This is a pre-print of the following article: Adrian Burton, Hylke Koers, Paolo Manghi, Sandro La Bruzzo, Amir Aryani, Michael Diepenbroek, Uwe Schindler, (2017) "The data-literature interlinking service: Towards a common infrastructure for sharing data-article links", Program, Vol. 51 Issue: 1, pp.75-100, https://doi.org/10.1108/PROG-06-2016-0048 **Please, cite this document by citing to the official article.**

The Data-Literature Interlinking Service: Towards a Common Infrastructure for Sharing Data-Article Links

Adrian Burton[1], Hylke Koers[2], Paolo Manghi[3], Sandro La Bruzzo[3], Amir Aryani[1], Michael Diepenbroek[4], Uwe Schindler[4]

[1] Australian National Data Service,
Melbourne, Australia
{adrian.burton, amir.aryani}@ands.org.au
[2] Elsevier,
Amsterdam, The Netherlands
{h.koers}@elsevier.com
[3] Institute of Information Science and Technology - CNR,
Pisa, Italy
{paolo.manghi, sandro.labruzzo}@isti.cnr.it
[4] PANGAEA,
Bremen, Germany
{mdiepenbroek, uschindler}@pangaea.de

Abstract. Research data publishing is today widely regarded as crucial for reproducibility, proper assessment of scientific results, and as a way for researchers to get proper credit for sharing their data. However, several challenges need to be solved to fully realize its potential, one of them being the development of a global standard for links between research data and literature. Current linking solutions are mostly based on bilateral, ad-hoc agreements between publishers and data centers. These operate in silos so that content cannot be readily combined to deliver a network graph connecting research data and literature in a comprehensive and reliable way. The RDA Publishing Data Services Working Group (PDS-WG) aims to address this issue of fragmentation by bringing together different stakeholders to agree on a common infrastructure for sharing links between datasets and literature. This paper presents the synergic effort of the PDS-WG and the OpenAIRE infrastructure to achieve these objectives by realizing and operating the Data-Literature Interlinking Service. The Service populates and provides access to a graph of dataset-literature links (at the time of writing close to five millions, and growing) collected from a variety of major data centers, publishers, and research organizations. To achieve its objectives the Service proposes an interoperable exchange data model and format, based on which it collects and publishes links, thereby offering the opportunity to validate such common approach on real-case scenarios, with real providers and consumers. Feedback of these actors will drive continuous refinement of the both data model and exchange format, supporting the further development of the Service to become an essential

part of a universal, open, cross-platform, cross-discipline solution for collecting and sharing dataset-literature links.

1 Introduction

Driven by innovations in digital technology and off-the-shelf availability of cheap storage solutions, research data is becoming ever more prominent in the way that research is performed and in the way that research findings are communicated. Research data holds a big promise, and improving the storing, sharing, and usage of data is seen by many as a powerful way to accelerate the pace of science, even fuel economic growth. As Neelie Kroes, then Vice-President of the European Commission responsible for the Digital Agenda put it: "Knowledge is the engine of our economy. And data is its fuel."

Challenges to realize the full potential of research data exist at different levels - from cultural aspects, such as proper rewards and incentives, to policy and funding, and to technology. The challenges are interconnected and impact a diversity of stakeholders in research data publishing - including researchers, research organizations, funding bodies, data centers, and publishers. It is essential that these stakeholders work together to address common issues and effectively push the envelope. ICSU World Data Systems (ICSU-WDS) and the Research Data Alliance (RDA) provide useful forums for these kind of collaborations, such as the Publishing Data Interest Group (IG). This IG addresses a range of issues in data publishing from a holistic and cross-stakeholder perspective, acting as the umbrella of Working Groups (WGs) that deal with data bibliometrics, data publication workflows, cost recovery, and services. Among these WGs, the Publishing Data Services WG (PDS-WG) brings together different parties in the research data landscape (e.g., data centers and publishers) with the objective of creating "an open, freely accessible, web based service that enables its users to identify datasets that are associated with a given article, and vice versa" [1]. The vision is that of moving away from the large set of bilateral arrangements that characterizes the linking eco-system today, towards establishing a common infrastructure recommending interoperability formats and tools enabling seamless exchange of article-data links between scholarly communication parties. Such a transition would facilitate interoperability between platforms and systems operated by the different parties, reduce systemic inefficiencies in the ecosystem, and ultimately enable new tools and functionalities to the benefit of researchers.

This paper presents in detail the ideas and implementation activities carried out by PDS-WG to realize a Data-Literature Interlinking Service (referred to as "the Service" in the following), as an extended and updated version of previous work described in [15]. In this process, the WG has joined forces with the OpenAIRE project[1] and infrastructure [10] in order to design, develop and deploy an operative and sustainable prototype of the Service. The Service has been conceived in such a way that its common data model

[1] OpenAIRE, http://www.openaire.eu

and exchange format can be refined over time to become community-driven standards, balancing between the information that can be shared across data providers and the information that is needed by consumers of the Service.

As a logical continuation of the development of the DLI Service, PDS-WG has joined forces with the THOR project, CrossRef, DataCite, and others to formulate the "Scholix" (short for Scholarly Link Exchange) guidelines, which are meant to drive the creation of an interoperability framework to exchange links between research data and the literature at a global scale.

Outline: Section 2 will advocate the need for data and literature links and present the driving motivation and methodology inspiring the realization of the Service in order to achieve the goals of the PDS-WG; Section 3 will present the general architecture of the Service, in terms of its functional requirements and data model as evolved today; Section 4 will present the current realization of the Service, some real-case consumers, and the next steps in the direction of improving service scalability and functional offerings.

