts: use cases for digital libraries, International Journal of Metadata, Semantics and Ontologies (IJMSO), ISSN 1744-263X, Inderscience, Cointrin-Geneva, Vol. 15, Iss. 2, pp. 133-143, https://doi.org/10.1504/IJMSO.2021.120285

This Version is available at:
http://hdl.handle.net/11108/506

Links between research artifacts: Use cases for Digital Libraries

Fidan Limani, Atif Latif and

Klaus Tochtermann

Abstract. The generation and availability of links between scholarly resources continue to increase. Initiatives to support it – both in terms of a (standard) representation model and accompanying infrastructure for collection and exchange – make this emerging artifact interesting to explore. Its role towards a more transparent, reproducible, and, ultimately, richer research context, makes it a valuable proposition for information infrastructures such as Digital Libraries. In this paper, we assess the potential of link artifacts for such an environment. We rely on a public link collection subset (>4.8 M links), which we represent based on the Linked Data approach that results with a collection of >163.8 M RDF triples. The incorporated use cases demonstrate the usefulness of this artifact in this study. We claim that the adoption of links extends the scholarly data collection and advances the services a Digital Library offers to its users.

Keywords: Research artifact links · Digital Library · Semantic Web.

1 Introduction

One of the areas that the technology (r)evolution affects is that of scholarly research and communication. This materializes with an ever-increasing set of research deliverables or artifacts – research data, source code, scientific workflows, open notebooks, etc. – becoming part of scholarly communication, as well as multiple dissemination channels and means of access for them are getting very relevant. The scope of these changes is quite broad and potentially includes all the phases of the research lifecycle, practically in an "as it happens" fashion (**priem2013**). As a result, a variety of artifacts is created during research that, depending on the community research requirements or incentives for any given artifact, are being published, curated, and archived. Their publication is either in the context of a more established artifact, such as a research article, or that of a self-sufficient one that "exists" on its right and serves an independent purpose, such as a data paper.

Open Science is another thread that knits these developments together. It recommends a different set of research practices that rely on current technological trends. Initially focusing only on publicly-funded research, but now being adopted as a practice in the wider community, Open Science models research as a collaborative effort of creation and shared access to its many research lifecycle phases (**openScience2020**). Additionally, initiatives like FAIR

(**fairPrinciples2020**) seek to increase the impact and (re)use of these many research artifacts, to the level that they become part of common research practices. In this way, technology and research practice changes eventually contribute to research reproducibility and transparency.

Libraries are an important component in the scholarly domain infrastructure. Frequently a common first stop for researchers as they provide artifact collections and services in information retrieval operations. Moreover, digital libraries are also considered as one of the key enablers for the Open Science movement (**libOpenScience**; **ogungbeni2018**). Facing these developments, infrastructures that collect scholarly resources need to continuously assess and decide on which artifacts to extend their collection with and which services to develop based on these artifacts to serve their new and/or existing audiences. After all, being able to adapt new research deliverables that can support more transparent, re-usable services for library users and researchers, in general, is always a desirable feature. Thus, demonstrating the value proposition by including research artifacts helps them make (or at least prioritize) this decision.

To this effect, research artifacts of interest need to be recognized and their value should be demonstrated within the scientific community. Such is the case with an artifact that links information about two related resources – publication and datasets. With a representation standard proposed and adopted by major parties, there are already services and the corresponding collection that can be used by other communities. We have made use of this link collection to showcase its value as an additional artifact for DL service.

As new research artifacts become available, libraries have an opportunity to strive for a more comprehensive research picture which would include different aspects of research and explore new use cases that benefit from these aspects. Such integration scenarios of heterogeneous resources – research publications and scientific blogs (**limani2017**), for example – have already been explored. We now turn to *links* between research deliverables as another artifact of interest for libraries. Our aim in this paper is to structure links via suitable semantic models and explore their potential (use cases) alongside existing library collections and services.

2 Research Motivation and Use cases

The increase of deliverables from different phases of the research lifecycle implies that there are many deliverables for the same research undertaking. This output can be used to further understand the ongoing research data activity, extend it, reproduce it, etc. Consequently, users need a more holistic view of a research body of work, an easier way to navigate the outcomes of a project, without limiting their search on what is a common artifact nowadays – the research article.

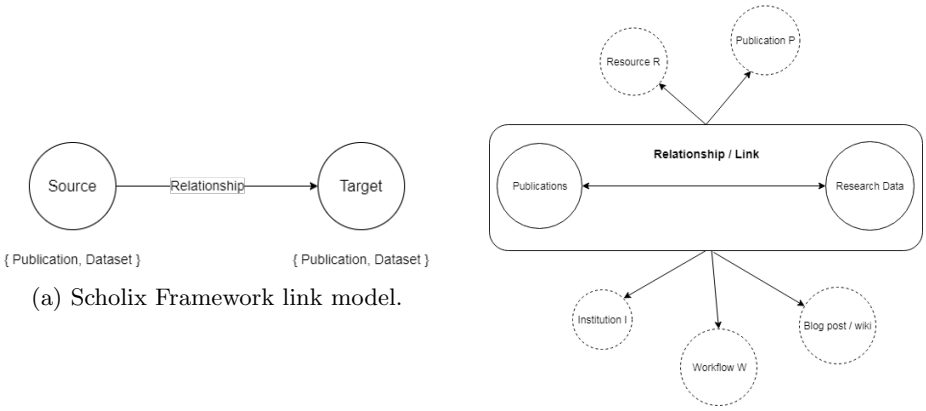
While several initiatives aim to package research deliverables into a single unit, we focus on one that establishes an explicit link between two artifacts.

