ICSU-WDS & RDA
Publishing Data Services WG

Summary & Recommendations

Prepared by Adrian Burton & Hylke Koers, March 2016

Outline

Executive Summary

The Publishing Data Services (PDS) Working Group, under a dual mandate from the Research Data Alliance (RDA) and ICSU-World Data System (ICSU-WDS), addressed the absence of a unifying framework to collect, manage and share information about links between research data and the literature. The WG holds that such a unifying framework would be of tremendous value for the stakeholders in the research data landscape and, ultimately, power new services to the benefit of researchers, as well as spur a transformation in research practices.

The main deliverables of the WG are twofold.

First, it is delivering a recommendation for a long-term approach to sharing information about the links between the literature and research data. Coined “SCHOLIX”, the proposed approach centers around the notion of hubs which aggregate links from their natural communities and utilise a common interoperability framework to exchange and expose those links.

Second, in synergistic efforts with OpenAIRE and PANGAEA, the WG has developed a prototype implementation of a data-literature interlinking system (the DLI system). This prototype serves as a reference system and is expected to develop into a key element of the proposed long-term approach.

The WG has brought together a substantial sample of over 1.4 million data-literature links, by involving a number of influential publishers, data centres, and global identifier service providers.

To realize the proposed model, further work on technology and standards, as well as advocacy to drive usage and adoption, is needed. It is recommended that this be coordinated through a follow-up working group under the umbrella of the Publishing Data IG.

Problem Analysis

Introduction

As research data is emerging as a first-class citizen of scholarly communication, it becomes essential to capture its relationship with other research products and entities. Without these relations, research data lives in isolation which makes it more difficult to find, access, and make sense of it. In particular, by creating bi-directional links between data and the literature, the visibility of both data and publications is increased, and data can be interpreted in the right context much more easily. Researchers strongly agree that it is useful to link underlying research data with the formal literature, as testified by the PARSE.Insight study concluded in 2010 (**1**; see also **2**).

While there are several organizations that track, store and expose links between research data and the literature, this is usually done in isolation and through ad-hoc bilateral arrangements (e.g. between an individual data center and a publisher, see **(3)** for a recent overview). There is currently no overarching standard or service to exchange links, or to combine links from different sources into a common resource. As a consequence, although different parties have a "piece of the puzzle" at different stages of the scholarly life cycle, those pieces cannot be readily combined into a rich and comprehensive network of published literature and data sets - thereby strongly limiting the value that can be realized from these connections.

The core objective of the "Publishing Data Services" WG, operating under a dual mandate from ICSU-WDS and RDA, has been to address the lack of a common standard or service to unify the data-literature landscape. From the RDA Case Statement (**4**): *"As a primary point of focus, the WG will address the problem of limited interoperability between data repositories, scholarly journal publication platforms, and tools for bibliometric analysis. Currently, there is no common framework for cross-referencing data sets and published articles, which creates barriers and inefficiencies for the interlinking and contextualization of journal articles and data sets. This is a problem because better connections between articles and data will improve the visibility, discoverability, and usability of scientific content and serve to accelerate science in the 21st century."*

Value Proposition and Use Cases:

The WG has worked towards an interlinking (or "cross-referencing") service for data sets and articles published in scientific journals. Borrowing language from the case statement (**4**) and a recent publication on this topic (**5**), the high-level value proposition of such a service for the different customer/user groups can be summarized as follows:

1. For **data repositories** and **journal publishers**: linking data and the literature will increase their visibility and usage, and can support additional services to improve the user experience on online platforms (for example, offering links to relevant data sets with articles, or offering links to the literature that will help place data in context). In contrast to the bilateral arrangements that we often see today between data centers and journal publishers, the proposed service will make the process of linking data sets and research literature a more robust, comprehensive, and scalable enterprise.
2. For **research institutes, bibliographic service providers, and funding bodies**: the service will enable advanced bibliographic services and productivity assessment tools that track datasets and journal publications within a common and comprehensive framework.
3. For **researchers**: firstly, the service will make the processes of finding and accessing relevant articles and data sets easier and more effective. Secondly it will make it possible for researchers to track long-term impact of their data (and publications), thereby providing additional incentives to share data.