2 The need for sharing Data-Literature links

The most immediate benefit in establishing links between articles and data is to increase visibility and discoverability, thus bringing data (and articles) more to the forefront and stimulating re-use. In addition, by providing links to the scholarly literature, data can be put in the right context that is often necessary to reproduce findings or re-use data properly (see also [5]). Researchers across disciplines strongly support the notion that there is value is creating links between data and the literature, as testified by results from the PARSE.Insight study[2], which was carried out with the help of EU funding in 2008—2010 : 85% respond *"yes"* to the question *"Do you think it is useful to link underlying research data with formal literature"* [5]. However, what is also clear is that in order to be fruitful, such linking needs to be done properly, by means of infrastructural solutions, delivering agreed-on policies, formats, and tools [3]. For example, a recent study in the astronomical literature showed that more than 50% of links from articles to data using a hard-coded HTTP web address were broken after 15 years [6]; and similar results have been reported in [12].

Many parties, in fact, are taking efforts to link up articles and data in a robust and future-proof way: a number of data repositories keep track of articles that cite, or refer to, their data; several publishers have some form of data-linking program to connect the articles they publish with relevant data hosted externally (see e.g., [7]); providers of bibliographic information are increasingly looking at data alongside the traditional article output; and organizations such as CrossRef, DataCite and OpenAIRE are developing systems to track or infer relationships between data and the literature (see also [8] for some examples of how data and literature publications are currently interlinked).

[2] *PARSE.Insight project*, http://www.parse-insight.eu/

However, these initiatives typically live in isolation, and there is no common framework for inter-linking datasets and published articles. As a consequence, although different parties have a "piece of the puzzle", those pieces cannot be readily combined to exploit at best the potential of a rich and comprehensive network of published literature and data sets. The work of PDS-WG is seeking to tackle the comprehensiveness and interoperability challenges underlying this scenario by realizing an open and one-for-all Data-Literature Interlinking Service. The Service will serve as a flexible sandbox where major scholarly communication stakeholders interested in sharing or consuming dataset-literature links will be able to do so while reporting their requirements, preferences, recommendations, obstacles to the PDS-WG. Such an incremental approach (see **Figure 1**) will enable the refinement of exchange formats, data model, and aggregation workflows implemented by the Service and, in the long run, to agree on common practices for sharing dataset-publication links.

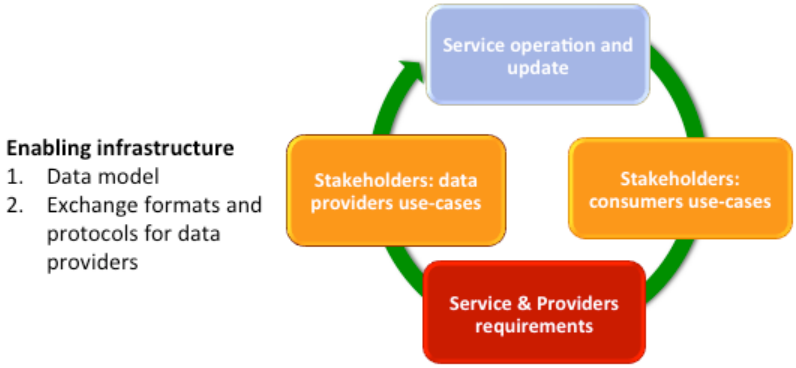

Figure 1 – Incremental cyclic methodology.

The operation of the Service will bring the following benefits (adapted from [1]):

1. *For data repositories and journal publishers*: it will make the process of linking data sets and research literature simpler and centralized, ensuring more visibility for data sources (and their "customers") as well as publication platforms.
2. *For research institutes, bibliographic service providers, and funding bodies*: it will enable the realization of advanced bibliographic services and productivity assessment tools that track datasets and journal publications within a common framework;
3. *For researchers*: it will make the processes of sharing, discovering, and accessing relevant articles and data easier, more efficient, and more accurate, thereby increasing scientific reward and enhancing its practices.

Furthermore, the operation of the Service provides an ideal testing ground to explore concepts and gather feedback from contributors, users, and stakeholders in general. These insights and feedback are instrumental in defining a vision and roadmap for a sustainable and robust global interlinking infrastructure.

As part of its final recommendations alongside the DLI Service, the PDS-WG put forward a long-term vision for linking research data and the literature under the name of "Scholix" - a framework for Scholarly Link Exchange [13]. In essence, Scholix is a proposed interoperability framework that relies on the notion of 'hubs' as natural aggregators for links between data and the literature. Such hubs would include CrossRef as a natural interface for journal publishers, DataCite for data centers, OpenAIRE for institutional repositories – as well as other, possibly more domain-specific, organizations that are well-placed to work with their communities to capture links between data and the literature.

Rather than a normative standard, Scholix represents an evolving lightweight set of guidelines including a conceptual model, an information model, information standards and encoding guidelines, and options for exchange protocols. Together, these guidelines will drive greater interoperability between the hubs, creating an open information ecosystem that will enable dedicated services to meet specific use cases for link consumers.

The development of the DLI has been instrumental in defining the Scholix vision, and going forward it is expected that the DLI will develop into an essential component of the infrastructure, with a dual role: one the one hand as an aggregation hub, and on the other as a user-facing service that will continue to provide access to the 'universe' (i.e., the information space across all hubs) of links.

2.1 Modus operandi

Four key principles underpin the thinking and the work carried out in the PDS-WG. First, the challenge of developing an open, universal interlinking system is as much of a "soft" (social) problem as it is a "hard" (technical) problem. The WG has therefore invested a considerable amount of time and effort in building a broad base of support through communication and outreach activities. Today all of the groups that were identified as key stakeholders - including data centers, publishers, providers of bibliographical information, funding bodies, etc. - are supporting the initiative, be it through WG membership, contributing a corpus of article/data links, participating in the technical work, or a combination thereof. The initiative is open and inclusive[3] and additional participation by other groups or individuals will be welcomed.