Specifically, the types of the artifacts in the link, forming this research unit based on this model, currently include articles and datasets.

In this section, alongside the general research motivation, and the rationale for our technology of choice in our approach, we present the new artifact that we want to explore in this study.

2.1 Links as Research Artifacts

Providing links between research resources as a way to enable reproducibility and credit researchers for sharing their data **burton2017b**, to mention but few of the benefits, 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.).



(a) Scholix Framework link model.

(b) Extending the link model.

Fig. 1: Linking research artifacts in a common unit.

Figure 1a) illustrates the link model conceptually (Section 5 provides more details in its representation) that we adopt to base our use cases. The model is quite straightforward: it represents a source artifact that relates to a target artifact, with means to specify the type of both artifacts and the exact way they relate to each other. The link instance brings different aspects of research – in this case that of an article and a dataset – into (a single) view.

2.2 Research Motivation

The motivation for this work is to provide users of a Digital Library (DL) the experience of a "research bundle" or linked research. These research bundles are established based on certain criteria and consists of different research artifacts that are brought together as a single unit or resource. The criteria can assert

different aspects of how the research artifacts relate to each other, such as citing each other, funded by the same organization, being on the same topic, and so on. Furthermore, we want to explore the complementarity of this artifact with existing collections in a DL, namely articles and datasets. Ultimately, we were driven by potential "one-stop-shop" research experience for DL users.

In a typical article search scenario, driven by data reuse or reproducibility, a user may want to also access its supplementing research data (RD). Similarly, a reverse "discovery path" would also work where a user can identify an article or data paper(s) contextualizing that RD. This could help her better understand the applicability, limitations, etc., of the RD of interest. Furthermore, the researcher might be interested in additional relevant RD to experiment with based on the methodology found in the article.

Exploring complementary scenarios for DLs based on the link collection is another motivation. Let's consider a DL that contains research articles, and a separate link collection of articles and datasets. For an article in the DL, we can search the link collection to find (that article) there. If present, we automatically have a dataset (to which that article links) to suggest to the DL user. Such complementarity is what we want to enable by providing a semantic structure to the links.

The research motivation, then, requires that we integrate the metadata of 3 different collections of research artifacts (articles, RD, and links), and represent them in a common model with a query capacity to support the use case scenarios, as demonstrated in Section 6.

2.3 Motivating Technology

Linked Data (LD) (**linkedData2020**) provides a conceptual and technological fit for our research undertaking – that of typifying the links between artifacts. Based on established semantic models – vocabularies and ontologies – links are structured to provide precise meaning. Moreover, being represented in RDF, the model is easily extendable in case it changes or the artifacts being linked evolve. 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 semantic models that reflect the metadata requirements and are well-established, open, well maintained, and documented.

As new research artifacts become available and of interest to be linked, the possibility to model additional aspects of the link model as an extension is also possible. Adopting the LD approach, we would easily accommodate other artifact types, including additional descriptions for the source, target, and the link itself, as shown in Figure 1b).

3 Related Work

The availability of different research artifacts is presenting new opportunities for scholarly research infrastructures. To illustrate it, we selected a subset relevant

to this work (see Table 1) from a more recent period that show a rich and active scholarly communication process. This includes 4 different research artifacts: publications from the engineering domain (adopted from **nsfSRIndicator2020**); dataset published in 5 (generic) data repositories (**assante2016scientific**), links between scholarly resources (from the ScholeXplorer project, which we use in this work); and scientific APIs from the *science* category (**programmableWeb2021**), which further shows the diversity in available scholarly resources. Data on publication trends for all these artifacts is not always publicly or readily available for a time frame of interest, as could be noticed from the table.

Year	Publications	Data	Links	Scientific APIs
2009	1.8+ M	/	/	3
2012	2.1+ M	30.042	/	255
2013	/	111.037	/	80
2014	/	98.003	/	41
2015	2.2+ M	77.063	/	18
2018	2.5+ M	/	62+ M links b/w 9+ M objects	54
2019	/	/	240+M links b/w 17+ M literature objects and 50+ M datasets	40
2020	/	/	445+Mi links b/w 17+ M literature objects and 50+ M datasets	82

Table 1: Recent publication trends for few research artifacts.