Refining these high-level value propositions, the WG has collected a set of concrete use cases which specify a particular user need around data/literature links. An extensive list of use cases was contributed by the NDS-OLDRADA project (See Appendix A). To prioritize this list, the WG organized three different sessions to hear from prospective users of the service what their most immediate needs are. The outcome of this exercise was a set of four, largely independent, priority use cases. These priority use cases are summarized in the table below.

Use Case	**Details**
Live linking	
As a *publisher,* ***I want to*** *know about relevant data for an article that I published* ***so that I*** *can present links to such data sets to the users on my platform* - OR - ***As a*** *data center,* ***I want to*** *know about relevant articles for a data set that I published* ***so that I*** *can present links to such articles to the users on my platform*	● Needs to be **on-demand, real-time query**. **Performance** is critical. ● Publisher or data center platform should be able to **control UI** for smooth platform integration. ● No need for the service to do any filtering; just return all linked data sets and client can filter as needed.
Overview	
As a *data center,* ***I want to*** *obtain a full overview of article/data (and data/data) links for the data sets relevant to me* ***so that I*** *can demonstrate the utility of my data*	● Query should be **on-demand**, **complete**, and **up-to-date**. ● **Precision** and **comprehensiveness** are key ● Ideally **on-demand**, pull mechanism.

Notification	
As a *data center****, I want to*** *be alerted that an article may be citing/referencing our data* ***so that I*** *can validate that link and then add it to our own database.*	● For an alerting mechanism, **recall is more important than precision** (since the data center will still validate) ● Should be **push** notifications. ● Data center needs to be able to selectively receive notifications for their data repository only, **need "data center" metadata**. ● This service is not so sensitive to comprehensive coverage
Exploration	
As *a researcher interested in a particular topic of study,* ***I want to*** *be able to explore a relevant article/data graph* ***so that I*** *can find the articles or data sets that I am interested in.*	● General "research" use case**,** could apply to individual researchers, data repositories, and others. ● Requires a lot of **freedom to do exploration** at the user's terms ● Would expect the user in this case is highly tech-savvy and will want to create their own search logic using a minimal **"hopping service"** that exposes a set of links given an article or data set PID.

Overview of current practices & solutions

Current practices and approaches towards literature-data linking are very heterogeneous, both in terms of how links are created as well as how they are managed and shared. This heterogeneity can be seen both at the social level (different practices through which researchers create links between data and the literature), as well as at the technical level (how are links collected, stored, and shared).

At the social level, the practices that researchers follow to create a connection between a research data set and a literature publication remain diverse, with different attitudes towards some important aspects:

1. **Choice of persistent identifier.** While data DOI's are gaining traction, other (non-DOI) persistent identifiers remain prevalent in some disciplines, for example Genbank and Protein Data Bank accession numbers in the life sciences.
2. **Type of data referencing in the literature,** ranging from formal data citation (as recommended in the Joint Declaration of Data Citation Principles[1]) to informal references to data sets included within the main text of an article.

[1] See https://www.force11.org/group/joint-declaration-data-citation-principles-final

3. **The moment at which the link is asserted**, which could be at publication of the article or data set (“*T=1*”) but also at a later moment (“*T=2*”), for example by adding a reference to the literature in a data sets metadata.

In addition to this, there is another broad set of questions around the actor and process of depositing links - which can range from an individual researcher asserting that “my data set X underlies my publication Y”, to scientific data officers that make such claims on behalf of a group of researchers, to text-mining projects that infer connections between objects with certain degrees of confidence.