Second, the WG is prioritizing its efforts towards building, a working prototype of the Service that can be used to demonstrate value to the intended users and stakeholder groups. This work is carried out in synergy with the OpenAIRE infrastructure, PANGAEA data archive, and ANDS data archive network. As with any demonstrator system, coverage and functional scope are initially limited but the ambition is to develop a service that will be of direct value in real-world situations. The admittedly

[3] A set of "guiding principles" that includes statements on the open character of the project can be accessed through the WG's RDA website: https://www.rd-alliance.org/groups/rdawds-publishing-data-services-wg.html

important set of questions around longer-term sustainability and governance of the Service is deferred to a later stage of the WG's lifetime. Specifically, a pragmatic, ground-up approach was followed: aggregate as many corpora of literature-data links as possible, harmonizing them into a common data model, and making them available online through an openly accessible Service. That means that in the initial stage of operation the WG admits a considerable effort to ingest heterogeneous information from contributors. In the long run, the expectation is that the Service will help at establishing exchange standards that will reduce conversion costs and lead to a more scalable approach. To this aim the Service will enable a "test & learn" approach, by facilitating the extension of the common data model and schema over time.

Third, the WG takes a generic, one-size-fits-all (as opposed to e.g., domain-specific) approach as much as possible to avoid fragmentation and preserve the value that lies in developing a comprehensive solution for all articles and all datasets. This approach necessarily means that the Service common data model is relatively discipline-agnostic, leaving domain-specific metadata a responsibility of the data repositories. This view also fed into the Scholix infrastructure, which is essentially domain-agnostic in its standards and infrastructure, yet leaves room for domain-specific information to flow into the system.

Finally, the WG places significant emphasis on provenance, reliability, quality of data-literature links and the associated metadata, considered of great importance for most key use cases (e.g., linking from online publishing or data platforms, bibliometrical analyses). This principle is reflected in the Service operation, which ensures that: (i) links are contributed by trusted sources, rather than inferred by the system, and (ii) the origin and completeness of links and metadata is tracked at a high level of detail and granularity.

2.2 Related Work

The ambition of enabling the realization or de-facto realizing a Data-Literature Interlinking Service is not unique. A number of related initiatives and organization active in this space aim at defining models, protocols, and services, with a focus specific to various research disciplines, kinds of dataset involved, and consumers. That list includes (but is not limited to) CrossRef[4], DataCite[5], OpenAIRE[6], RMap[7] [16], the National Data Service[8], bioCADDIE[9], the Open Science Framework[10], THOR[11], SILK framework[12]

[4] CrossRef, http://www.crossref.org
[5] DataCite, http://www.datacite.org
[6] OpenAIRE, http://www.openaire.eu
[7] RMap, http://www.rmap-project.info
[8] National Data Service, http://www.nationaldataservice.org/
[9] BioCADDIE, https://biocaddie.org/
[10] https://osf.io/
[11] THOR EC project, http://cordis.europa.eu/project/rcn/194927_en.html
[12] Silk Framework, http://silkframework.org/

and LIME for RDF Linked Open Data, and the RDA Data Description Registry Interoperability (DDRI)[13] WG which has developed RD-Switchboard.org[14].

Of particular interest to the work presented here are the following efforts, the first three closely involved in discussions around the Scholix framework and guidelines. Their developments fit in naturally with the proposed long-term infrastructure, and will feed into further enhancements of Scholix high-level interoperability standards.

CrossRef and DataCite Event Data. Crossref and DataCite are collaboratively working on providing article/data links via the Event Data service. Event Data is partly shared infrastructure between the two organizations, and partly independent services by Crossref and DataCite, as Event Data is a generic service for links between DOIs and other resources, some of which fall outside the scope of Scholix. The Event Data service follows the Scholix specification for describing assertions, and makes the assertions available to other Scholix Hub partners. Publishers and data centers submit article/data links via the DOI metadata they deposit with Crossref and DataCite, respectively. Crossref and DataCite Event Data will become production services in 2017. While CrossRef and DataCite provide the mass of links available at publishers and data centres, the DLI Service acts as a binder and merger of this information.

OpenAIRE infrastructure. OpenAIRE developed a robust infrastructure to perform large-scale analysis of scientific documents and aggregate relations to funder information and datasets through harvesting from data sources and by text-mining a collection of around 5M OA publications. One of the objectives of the OpenAIRE2020 project is to act as a broker of links between publications and datasets [17] by adhering to the Scholix recommendations. OpenAIRE aggregates and provides the links available at repositories and many other links inferred from article full-texts to DOIs. As such it offers to the DLI service publication-dataset links that potentially differ from DataCite and CrossRef, providing the missing slice of the cake.

Research Data Switchboard. One of the third party tools that resembles the DLI Service and will in fact support the Scholix framework is the Research Data Switchboard, an open and collaborative software solution that addresses the problem of cross-platform discovery of research data. The system connects datasets together across multiple registries on the basis of co-authorship or other collaboration models such as joint funding and grants. The best metaphor for it is the "SEE ALSO" section in online bookstores, where customers are invited to look at other products by the same author, related topics or similar publishers. The outcome of an RD-Switchboard is a database conforming to the Research Graph schema[15] – an instance of such database has been created by the participants in the Research Data Alliance Data Description Registry

[13] *DDRI*, https://www.rd-alliance.org/group/data-description-registry-interoperability.html
[14] See http://www.rd-switchboard.org/
[15] *RD-Switchboard Research Graph,* http://researchgraph.org/schema/

Interoperability (DDRI) Working Group.[16] The resulting graph can then be queried and visualized via advanced graphical interfaces, whose implementation is based on the Force Directed Graph Drawing Algorithm [9], see Figure 2. At the time of writing this article, the RD-Switchboard source code has been adopted by the following institutions: NCI – National Computational Infrastructure, Australia, ANDS – Australian National Data Service, University of Sydney, Australia and National Institute of Informatics, Japan. The NCI adoption is the most advanced at this point. NCI uses the RD-Switchboard graph database to identify missing connections, improves metadata content, and discover new links between datasets, organizations and researchers.
By focusing on collection and inference of publication-dataset links relative to specific institutions (e.g., to their authors) or groups of institutions, RD-Switchboard has been realized as a goal-driven solution. Different instances of the system may therefore become precious provider data sources for the DLI service, by identifying links unavailable to publishers and data centres.