Many initiatives provide models for a more complete (and complementary) research picture by bringing the different research deliverables together, and linking research resources is one such initiative. Mayernik et al. (**mayernik2016linking**) report on the challenges and opportunities for linking resources across institutional repositories. Burton et al. (**burton2017a**) present the Scholix initiative: a framework to support linking resources between providers (hubs) of scholarly literature. In another work, (**hoekstra2014linkitup**) explore linking from FigShare (**figShare2020**) 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 (**researchObject2020**) and RMap (**rMap2020**) bring research deliverables of different types in a common unit that interested parties can act upon. In this way, they recognize and handle a broader research perspective and artifacts, from workflow to software to presentation slides, etc., as a means to provide a richer scientific context for users. Moreover, being extensible enables them to accommodate new artifacts, depending on the requirements (see (**stocker2017data**) for such an example).

Kramer et al. (**kramer2012using**) focus on relating semantified (represented in RDF) datasets to relevant resources (publications, organizations, stud-

ies, people, etc.) in the domain of social sciences, and describe 5 use cases that benefit from this undertaking. Moreover, Wiljes et al. (**wiljes2013towards**) 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. (**kauppinen2016linked**) 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, and so on). Finally, Fathalla et al. (**fathalla2017**) in their work bring fine-grained access to the constituting parts of survey articles, as one of the research outputs in scholarly communication. Via an ontology designed for this purpose, hey show the benefits for researchers during the literature search on a certain topic.

4 Dataset Selection

In selecting the datasets we considered 2 aspects: they should support use cases relevant to the DL, and they should showcase the importance of linked artifacts. In this section, we present two such datasets selected for this paper and describe their features, as well as their relevance to this work.

We can categorize the two link datasets as being part of a (1) intra-institutional collection, or (2) a public collection. The former is modest in size and consists of publication and data resources from a single domain (economics). Originally not linked with each other, they are part of the same DL ecosystem and are governed by a single institution. This selection enabled us to explore use cases that are specific for the domain of the DL. The latter is large in size, cross-domain, and contributed to by many institutions. Such a collection enabled us to implement disciplinary and cross-disciplinary use case scenarios. Let's next present some details about both collections, including their usage rationale.

a) Intra-institutional link collection The ZBW (**zbw2020**) has two subject portals that respectively deal with publications and research data. Researchers are encouraged to submit their articles and RD in these repositories, but there is no (explicit) linking of the two required nor provided. We apply a simple approach to establish links between these artifacts. Let's see a short description of each subject portal and the linking approach for these two collections:

- EconBiz (**econBiz2020**): A subject portal focusing on publications/articles from the domain of economics and business administration. Its collection consists of many types of publications, such as conference or journal papers, book chapters, master and Ph.D. thesis, working papers, etc. Currently the collection stores more than 10 M publications across participating databases, with a minimal collection of RD (not considered in our experiment).
- JDA (**jda2020**): 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 provides a platform for storage, dis-

semination and access control for datasets of interested journals. At the time of harvest, the JDA collection contains 151 datasets from 7 journals.
– Publications-to-dataset linking: We aim to link publication and RD that stem from the same research work. In case an author has a publication in EconBiz and a dataset in JDA, we do the matching based on the degree of overlap between the publication and dataset title. The result from our matching processes resulted in 115 JDA entries links to EconBiz publications.

b) Public link collection: The ScholeXplorer Service For this work, we use the link collection from OpenAIRE's Data Literature Interlinking service – ScholeXplorer (**scholExplorer2020**). This service currently interlinks more than 1.3 M publications and 8.2 M datasets, all via more than 56 M bi-directional links. There are two types of links in this collection: RD-to-publications and RD-to-RD ones. This collection is modeled according to a common link metadata schema, which we introduce next.

5 Domain Modeling: Publications, RD and link metadata

Many initiatives that model research resources bundles / linking already exist and new ones continue to emerge (we referred to some of the more established ones in Section 3). We choose the Scholix Framework (**scholix2017**) to model the links derived from the above-mentioned datasets. In this section, we briefly present the metadata characteristics of our link datasets and proceed to represent them based on the capabilities of the Scholix framework.

Across various domains and research practices, research artifacts typically contain different metadata descriptions. As is often the case during modeling tasks, at times we struggle with providing the minimum-required metadata, and at times we have to leave certain elements out of the model to reach this balance. Thus, there is a need to balance this diversity and strive for a common metadata set that is sufficient to support the use case requirements and is well represented in the chosen (link) model. In the following, we discuss the metadata specification that we choose to model with the Scholix framework for our datasets i.e. publications, RD, and links.

5.1 Publication and dataset metadata features

Publications and RD contain common descriptive metadata, such as `title`, `creator`, `identifier`, `publisher`, `publication date`, `license` information, etc. The metadata features that we consider from both collections determine the use case scenarios that can be implemented; as more metadata becomes available for both resource types, the number of possible use cases will increase correspondingly. In some cases, though, we find that metadata used by certain communities (for administrative or other uses) are too fine-grained for the scope of the model. In such cases, we do not model them unless they have immediate support for our use cases.