At the technical level, a number of solutions with limited scope (often on a bilateral basis) have been proposed or developed to interlink data and the literature (see **3** for an overview). Building on top of those, several organizations and groups have developed solutions, or expressed an interested in doing so, to bring links from different sources together. Without attempting to be comprehensive, these include the National Data Service, ICSU-World Data System, Open Science Foundation, BioCADDIE, and THOR. Domain-specific approaches include Europe PubMed Central (for life sciences), the Neuroscience Information Framework, and the NASA Astrophysics Data System. Here we’d like to highlight three efforts that are particularly noteworthy in the present context because of their potential to organize links at large and across different domains:

1. CrossRef & DataCite have set up a common DOI event tracker service, which enables the exchange of DOI-DOI links between the CrossRef (mostly literature publications) and DataCite (mostly data sets) infrastructures.[2]
2. OpenAIRE developed a robust infrastructure to perform large-scale analysis of scientific documents and establish relations to funder information through inference. One of the objectives of the OpenAIREplus project is to extend this to relations between data and the literature.[3]
3. The RMap project[4], funded by the Alfred P. Sloan Foundation, developed data models and standards to map and preserve relationships between scholarly resources, including literature publication and their underlying data.

Gap analysis

While there is clear merit in each of the efforts mentioned above, their scope is limited in one way or the other - be it focusing on a certain domain only (for example, life sciences or astronomy), or concentrating on only a subset of all possible links (for example, DOI-DOI links). What is still missing from the picture is an overarching, universal approach that addresses the challenge in a way that is fully *inclusive* with regard to current practices and that is valid across domains.

[2] See http://crosstech.crossref.org/2015/03/crossrefs-doi-event-tracker-pilot.html

[3] See http://www.openaire.eu/en/component/content/article/76-highlights/326-openaireplus-press-release

[4] See http://rmap-project.info/rmap/

What is lacking, thus, is a universal “system”[5] that enables interested parties to contribute links to a common pool where others can retrieve them. Or, as Callaghan et al. wrote as one of their three key recommendations in **(3)**:

*“**Role of a centralised, third-party registry:** There is a role for a centralised, third-party registry and metadata broker in data publication to simplify the process of passing information between data repositories and journals. As yet this registry does not exist, though some existing initiatives (DataCite, OpenAIRE) provide some aspects of the service that would be required of this registry. Although not data-related, CrossRef also provide some aspects of this registry service. We recommend that this be investigated through the Publishing Data Interest Group of the Research Data Alliance.”*

This, in essence, is the charter which the Publishing Data Services WG has taken up.

Recommendations

General considerations

In the process of conceiving a data/literature interlinking system that delivers on the value proposition and use cases as described in the above, the working group has formulated a set of recommendations. These have crystallized during the course of the WG, and have been discussed in detail and refined during a Workshop held in Amsterdam in January 2016 to discuss a vision and approach for a long-term, sustainable linking system (see also **6**). In summary, the WG recommends that a system to interlink research data and the literature should have the following qualities:

1. It should be **universal,** in the sense of being unrestricted in terms of scientific discipline or geography, and connecting with a range of existing practices so as to strive for **comprehensiveness.**
2. It should be **inclusive** and **participatory**, supported and used by all stakeholder groups.
3. It should be **open**, with terms of access and participation that are **non-discriminatory**
4. It should be **trustworthy,** providing users a level of **quality** appropriate to their purpose. Given the various use cases, this is best done through **meticulous provenance and metadata** (as opposed to a “filtering at the gate” approach)
5. It should be based on common **standards**, and **extensible** to other sorts of links (for example research ID’s, funding bodies, but also to e.g. tweets or Wikipedia entries).

[5] Where, at this point, we use the word “system” in a very general context - meaning anything from a set of common standards to a full service.

In addition to the above, any long-term system needs to be set up and operated in such a way that it is **sustainable.** While more detailed recommendations in this regard will be discussed in the next section, it should be mentioned here that the WG recommends that a sustainable, robust, future-proof system is best realized through the combination of:

1. An **enabling infrastructure** and standards-based **interoperability framework**;
2. On top of which various **services** can be built that address one or several use cases.

Enabling Infrastructure and Interoperability Framework

The WG organized a focused workshop in January 2016 with a limited number of people representing some of the key infrastructure providers in scholarly communication. The purpose of the meeting was to come to a common, shared vision on a sustainable system to interlink research data and the literature - and agree on next steps, joining efforts to get there.

A full workshop report, including more detailed recommendations, is available as a separate document **(6)**. We copy the summary here:

The WG proposes an approach to sharing information about the links between the literature and research data. A set of hubs will collect literature-data (as well as data-data) links from their natural communities using minor extensions to existing local procedures and, in some cases, inference. The hubs agree on an interoperability framework with a common information model and open exchange methods, optimised for exchanging information among the hubs. The hubs will serve as an enabling global information infrastructure for the development of (third party) services.