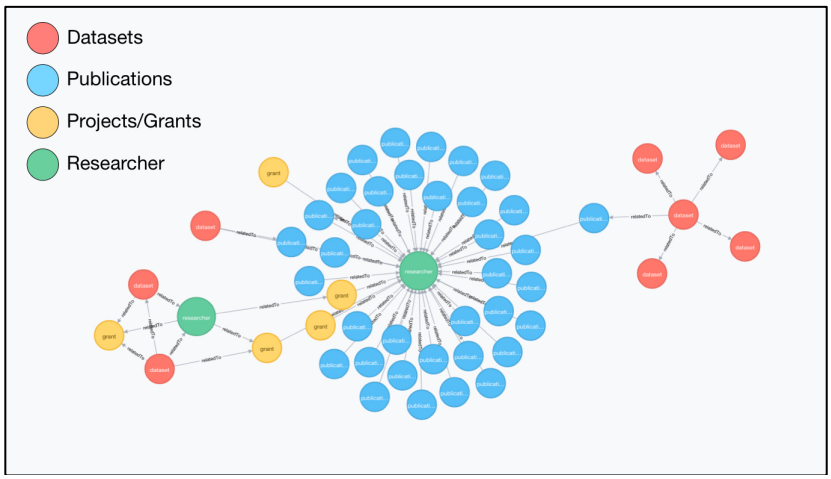

Figure 2 – Screenshot of the RD-Switchboard graph visualization using Neo4j Browser

Linked Open Data solutions. With specific focus on Linked Open Data, two known solutions to linking are the SILK framework (The Linked Data Integration Framework) and LIMES (LInk discovery framework for MEtric Spaces). Although these approaches tackle issues that may recall technical challenges underlying the DLI Service realization, their main focus is at a higher level of data abstraction, i.e., generic LOD collections (generic links), and at a higher level of integration, i.e., interlinking and identifying similarities between LOD data sources. The DLI service acts as an aggregator of publication-dataset and dataset-dataset links, collected from non-LOD data sources (publishers, repositories, and data centres), and its main goal is to act as a provider for the resulting graph, which today cannot be accessed from a single entry point.

[16] *Data Description Registry Interoperability (DDRI) Working Group*, http://rd-alliance.org/groups/data-description-registry-interoperability.html

3 The Data-Literature Interlinking Service architecture

The Data Literature Interlinking (DLI) service ("the Service") aims to populate and provide access to the *DLI information space*, a graph of relationships between dataset and literature objects, and between dataset and dataset objects. Objects and relationships are provided by data sources managed by publishers (e.g., Elsevier, Thomson Reuters), data centers (e.g., PANGAEA, CCDC), or other organizations providing services to store and manage links between datasets and publications (e.g., DataCite, OpenAIRE). The Service aggregates content harvested from the data sources and offers programmatic access (APIs) to the resulting information space. Such APIs offer full-text search by field or free keywords and bulk access to the collection (e.g., OAI-PMH protocol). They enable the construction of services on top of the Service – for example the DLI Service end-user search and statistics portal – and serve content to third-party community services – for example the RD-Switchboard Service developed by ANDS.

The Service is intended as a flexible playground where data curator users can monitor the aggregation outputs, collect feedback from data providers and service consumers, and refine ingestion workflows, common data model and exchange format accordingly. The expectation is that by means of such incremental and agile methodology, and involving pro-active consumers of the Service, this activity will converge to an ideal data model and exchange metadata format for description and exchange of links between datasets and publications. The following sections present the functional requirements of the Service and the initial DLI information space data model.

3.1 Functional requirements

This section discusses the general functional requirement identified by a study of the problem, based on several potential or candidate use-cases and on the experience of the OpenAIRE infrastructure, operating similar aggregation services for scholarly communication. In particular, the Service will support four categories of users, at different levels of abstraction:

- *Data source managers*, users operating data sources and therefore serving content to the Service. Examples of data sources are: scientific publishers, data centers, repositories, or aggregators of these. As such a data source may be a *publisher* of information (i.e., the keeper of the original digital data source) or the *provider* of information (e.g., an aggregator like DataCite). Their intent is to serve their user community at delivering scientific output to the world while gaining visibility at the same time;
- *Data curators of the Service*, users operating the Service, hence in the need of user-friendly tools to configure, orchestrate, and monitor data source aggregation activities in order to guarantee an expected QoS;
- *Third-party service developers*, users willing to (bulk) collect the DLI information space to process and offer it to their users or to extend their services by interactively searching the information space.

In order to serve such users, the Service needs to provide the following functional areas, as depicted in Figure 3: *aggregation of content, population of graph,* and *supporting access to the graph*.