When it comes to publications and RD, in addition to the resource properties of Scholix collection (see 5.2), we also include the following metadata extension: `subject terms` and `number of files` of a dataset. Subject terms denote the subject of a resource and provide a terminology linking capacity for our datasets – a nice feat to explore use cases that involve both research publications and RD. The number of files metadata element denotes the number of files that constitute an RD, which enables the use case scenario where such an aspect is important to users (for example filtering results).

5.2 Link metadata features

We model the links based on the **link information model** from Scholix Framework. The model captures common attributes for research resources (publications and datasets) and the links between them, which makes it relatively easy for communities to apply. Table 1 lists the requirement of this framework: the properties in bold are mandatory, whereas the rest are optional.

Table 2: 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 the other resource, etc.); license URL provides license information for the link (excluding the resources being linked). The link attributes are important to users and could be used for cases such as data provenance, information quality (depending on who the provider is), relationship nature of linked resources, licensing arrangements, and so on.

According to this model, the link is one-directional (`Relationship Type` property), i.e., from source to destination. Before Scholix v3, there was an `Inverse Relationship Type` property also included in the schema to denote a bi-directional link. This provided the benefit of exploring the links based on two relationship types. For example, if a source `cites` a target, then the target `is cited by` that source, and we could rely on either assertion for our use

cases. Adding the inverse relationship property in the final semantic model, although not conforming to the latest Scholix Framework version, could be easily achieved during the link conversion if there is a rationale for such a use case. For the requirements of this work, as discussed in 5.1, we include two more metadata elements for the resources being linked, in addition to what the Scholix information package offers.

6 Getting Semantical: RDF Modeling

There are two main entities to model in a link: the link itself and the resources (source and target) it links. Most of the metadata for these entities, being of a descriptive nature do support functionalities such as discovery, (resource) identification, etc. Semantic models that can express these metadata are well established and documented, often overlapping, thus "competing" to represent the same metadata element. The challenge here is to choose the most fitting selection among (whenever the case) concurrent alternatives. Semantic models can overlap from the conceptual aspect (represent the same metadata), and/or from the aspect of coverage (support the same metadata to a different extent, i.e., their domain or range) for given metadata requirements. Let us next discuss our choices during link modeling.

The Dublin Core Metadata Initiative (DCMI) **dcmi2021** and the Bibliographic Framework Initiative cover almost all the metadata for a link resource. The former provides most of the metadata properties, whereas the latter covers the type of resource (publication or dataset). We want to note that they overlap with some of their classes: they both support publications and datasets, but the latter supports a larger array of works and is an initiative from the library community for a more web-of-data representation of library catalog metadata (**mccallum2017bibframe**). This makes it more suitable in the context of semantic modeling, as well as for potential future extensions of the link model for the inclusion of new types of scholarly artifacts. To complete the resources semantic representation, the DataCite Ontology (**peroni2016**) covers the resource identifier. On the other hand, when modeling the link itself, the Citation Typing Ontology (**cito2020**) is used. Its properties enable us to define the relationship type between two linked entities. This also includes two relationship types we rely on from the Functional Requirements for Bibliographic Records (FRBR) model (**frbr2020**). The Europeana Data Model (**isaac2013**) models the link provider attribute, which completes the metadata representation requirements of a link instance. Table 2 contains the mapping between the link metadata and the semantic models used.

Another important aspect of the model is that of provenance. Specifically, we capture the following aspects: the workflows we use to do the conversion (`SoftwareAgent` class), the process of generating a new dataset in RDF (`Generation` class), the time the process took for the conversion (`Activity` class), and the resulting (RDF) conversion (`Collection` class). Due to space limitations, we leave the corresponding properties used with these classes out of this part.

Table 3: Classes and properties used to model Scholix links

Vocabulary / Ontology	**Usage**
CiTO (http://purl.org/spar/cito/)	Represents the link itself, technically a *citation* between two resources (`Citation` class) – its source and target (`hasCitingEntity` and `hasCitedEntity` properties, correspondingly), the type of their relation, i.e., does the source cite, support as dataset, etc., the target (`hasCitationCharacterization` property), and the date the link was established (`hasCitationCreationDate` property).
DC Metadata Initiative (http://purl.org/dc/terms/)	Represents bibliographic descriptions of a link, via `creator`, `title`, publication `date`, `publisher`, `license`, including the three additional metadata represented via `subject` and `extent` properties, and the `SizeOrDuration` class.
BibFrame (http://id.loc.gov/ontologies/bibframe/)	Represents the type of the resource being linked. Considering the two types of resources used in our link collections, we rely on its `Publication` and `Dataset` classes.
DataCite Ontology (http://purl.org/spar/datacite/)	Covers two cases of identifiers: DOI-based for which we use the `PrimaryResourceIdentifier` class, and all the other cases for which we use the `AlternateResourceIdentifier` class.
FRBR (http://purl.org/vocab/frbr/core#)	Used to represent two relationship types between linked resources, that of source supporting the target resource (`supplement`), and that of source being supported by target resource (`supplementOf`).
Europeana Data Model (http://www.europeana.eu/schemas/edm/)	The `Agent` class specifies an entity that (establishes and/or) provides (`provider` property) the link. These two elements are used to complete the representation of a link entity.
PROV Ontology (http://www.w3.org/ns/prov)	This ontology provides provenance for the link collection in RDF as a whole, and is not applied to individual links.