This proposed model is depicted in Figure 1.

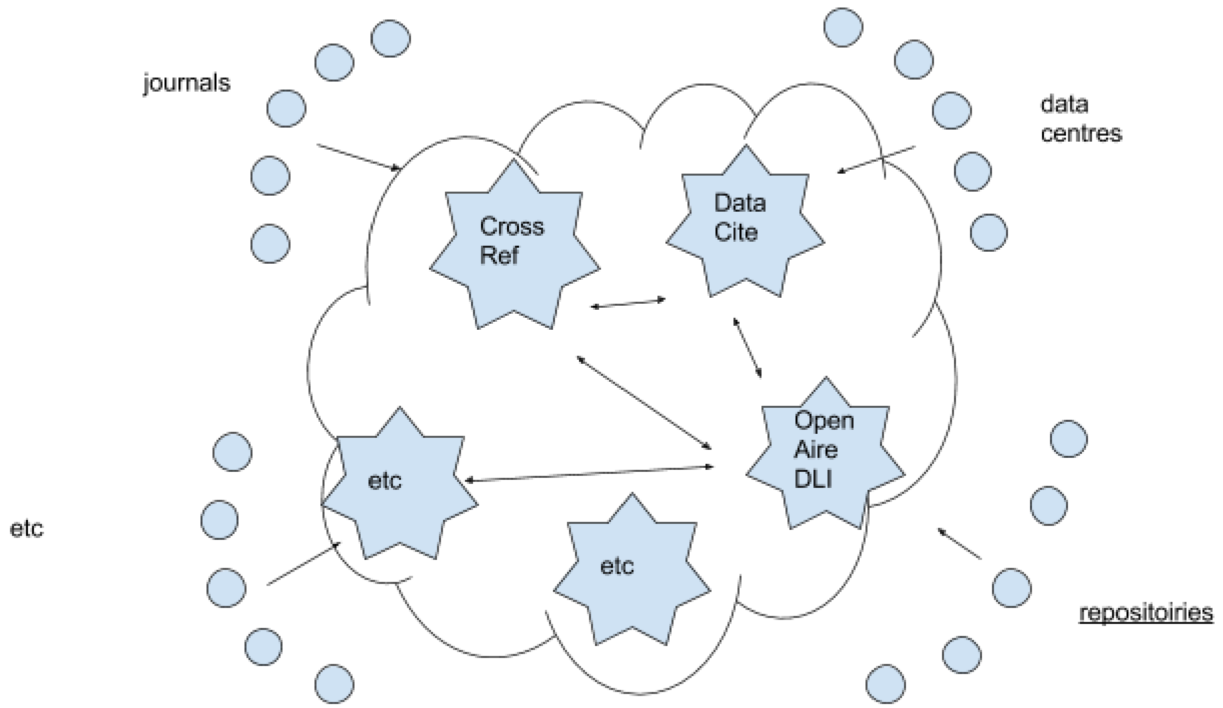


Figure 1: A visualization of the approach to sharing information about the links between the literature and research data as proposed here: a model with a limited number of hubs that connect with their natural communities and share links between themselves using a common interoperability framework. This model will be referred to as “SCHOLIX”.

It was decided to name this approach: **SCHOLIX** - an approach to **Scho**larly **Li**nk E**x**change.

Deferring a detailed discussion and motivation of this recommendation to (**6**), some key benefits for the “multi-hub” approach are:

1. **Minimizing the effort for link contributors:** Link aggregation can be set up as extensions of existing infrastructure and content streams (for example, journal publishers would be able to share links through CrossRef by extending workflows that already exist between them). Compliance to interoperability standards is only required by the relatively small number of hubs.
2. **Efficiency through specialization.** A hub can specialize in collecting links from a certain domain, or restrict itself to any particular area of the full space - for example a particular research domain, grey vs. formal literature, links contributed at “*T=1*” vs. “*T=2”*, etc.
3. **Avoid the risk of a single point of failure**: Both at a technical level, but also at an organization level, thanks to distributed ownership over several parties with a vested interest.
4. **Extensible and flexible framework** that facilitates the connection of new hubs and provides a “network effect” benefit for every new node.