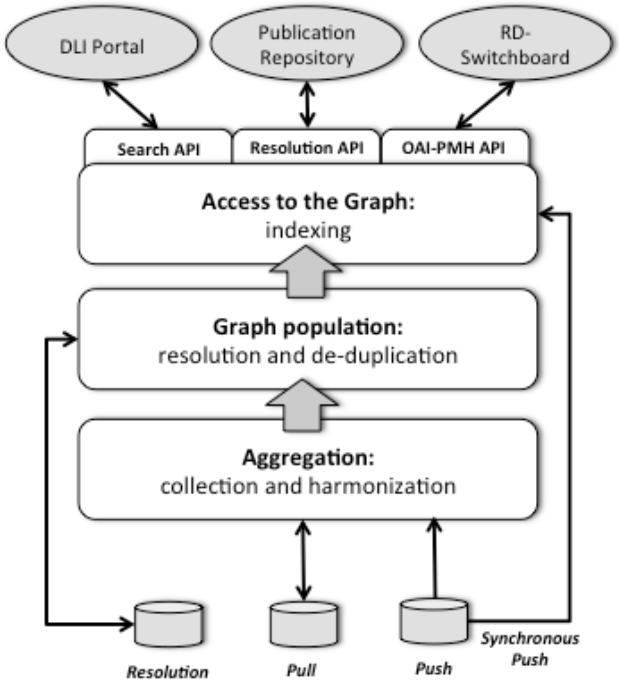

Figure 3 – High-Level Functional Architecture

The Aggregation functionality should be considered as independent from Graph Population and Access to the graph, which instead are typically sequential: the aggregation process is continuous and involves metadata collection and harmonization from an arbitrary number of data sources; graph population takes place at given time intervals, to be decided by Data Curators, generates a graph that is a snapshot of the harmonized metadata currently available, and indexes the graph in different back-ends in order to support a variety of access modalities. Due to the process of graph resolution and de-duplication, graph population can hardly be performed incrementally [21].

Aggregating content from data sources. Data sources are intended as providers interested in feeding object-to-object relationships to the Service. Data sources publish and/or deliver so-called *metadata packages* (records) that encode the description of how a *source object* is interlinked via relationships to a set of *target objects*. Examples are XML files (e.g., DataCite records), JSON files, Excel, CSV, etc. Typically, objects are uniquely identified by a persistent identifier (PID) together with a namespace (e.g., DOI, PMCID, URL) enabling its interpretation. The namespace can be implicit or explicit in the metadata, i.e., when the data source handles objects across several namespaces. Data sources are classified in two main classes: *data publishers* and *data*

providers. A data source is a publisher when it is in charge of storing, curating, and making accessible the original digital objects, i.e., metadata and files. A data source is a provider if it exposes the metadata of digital objects, for use of the Service. For example, DataCite is a data provider for the Service, since it delivers metadata about objects kept on a list of data publishers, while PANGAEA is both a data publisher and a data provider.

Data sources can contribute metadata packages to the Service, hence interact with it, according to four modalities:

- *Bulk Pull*, i.e., the Service harvests a collection of metadata packages from the data source, which offers standard APIs for this;
- *Bulk Push*, i.e., the data source transfers a collection of metadata packages to the Service, which offers standard APIs for this;
- *Synchronous Push,* i.e., the data source of type publishers sends a metadata package to the Service whenever a new package is deposited at the data source; it is a requirement for the data source to make the package visible through the Service APIs in almost-real time; the Service offers APIs enabling this interaction and immediately publishes objects and relationships in the metadata package to third party services; as a consequence, objects in the metadata package are subject to resolution, but bypass de-duplication, which cannot be incrementally applied;
- *PID Resolution*, i.e., the information space includes an object PID without metadata fields (for example the target object of a metadata package); the Service identifies a "resolver" data source (e.g., DataCite, CrossRef, PDB) where the full metadata package can be found, and sends it a request for resolving the PID.

In an optimal world, data sources should deliver metadata packages that conform to DLI exchange format and data model recommended by the Terms of Agreement of the Service. Format and model would be entitled to become a standard or best practice for sharing dataset-literature links. In the initial stage of operation, however, the Service cannot expect data sources to conform to such format. It must therefore provide "metadata harmonization" mechanisms able to map metadata packages, whatever native data model and exchange format they conform to, onto the DLI exchange format. Identifying the DLI data model and exchange format is a core activity in the design of the Service, whose technology should in turn be able to dynamically adapt to their evolution over time.

The Service should keep fine-grain *provenance* information, in order to describe for each object and relationships of the graph their exact origin and status. Provenance gives visibility to all data sources directly (e.g., registered to the Service and providing metadata) or indirectly (e.g., providing content to data sources registered to the service) contributing to the construction of the Service information graph. Moreover, it facilitates the identification of issues in the graph and the prompt identification of errors and relative reporting to the original keeper of the information. Provenance should therefore include information about: data provider and data publisher of objects and relationships, date of collection, modality of collection (e.g., bulk pull, bulk push, synchronous push, PID resolution), and completeness of the metadata (e.g., only PID, full metadata).

Population of the information space graph. The aggregation process will continuously operate over time, maintaining for each data source the corresponding collection of harmonized metadata packages conforming to the DLI format. Independently from this process, the information space graph population process consists in collecting the DLI records relative to all data sources and building a graph out of the links they contain. To this aim, the Service requires tools for converting, i.e., "un-packaging", DLI records onto the objects and relationships of an aggregated graph. For example, Figure 4 illustrates the sub-graph resulting from the un-packaging of a DataCite metadata record relative to the dataset **d1** collected from the data source **R**; the record **d1** contains links to another version of the dataset **d2** and to the supplemented publication **p.** Objects of the graph must then be de-duplicated and resolved, in order to disambiguate the graph and complete its missing information.