However, they can be observed either in our source code or in the generated RDF collection.

With the modeling requirements covered, Figure 2 shows the resulting semantic model describing a link. There you can see the application of the exact classes and properties used to represent the Scholix Framework link model. In it we follow the structure from Figure 1a): on the left-hand side we present the source, then follows the link, which concludes with the target resource. The model is easy to understand, but we just want to mention the `cito:hasCharacterization` property. CiTO represents this property using OWL2 punning; this property is provided as an object of a triple, so that's why we see an oval symbolizing the value for this property in the model. Moreover, under the `frbr:supplement` value, we also listed the rest of the values used for the relationship type throughout the whole link collection.

It is important to mention that in the current semantic web landscape, there is always the option to extend existing semantic models or develop custom ones for the problem at hand. However, amid multiple competing link models avail-

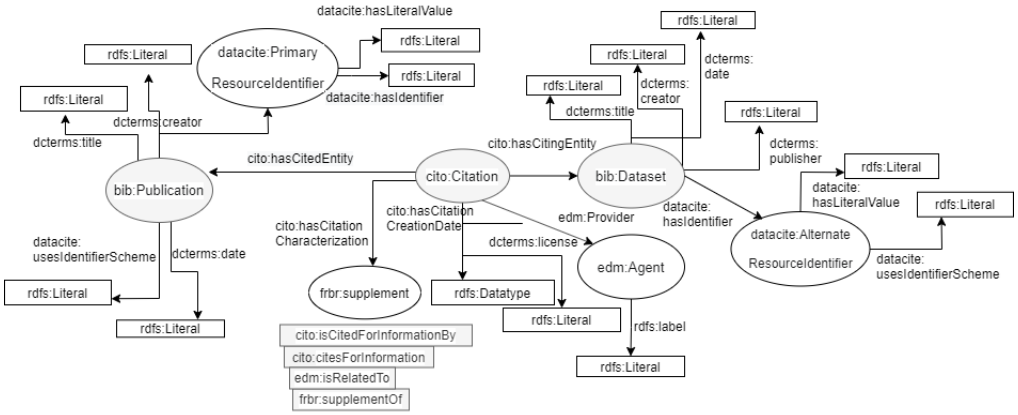


Fig. 2: Semantic model that describes Scholix links

able, a standardized semantic version has yet to come. Thus, we rely on existing vocabularies and ontologies that are not necessarily conceptualized with linking research resources as a key driver but represent a good fit for link model requirements we explore. Moreover, some link collections had the metadata elements that were not envisioned in the Scholix Framework model (dataset size and time, for example). Due to the nature of use cases that these metadata provide in this study, we were able to represent them in the RDF model with relative ease (no pre-existing schema to change, for example). In our case, we provided these additional attributes to some linked resources (datasets), but it just shows how the technology we rely on for the link representation enables model extensions. This is important for at least two reasons: (1) the Scholix Framework that we choose as a model to represent links has a minimal set of mandatory attributes, and (2) links can come from different providers and contain metadata attributes that are not part of Scholix Framework and need to be included in the model.

7 Link harvesting, conversion, and storage: A Workflow

To enable the use cases we discuss in this paper, there is a set of data processing activities that we first implemented. These activities can be grouped in 3 broader categories: (1) access link collection; (2) link harvest (and creation, where applicable), pre-process, and parse; and (3) link conversion to RDF, their storage in a triple store, and RDF data dump generation for later (re)use. Let's briefly present the activities for each category as implemented in the workflow in Figure 3.

Link access We access the datasets in two ways: the EconBiz (for publications) and JDA (for datasets) collections via their corresponding APIs, whereas the ScholeXplorer collection via its publicly-available data dump (**laBruzzo2018**). The latter collection also offers REST-based access, but due to its size and HTTP

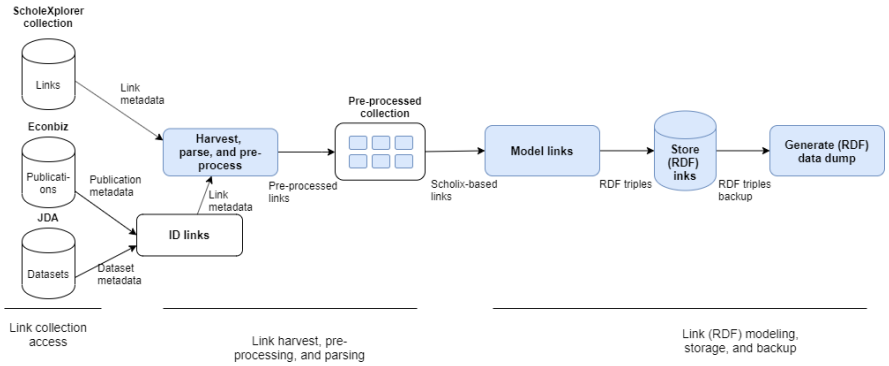


Fig. 3: Link identification, harvesting, conversion and storage workflow

request limitations, harvesting the links from local storage was the most effective implementation approach.