5. **Open and enabling information environment** that supports the creation of value-adding services by third parties.

There are two important points to note here. First, the “multi-hub” infrastructure is only one element to a successful data/literature interlinking system: it will need to be supplemented by services to deliver value to end-users (for example to meet the use cases mentioned in the above) and by culture change. Secondly, while we feel that the “multi-hub” infrastructure provides the right environment to deliver a data-literature system following the general recommendations outlined above, it should be noted that it is still a *conceptual* model . The key to make it work lies in the proposed interoperability framework, including foundational terminology, information model and standards, and encoding and exchange options.

The working group has made substantial first steps in recommendations in all these areas (enough to support negotiated information exchange between CrossRef, DataCite, and OpenAIRE), but a need for further specification remains. For further details on these interoperability recommendations, please see (**6**).

The Data-Literature Interlinking (DLI) Service

Summary

The DLI (Data-Literature Interlinking) Service was developed by OpenAIRE in a synergistic effort with the Publishing Data Services WG and PANGAEA. It can be seen as a prototype of the interlinking system as envisioned in these recommendations, paving the way for the proposed interoperability framework at a conceptual, technical, and organizational level.

The DLI was developed to provide an end-to-end solution to collect links from contributing organizations, store these in a central database using a common schema, and expose them to interested parties. The system has demonstrated great value in several ways. First, it is an operational system that - even with the limitations inherent to its current status as a prototype - is fully able to deliver value to end-users. It also helps to bring the concept across to interested parties in a tangible, “hands-on” way, demonstrating clearly the value they get out of an interlinking system. Finally, developing the DLI has been a fruitful learning exercise in many technical aspects, for example data modeling and metadata exchange format. Such learnings have fed into the recommendations presented here, and are expected to be of significant value as the proposed long-term linking framework will develop further.

The adaptive nature of the DLI system has allowed the working group to address the chicken and egg problem: how to make progress, demonstrate value and “learn by doing” *before* the emergence of standard behaviours or technical frameworks?

Technical details of the DLI system, including data model and content aggregation system, are described in the conference paper (**5**)[6] and interested readers are referred to that publication for details. In the context of the recommendations presented here, three key assets of the DLI system that should be mentioned are:

- **Link corpus:** The DLI system currently holds over 1.4 million unique article/data links (in addition to article/article and data/data links).These links are contributed by a number of partner organizations, many of which are through the WG. Contribution organizations include data centers, publishers, and infrastructure providers – forming an excellent representation of the various stakeholders. It should be noted that, in the absence of commons standards, the DLI system allowed for a lot of flexibility in how links are deposited. There are live interfaces with some systems, but there is also link data that has been contributed as a one-off process and which, consequently, will not automatically stay updated.
- **Standards**: In developing the DLI, a pragmatic "test and learn" approach was followed in established data schema, encoding standards and database technology. The current approach is described in details in **(5),** but it is important to note here that a lot of attention was given to metadata and provenance information. For every piece of metadata, be it about the link or about a research object (article or data set), the origin of that metadata is meticulously tracked and stored.
- **Services:** The DLI system supports a number of ways to share (collections of) links with interested parties. There is a web portal for human inspection[7], OAI-PMH provision to download the full data corpus, and API's[8] to retrieve data programmatically to interface with other platforms.

The DLI and the proposed interoperability framework

The DLI system, as described above, can be seen as a trailblazer system that paves the way for the full interoperability framework as proposed in this document. In the language of the proposed "multi-hub" model, the DLI system can be described as a single hub with an integrated service layer.

Going forward, it is anticipated that the key assets of the DLI system, as listed in the above, will remain of value as we evolve towards the "multi-hub" infrastructure. Since the DLI already offers an end-to-end solution for aggregating, organizing and exposing links, it provides a natural point of reference for other parties to build and connect the elements that are necessary to realize the long-term approach proposed here. As that long-term linking framework develops, it is anticipated that the role of the DLI will change gradually to the extent that:

- The link aggregation function of the DLI will become one of several hubs

[6] An updated description, including technical views on extended interoperability within the proposed framework, is expected in a forthcoming publication.