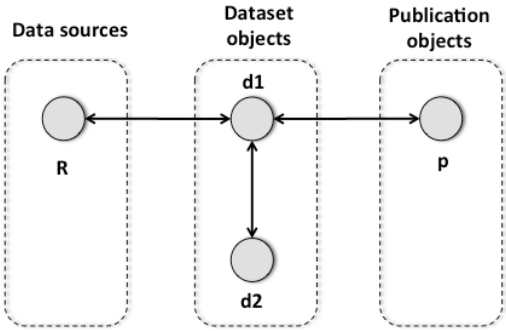

Figure 4 – The graph resulting from "un-packaging" a DataCite metadata record

*De-duplication***.** Different data sources may provide metadata information about the same objects and relationships. This duplication of information generates an ambiguous information space, which may lead to ambiguous search results and statistics. In particular, objects of the same typology (publications or datasets) in the DLI information space may be considered duplicates under two special conditions:

- Identity equivalence: Objects with the same PID (and namespace) collected from different sources; for example metadata packages relative to the same publications (lined to datasets) collected from a publisher and a data center;
- Property equivalence: Objects with different PID and/or namespace but bearing the same values for relevant properties; for example metadata packages relative to the same publication, one collected from EuropePMC (PMCID) and one from a publisher (DOI), or a publication pre-print from ArXiv and its corresponding published version on the publisher site. Equivalence by property matching can be assessed by exploiting the information in the original metadata; e.g., using string matching distances apt to the case, defining acceptance thresholds, and weighing the values of properties such as title, author names, and published year. Moreover, de-duplication must be flanked by human negative or positive feedback in order to improve the results and avoid mistake repetition.

The service requires de-duplication tools, capable of identifying groups of duplicates by matching their properties and merging them into one "representative" object. An

example is shown in Figure 5: two objects **d1** and **d2**, respectively collected from data sources **R1** and **R2**, are associated to publications **p1** and **p2**; the Service matches the pair and identifies they are equivalent, hence generates a representative object **d** which inherits provenance relationships to **R1** and **R2** and relationships to publications **p1** and **p2**. The new object will guarantee visibility to all contributing data sources and also group/centralize relationships to other objects. Similarly, **d** will keep all the pairs PID-namespace of the objects it merges in order to provide the broadest range of accessibility for the object.

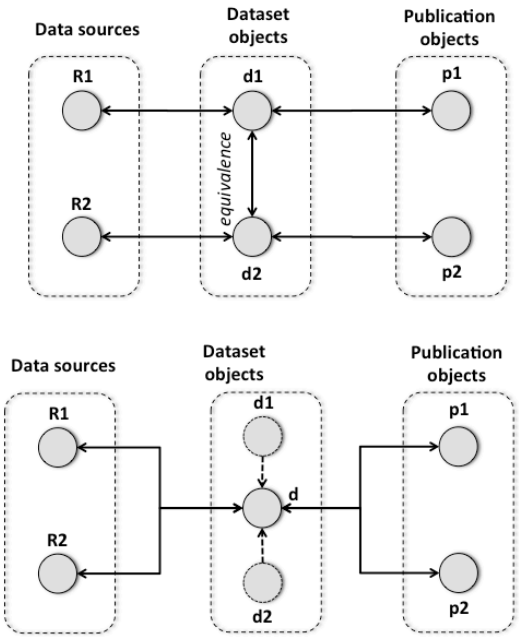

Figure 5 - Deduplication by merge

Resolution. The aggregation process may bring into the graph objects without relative metadata fields. For example, the un-packaging of the DataCite record describing **d1** in Figure 4 brings into the graph two objects **d2** and **p** whose only metadata field is the DOI (or other persistent identifier) currently available in the record. In order to complete such objects with more informative metadata fields, the Service must resolve the PIDs, i.e., refer to a third-party resolver service capable of providing and returning the relative records. In order to provide a flexible functionality, the Service should support a "resolution framework" made of:

- A list of "resolver" data sources: a resolver data source provides APIs capable of resolving PIDs for given namespaces and return the relative DLI record; e.g., for example the DataCite Resolver data source is capable of resolving PIDs such as 10.1000/xyz123 with namespace "doi" and return the relative DLI record;
- A unique entry point to all resolvers, whose API accepts a PID and a namespace and returns the DLI record corresponding to the PID, by trying all resolver data sources compatible with the given namespace.

Access to the information space graph. The Service should offer access to the information space graph by means of several possible strategies, in order to satisfy the needs of different consuming services. Initially, four use-cases have been regarded:

- *Search and browse access to the graph:* the Service should provide APIs to enable third-party services Google-like searches over all objects and browsing of the graph (i.e., navigating from one object to its linked neighbors);
- *End-user access to search and browse functionalities:* the Service should include a portal enabling users to exploit at best the search and browse functionality above;
- *Bulk access to the graph:* the Service should provide APIs to enable third-party services to bulk access objects in the graph using protocols such as OAI-PMH or similar;
- *Resolution of PIDs:* the Service should provide APIs to enable third-party services to run concurrent and high-frequency queries over the graph in order to resolve PIDs, i.e., given a PID the search returns the set of relationships originating from such object available in the index.
- *Data source resolution:* the Service should provide APIs to enable third-party services to "resolve the links" of very large sets of PIDs, relative to the objects of a data source, i.e., given a set of PIDs the Service returns in bulk the set of relationships originating from each PID in the set; this functionality is demanded by data sources willing do enrich their objects with relationships to other datasets or publications available at the Service;
- *LOD exports*: providing a Linked Open Data SPARQL end point and dump, enabling third-party services to search and navigate the graph or download it to integrate it with the Open Data cloud.