Link harvest, pre-processing and parsing When it comes to the first dataset, it is during this phase that we first identify and establish the RD-to-publication links, and then harvest them. As mentioned in Section 4, although complementary, there are no explicit links between publications and RD in the corresponding repositories. On the other hand, the ScholeXplorer collection comes as a data dump of 30 (compressed) files, of which we harvest only one as this suffices for our use cases.

The number of links that each ScholeXplorer file contains is significant. With 4.2 M links per file on average, this represents a hurdle, especially for the later phases of the workflow, i.e., RDF conversion and storage. We introduced the pre-processing step in the workflow to address this challenge. During this step, one can split the original files into a series of smaller ones that are then easier to process. This is important for those interested to reuse the workflow but do not have access to a more powerful machines to run it. To test this approach, we pre-processed the original ScholeXplorer collection to a set of smaller files, consisting of 300-K-batch links each, and ran the workflow on a personal machine with moderate configuration (i5 CPU, 8 GB RAM, and an SSD HD).

The final task of this phase is that of parsing the link metadata. When considering that different providers contribute their links to the ScholeXplorer collection, without central coordination (metadata standard, allowed values, controlled vocabularies, and so on), differences in metadata provided are to be expected. One such example is the processing of publication dates – both for the links and resources. While parsing the datasets, we encountered and handled more than 30 publication date formats, including values for months in English, German, and French. Rather than ignoring these cases and leaving them out, we opted to parse them all and make them available for our use case scenarios.

Link modeling and storage We convert the link collections to RDF based on the semantic model presented in Section 6. For the RDFizing process, we rely on Apache Jena Framework (**apacheJena2020**); we store the resulting collection in separate named graphs, as this provides for an easier management of the collection (update, maintenance, provision of more granular access, etc.). In our case, we have a named graph for every file of links, which enables us to trace harvested links to its source. Additionally, we provide the dataset (conversion) provenance information, also stored in a separate named graph, for ease of dataset versioning and annotation. We store the RDF links via a batch process (**openAireRDF2020**): we accumulate RDF links into batches of 50 K links before issuing database write operations, which results in a better overall run-time performance of the workflow.

The three workflow phases are logically structured to insulate single features from the rest of the implementation. This was done to enable easier workflow extensions in the future. For example, if there is a new access approach available – for new or existing datasets, RDF semantic modeling, or storage approach – these features could be incorporated in the corresponding parts of the workflow, minimizing potential changes in other parts of the workflow.

Table 4: ScholeXplorer harvest

Datasets	8.687.604
Literature	657.547
Unknown type	289.875
References	2.193.806
IsReferencedBy	1.810.862
IsSupplementedBy	60.134
IsRelatedTo	676.562
IsSupplementTo	76.149
Total links	4.817.513
Total RDF triples	163.854.345

Table 5: JDA-to-EconBiz harvest

Datasets	115
Literature	115
IsSupplementTo	115
Total links	115
Total RDF triples	4.549

To wrap up the section, few features of the two datasets are in order (see Table 4). The number of datasets and literature resources linked in the 4.8 M link collection, including the types of relationships, and the total RDF triples generated by the workflow are shown. Please note that there were cases in the collection where the type of the resources being linked was not specified ("unknown" in the table). Similarly, Table 5 includes information about the rather minimal link collection between data and publication repositories of a single institution. Their importance is mainly in enabling few of the use cases in this work, as well as in demonstrating the straightforward extension capabilities of the Semantic Web (the inclusion of attributes other than what the Scholix Information Package provides) as the technology adopted in this work. These additional attributes also enable additional use cases, which was the main rationale for (creating and then) their inclusion. Finally, we ran our application on a server with the following specifications: 4 CPUs (4 x 18 cores) and 1 TB of RAM, running on a CentOS, whereas the (Java) application was configured to run with 80 GB of

heap space. The disk space used to store the resulting RDF collection is the only hurdle from running the workflow on a personal machine.

8 Use cases: Explored scenarios

The use cases revolve around search and filtering and are based on the available metadata for links, publications, and RD. We group the potential scenarios in two broader groups: (a) scenarios that include only the link collection, and (b) scenarios that combine link and collections of other research artifacts.

(a) Searching (through) links The scenarios in this group provide the common ones that are typically conducted with any collection. These scenarios enable the search and filtering of links, including the resources linked (both, or only one – source or target), based on the available metadata used to represent them. Some of the examples include:

- List resources that are linked by the same publisher, publication date, domain, author(s), and so on. For the links that also include subject terms or file size (for RD) metadata for either source or target of a link, these metadata can also be considered during search.
- Based on available links, provide information on researchers who use those publications/RD, as well as the level of interest expressed. This could be of interest to both individual researchers and institutions tracking their research impact.
- Show resources (links or linked artifacts) based on criteria such as subject coverage, resource type, number of files a resource has, etc.
- Search or filter based on available metadata relevant to a researcher/research community. This is where the available metadata coupled with what the link model provides or is extended to provide (as in our case) comes to fruition.