[7] See http://dliservice.research-infrastructures.eu/

[8] The API has been developed by PANGAEA and, technically speaking, runs on an independent infrastructure.

- The multi-hub model will provide an open, sustainable and comprehensive pool of information to be aggregated.
- Data standards developed in the process of realizing the DLI will feed into common standards that are to be established to govern the communication between the hubs.
- The DLI system will connect with other hubs using these standards to exchange links.
- The DLI inference engine will continue to add to the explicit assertions.
- The DLI service layer which exposes links (using a portal, OAI-PMH, and API's) will provide a "one for all" interface that offers access to all links across the hubs.
- The DLI service will provide a public face to the SCHOLIX standards framework and be a model for other services.

Adoption and Implementation

Even within its 18-month RDA lifetime, the WG has seen its work already being adopted:

- Fifteen organisations (including representatives of publishers, data centres, registries, and global service providers) have contributed over 1.4 million data-literature links into the system.
- The proposed long-term approach to sharing information about the links between the literature and research data (the "SCHOLIX" model) is adopted by CrossRef, DataCite, and OpenAIRE.
- Europe PubMedCentral has adopted the DLI metadata standards to describe article/data links in their system, such that the DLI can harvest these links and include them in the service. This paves the way for fuller interoperability between that system and the DLI.
- The RD-switchboard, which was delivered by the RDA WG "Data Description Registry Interoperability (DDRI)", is ingesting links from the DLI to combine with other content sources to obtain a more comprehensive network of connected scholarly entities.
- The WG received an ad-hoc request to provide a listing of related data for a set of articles. This was delivered using the DLI's OAI-PMH interface.

In addition to this, a number of data repositories are currently exploring how to connect with the DLI system to offer a "related publications" kind of functionality to users of their web portals.

Any party that is interested to explore the current corpus of links in the DLI system can freely connect with the system using any of the existing services (web portal, OAI-PMH delivery, or API connections).

Next Steps

In summary, the Publishing Data Services Working Group is delivering (i) a recommendation for a long-term approach to sharing information about the links between the literature and research data, and (ii) a prototype implementation of a data-literature interlinking system that is expected to develop into a key element of the proposed long-term model.

In order to drive the agenda forward and realize the proposed long-term interlinking framework, the following steps are proposed:

1. **Establish common standards for data modeling and exchange**. This will leverage the work carried out by this WG to realize the DLI system, as well other efforts undertaken by groups and prospective hubs who have worked on data modelling aspects of data-literature linking.
2. **Implementation of common standards by prospective hubs** (CrossRef, DataCite, OpenAIRE, and others).
3. **Support adoption from stakeholder groups** (publishers, data centers, institutional repositories, etc.). It's recommended to this in collaboration with advocacy fora like Force11 or CODATA, as well as through publisher (e.g. STM) and data center peak bodies (e.g. ICSU-WDS).
4. **Capture and maintain documentation** and best-practices in a neutral, public place.
5. **Investigate and propose an organizational structure** to sustain the system, and investigate resourcing options.

The WG recommends that these activities be coordinated through a new joint RDA / ICSU-WDS Working Group that sits under the umbrella of the Publishing Data Interest Group. Its charter should focus on realizing the SCHOLIX model by coordinating technical work to realize the proposed interoperability framework and by fostering adoption and usage of the emerging system.

References

(1) PARSE.Insight, Permanent Access to the records of Science in Europe.