3.2 Data model

The conceptual data model of the DLI information space is depicted in **Figure 6**. The model (as well as the corresponding exchange format defined in the following section) is intended as an initial starting point, but is bound to be refined, as new requirements from service stakeholders and consumers will surface. Objects can be of three types, *publications* (intended as scientific literature), *datasets*, and *unknown*. Objects are of type *unknown* when it is not possible to understand if they are publications or a datasets. The data model currently includes *title*, *authors*, and *publication date* of the objects, but this choice may be revised in the future, to meet feedback and evolving requirements from Service users. Relationships between them are directed and bidirectional; e.g., if an object A has a relationship *isCitedBy* to an object B then also the inverse relationship *cites* will be found in the information space. Relationships bear semantics (field *Relation_semantics*), expressed by a label that belongs to a given ontology (*Relation_Semantics_Scheme* field), e.g., DataCite vocabulary.

To model the graph resulting from deduplication, *publications* and *datasets* have corresponding subclasses *representative publications* and *representative datasets*. A *representative publication* (*dataset*) is a *publication* (*dataset*) obtained by merging a number of *publications* (*datasets*) to which is related by a *mergedBy* relationship. In particular, the publications (datasets) merged by a representative publication (dataset)

are virtually "deleted" (*status* field), as well as their outgoing and incoming relationships, to enable a view of the disambiguated graph, made of "active" objects and relationships.

The data model includes the possibility of having objects whose provenance is that of data sources willing to benefit from "data source resolution" functionality. Such objects have status "intersect". Their peculiarity is that they should not be visible to consumers of the graph, since they do not belong to a provider data source, but should still be de-duplicated and merged with other objects, i.e., associated as a *mergedBy* object of representative objects. As described below, data sources "to be resolved" will have access to the representative objects that *merge* (i.e., include) their objects, hence indirectly to all possible relationships inherited by other equivalent objects aggregated by the Service.

Objects and relationships are into the system because either (*i*) they have been pulled as collections from provider data sources, (*ii*) pushed as collections by provider data sources, (*iii*) pushed individually by provider data sources, or (*iv*) obtained by resolving a PID using a resolver service. In order to keep track of their provenance, items are equipped with *provenance information* that consists of:

- A reference to the originating data source;
- The time of ingestion of the item into the system;
- The modality of bringing the item into the system: "pull", "push", "synchronous", "resolved").
- The completions status, described by a field *completion_status* in provenance, which tracks down whether the data source has contributed full object metadata description or only a PID-namespace. This way the Service can identify which objects are "incomplete" and should be subject to subsequent resolution attempts.

When the same items are provided by different data sources (duplicates) and are merged together into one representative item to disambiguate the information space, then the resulting "representative" item keeps provenance information about all the items it merges.

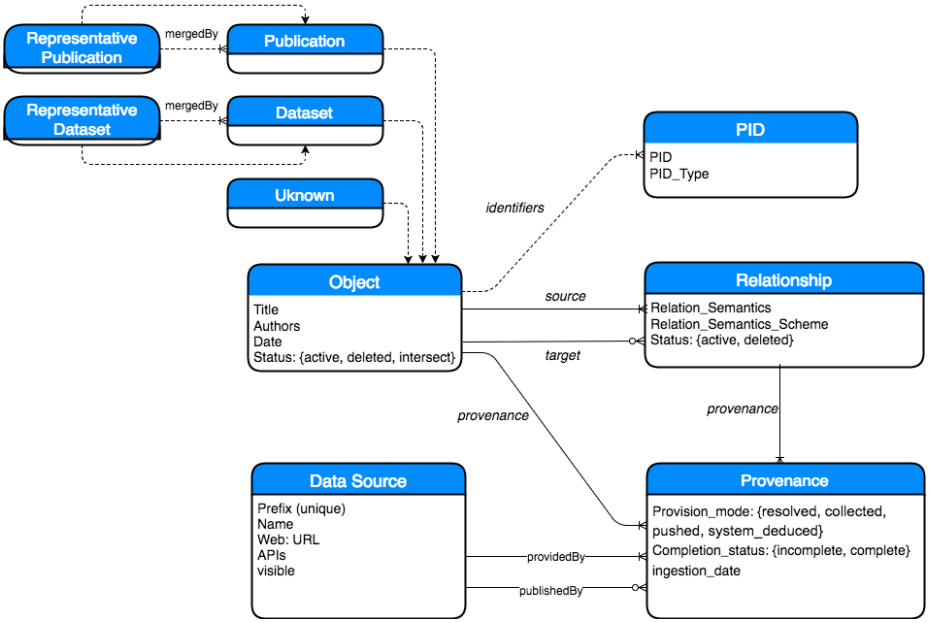

Figure 6 – Conceptual Data Model

4 The Data-Literature Interlinking Service implementation

The first implementation of the Service is powered by the D-NET Software Toolkit ([2],[18]). D-NET is today the platform underlying the production systems of several aggregation infrastructures (e.g., OpenAIRE, EFG/EFG1914[17], HOPE[18], EAGLE[19]) and repository federations (e.g., CEON Poland[20], MINCYT Argentina[21], FECYT Spain[22]). The software is devised to enable the construction and monitoring of aggregative data infrastructures, by combining and orchestrating a set of highly configurable D-NET data aggregation services (and/or third-party web services) into autonomic workflows. D-NET services offer high-level data processing actions by embedding (hiding) the power and complexity of standard and cutting edge data storage and processing technologies. For example, "metadata data storage" functionality is possible via relational databases (Postgres), XML databases (Exist), column stores (MongoDB, HBASE); "data processing" is available via general purpose services, such as XSLT engines, Groovy Engines, Hadoop MapReduce, which are configurable to match different data models and embed customizable algorithms for metadata transformation,

[17] European Film Gateway, http://www.europeanfilmgateway.eu/
[18] Heritage of People's Europe, http://www.peoplesheritage.eu/
[19] Europeana Eagle Project, http://www.eagle-network.eu/
[20] CEON, http://ceon.pl/
[21] Sistema Nacional de Repositorios Digitales de Argentina, http://repositorios.mincyt.gob.ar/
[22] Recolecta, http://recolecta.fecyt.es/

de-duplication, and inference by (text)mining collected files or metadata; "metadata indexing and access" is available via full-text indices (Apache Solr) or graph databases (Virtuoso). Based on their functional and non-functional requirements (e.g., scalability, efficiency) developers can pick and configure the Services they need and build automated workflows that carry out arbitrarily long data processing tasks. D-NET also supplements system administrators with tools for autonomic monitoring the consistency and quality of workflows [19] and the data resulting from their execution [14].