Let's take the search based on the link resources subject. Searching for resources (source or target) that match the subject term `economics`, for example, we get 6 matches from the first collection: 4 datasets (link source) and 2 publications (link target) from the collection. Out of them, 2 datasets (JDA) are used as the data source for the corresponding 2 publications (Econbiz), i.e., they refer to 2 link instances that bring a dataset and publication from the same subject together. Searching the second collection with the same terms provides 141 matches. Although a far bigger collection, its providers are mainly from life sciences, with little coverage of the economics domains. For the former collection, we rely on the subject terms of resources (as represented in the link extension we adopted), whereas for the latter we rely on the title of link resources to check if it contains the said term.

(b) Searching over heterogeneous collections The nature of the link lends itself complementary to other resources, thus it supports search scenarios that

involve heterogeneous resources. A link collection could provide a useful recommendation to external collections with the matching resources, in this case, publications or RD. Let's see two cases for this scenario: publications and datasets collections.

Searching for RD In this scenario, a user searches for RD that (as a primary source of data) directly supports a research article a user is interested in. She has access to a collection of publications and another one of the links. If a publication from the former matches a publication that is part of a link (source or target) in the latter, she can consider the other resource (the dataset in this case) it links to as a relevant item. The way the matching occurs is not important at this point and can include different metadata, such as title, publication date, author, and so on, as seen before. As an example, using a publication collection from the ZBW – EconStor (**econBiz2020**), with over 3.6 M RDF triples – gave back 34 results with the subject term metadata `economics`, which a user can additionally consider. For any of these publications, we can search a link collection for matches from the links, which would in turn provide us additional information in terms of RD.

On another note, the user can rely on the subject terms (or terms present in the tile, abstracts, or descriptions of resources) to search for a field of interest across resource providers for a more interdisciplinary search scenario. Given resource collections from the domain, this would allow one to search for fishing quotas for a given fish type, the market fluctuations in a certain period, as well as the impact of climate conditions on its habitat.

Searching for publications (and even more data) In this scenario, the user has access to an RD repository and wants to find relevant/related publications and/or RD from the link collection. Similarly to the previous case, if there is a match of a dataset item to a resource in the link collection, the publication or RD resource the link points to can be recommended to the user. Moreover, users might want to see what are the disciplines that certain RD are currently "trending" in. The link model we use provides a good measure for this, as it shows citations between publications and datasets that can be quantified in the context of popularity or trending of a resource, which could be a useful feature during search and filtering.

When it comes to the search scenario based on the research discipline, the subject terms play an important role. As with the previous search category, they enable one to search across collections from different institutions that provide resources – be it links, publications, or RD. As mentioned, the available metadata in our collection provide many filtering capabilities for this scenario such as restricting links based on a certain time frame, the type of resources they link to, the institution that publishes them, and so on.

We used the Apache Jena framework (and its Fuseki server) to realize the mentioned use cases. To try out the explored use case scenarios, interested parties can follow the instructions in the workflow code implementation[1], as well as the sample (SPARQL) queries that are part of the source code implementation.

[1] https://bitbucket.org/fidanLimani/semanticlinkrepresentation

9 Conclusion

The evolving technology and research practices, such as Open Science, are pushing for dissemination of and access to as many research artifacts as possible. In this work, we explored one such artifact – links between scholarly resources – as value drivers to information infrastructures, in particular DLs. We harvested more than 4.8 M links from 2 sources, and adopted suitable models – via ontologies and vocabularies – to structure them in a common (semantical) machine-readable representation. Finally, we explored different use case scenarios of interest to a DL environment. The key implications of this work include recognizing more package-like provision of research outcomes, materialized via links in our case, and identifying potential use cases to be considered in a DL environment.

With the initial results, 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 issue that we identified during our work was the metadata inconsistencies for the different resources (identification schemes, metadata variety, etc.), which we needed to handle as part of our workflow implementation.

In our future work, in addition to publications and datasets, we plan to "package" more resource types 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 collection. These analyses are especially important as the link collection grows and includes new resource types. Finally, as far as RDF modeling goes, at times, it felt like there is a lack of a fitting ontology to represent the Scholix Framework link model adopted in the study, and we see this gap as a beneficial follow-up work soon.

References

Apache Jena, https://jena.apache.org

Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). 'Are scientific data repositories coping with research data publishing?'. Data Science Journal Vol.15. pp. 6. https://doi.org/10.5334/dsj-2016-006.

Bibliographic Framework Initiative, https://www.loc.gov/bibframe/.

Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., & Schindler, U. (2017). 'The data-literature interlinking service: Towards a common infrastructure for sharing data-article links'. Program Vol. 51, No. 1, pp. 75 – 100.