(2) *"Abelard and Héloise: Why Data and Publications Belong Together"*, Smit, E. (2011). DOI: doi: 10.1045/january2011-smit

(3) *"Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples. International Journal of Digital Curation",* Callaghan, S., Tedds, J., Lawrence, R., Murphy, F., Roberts, T., Wilcox, W. (2014). DOI: 10.2218/ijdc.v9i1.310

(4) Publishing Data Services WG Case Statement

(5) *"On Bridging Data Centers and Publishers: the Data-Literature Interlinking Service"*, in: Metadata and Semantics Research, Proceedings of 9th Research Conference, MTSR 2015, Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., and Schindler, U., (2015), DOI: 10.1007/978-3-319-24129-6_28

(6) Interoperability Framework Recommendations - output of the RDA/ICSU-WDS Publishing Data Services Working Group

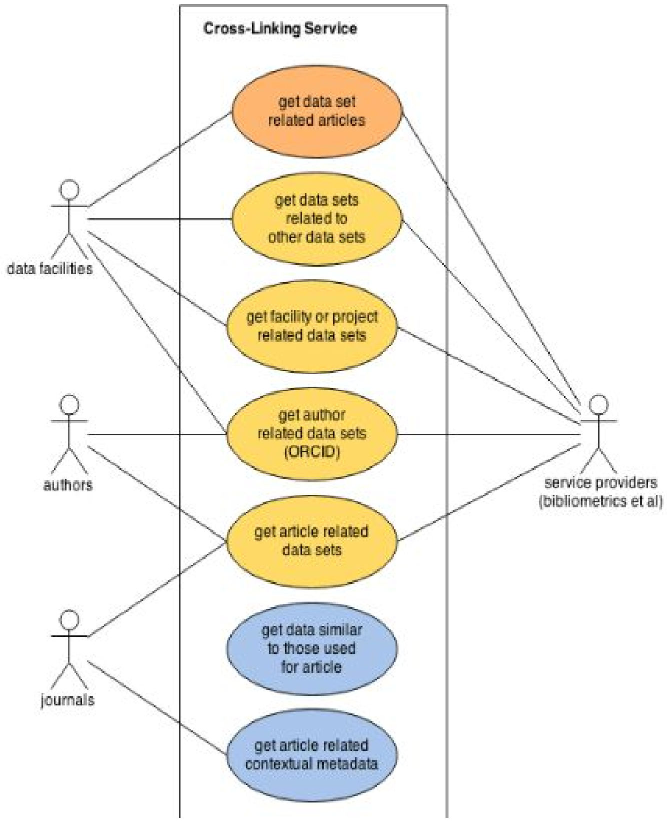
Appendix A: Overview of use cases

The following list was contributed by the National Data Service OLDRADA project[9]:

- I want to query the system with my deposit ID as input.
- For one of my dataset ID's, I want ID's of associated articles.
- I want the data to be available in a way that makes easy to incorporate into my platform.
- I want to deposit a batch of article-ID / dataset-ID associations.
- For one of my dataset ID's, I want ID's of associated datasets.
- I want to upload associations myself using an API.
- I want the project outputs to be available long term and interoperable and adoptable by other publishers.
- For one of my article ID's, I want ID's of related datasets.
- I want to modify or remove an association previously deposited by me.
- I want to download a complete snapshot of the OLDRADA dataset.
- I want some basic information with each link returned (title, author, funder etc).
- I want to restrict any query to a specific document or dataset identifier.
- I want to know the type of association (article only refers to data, or data is underpinning article, or data was curated from article, or ...).
- I want my dataset to be identified by a URL or by base-URL plus ID.
- I want all my datasets to be fully discoverable (e.g., from Google).
- I want to be crawled to get my associations uploaded.
- Datasets need to have some limited provenance associated.
- For one of my (article or dataset) ID's I want to know how often it has been used as input parameter.
- I want to upload associations myself using FTP.
- I want to variably manage the incoming requests.
- I want to search using basic information like titles, authors and funders.
- I need to be able to distinguish different versions of a dataset.
- In any results list, I want the results to be completely de-duplicated.
- For one of my data ID's I want to know how often it has been returned in search results.
- Deposits from my deposit ID should make all previous deposits invalid.
- I want an embargo period on data-article associations.

And this high-level diagram was contributed by Michael Diepenbroek:

[9] With thanks to IJ.J. Aalbersberg for collecting and sharing this list.


Cross-Linking Service
get data set related articles
get data sets related to other data sets
get facility or project related data sets
get author related data sets (ORCID)
get article related data sets
get data similar to those used for article
get article related contextual metadata
data facilities
authors
journals
service providers (bibliometrics et al)