La Bruzzo, S., & Manghi, P., (2019). 'OpenAIRE ScholeXplorer Service: Scholix JSON Dump [Data set]'. https://doi.org/10.5281/zenodo.2674330.

Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., ... & Schindler, U. (2017). 'The Scholix framework for interoperability in data-literature information exchange'. D-Lib Magazine Vol. 22, No. 5/6, pp. 11

CiTO, the Citation Typing Ontology, https://sparontologies.github.io/cito/current/cito.html

de la Fuente, G. B. (2016). 'Libraries: roles and opportunities on Open Science', https://www.fosteropenscience.eu/content/libraries-roles-and-opportunities-open-science.

de la Fuente, G.B. (n.d.) 'What is Open Science? Introduction', https://www.fosteropenscience.eu/content/what-open-science-introduction.

Dublin Core Metadata Initiative, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

EconBiz - Find Economic Literature, https://www.econbiz.de/

Expression of Core FRBR Concepts in RDF, https://vocab.org/frbr/core#

Fathalla, S., Vahdati, S., Auer, S., & Lange, C. (2017). 'Towards a knowledge graph representing research findings by semantifying survey articles'. International Links between research artifacts Conference on Theory and Practice of Digital Libraries pp. 315–327. https://doi.org/10.1007/978-3-319-67008-9_25.

FigShare, https://figshare.com/

Hoekstra, R., Groth, P., & Charlaganov, M. (2014). 'Linkitup: Semantic Publishing of Research Data'. Semantic Web Evaluation Challenge pp. 95 – 100. https://doi.org/10.1007/978-3-319-12024-9_12.

Isaac, A. (2013). 'Europeana data model primer'.

Journal Data Archive, https://journaldata.zbw.eu/.

Kauppinen, T., Baglatzi, A., & Keßler, C. (2016). 'Data-Intensive Science'. https://doi.org/10.1201/b14935

Khan, B., Robbins, C., & Okrent, A. (2020). 'Science and Engineering Indicators 2020: The State of U.S. Science and Engineering'. National Science Board, Science & Engineering Indicators, https://ncses.nsf.gov/pubs/nsb20201/.

Kramer, S., Leahey, A., Southall, H., Vompras, J., & Wackerow, J. (2012). 'Using RDF to describe and link social science data to related resources on the web'.

Leibniz-Informationszentrum Wirtschaft, https://www.zbw.eu.

Limani F., Latif A., Tochtermann K. (2019). 'Scholarly Resources Structuring: Use Cases for Digital Libraries'. Garoufallou E., Fallucchi F., William De Luca E. (eds) Metadata and Semantic Research. MTSR 2019. Communications in Computer and Information Science Vol. 1057. https://doi.org/10.1007/978-3-030-36599-8_22.

Limani, F., Latif, A., & Tochtermann, K. (2017). 'Bringing Scientific Blogs to Digital Libraries'. WEBIST, pp. 284–290. https://doi.org/10.5220/0006295702840290.

Linked Data Principles, https://www.w3.org/wiki/LinkedData.

Mayernik, M. S., Phillips, J., & Nienhouse, E. (2016). 'Linking publications and data: Challenges, trends, and opportunities'. D-Lib Magazine Vol. 22, No. 5/6, pp. 11.

McCallum, S. (2017). 'BIBFRAME development'. Italian Journal of Library, Archives and Information Science pp. 71 – 85. http://digital.casalini.it/10.4403/jlis.it-12415.

Ogungbeni, J. I., Obiamalu, A. R., Ssemambo, S., & Bazibu, C. M. (2018). 'The roles of academic libraries in propagating open science: A qualitative literature review'. Information Development, Vol. 34, No. 2, pp. 113-121.

Peroni, S., Shotton, D., Ashton, J., Barton, A., Gramsbergen, E., & Jacquemot, M. C. (2016). 'DataCite2RDF: mapping DataCite metadata schema 3.1 terms to RDF'. Figshare, DOI. Vol. 10, pp. m9.

Priem, J. (2013). 'Beyond the paper', Nature, Vol. 495, No. 7442, pp. 437–440.

Programmable Web, https://www.programmableweb.com/category/science/apis?category=20070.

Research Objects, Enabling reproducible, transparent research, http://www.researchobject.org/.

RMap, https://rmap-hub.org/.

ScholeXplorer - The data literature interlinking service, https://scholexplorer.openaire.eu/.

Scholix: A Framework for Scholarly Link eXchange, http://www.scholix.org/.

Stocker, M. (2017). 'From data to machine readable information aggregated in research objects'. D-Lib Magazine, Vol. 23, No. 1, pp.1.

The FAIR Data Principles, https://www.force11.org/group/fairgroup/fairprinciples.

Wiljes, C., Jahn, N., Lier, F., Paul-Stueve, T., Vompras, J., Pietsch, C., Cimiano, P. (2013). 'Towards linked research data: An institutional approach'. CEUR Workshop Proceedings Vol. 994, pp. 27 – 38.