The Service adopts D-NET to implement the functionalities described in the previous section, except for the following, which will be increasingly added to the system over time:

- Data sources are only of type "pull" and "resolution";
- The semantics of relationships is limited to the subset of DataCite (i.e., no support for multiple vocabularies): *references, cites, isSupplementTo, isReferredBy, isCitedBy, isSupplementedBy* and otherwise mapped onto the *unknown* value;
- De-duplication is implemented only at the level of PID equivalence;
- The information graph is not yet available as LOD.

The reason behind these "limitations" is merely pragmatic. The BETA version of the service had to prove its feasibility, in terms of scalability and efficiency, and its benefits, by delivering enough content to drive and inspire real use-cases. In this initial phase, this led to the deployment of D-NET aggregation services enabling data curators to test the efficient population of the information space by collecting content from heterogeneous data sources. Introducing "push" sources, hence the complexity of an FTP entry point, was not necessary at this stage; similarly, the adoption of a richer vocabulary, would have only complicated the aggregation process. For the same reason, the de-duplication process has been limited to PID equivalence, which is a sub-case of property equivalence where only PID and namespace properties are regarded in an identity match. The introduction of property equivalence by similarity will be introduced once the Service will be deployed in production, to likely introduce content quality issues to be solved by properly configuring similarity functions and thresholds. Finally, LOD export is a problem on its own as it requires the identification of an RDF schema that matches fit-for-purpose data model while satisfying LOD cloud integration needs. As such, it is the kind of action to be addressed once the Service is in production and a stable data model of reference is in place (e.g., output of Scholix initiative).

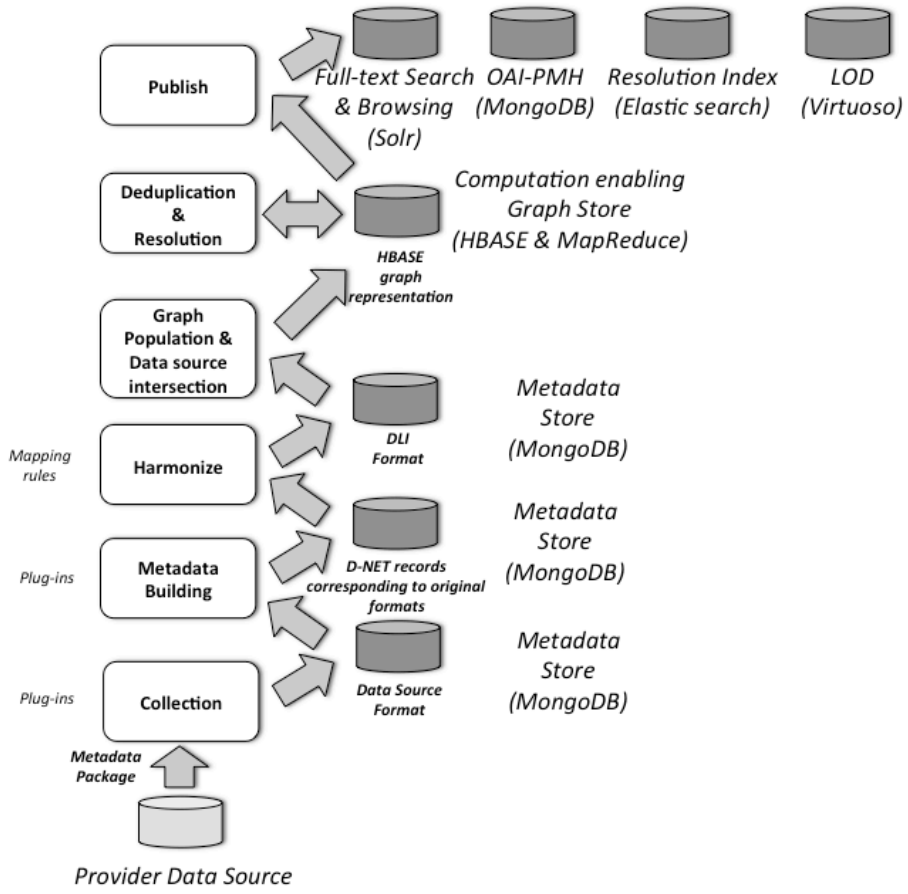

Figure 7 - Data source aggregation workflow.

The following sections describe how the data source aggregation workflow has been implemented in D-NET as part of the Service functionality.

Aggregating content from data sources The system offer administrative user interfaces for handling the registration of data provider data sources and configuring/scheduling the aggregation of their metadata content. Data sources register to the Service by submitting a profile describing their general properties (e.g., name, location, etc.) and technical interoperability properties (e.g., data collection APIs, data collection modality). Each registered data source is associated with an autonomic workflow (see Figure 7) that, at given time intervals, collects its metadata packages and processes them (harmonizes them) to generate a corresponding DLI record and then populate the DLI information space graph:

- *Collection*: D-NET includes a Collector Service capable of handling a number of standard protocols, such as OAI-PMH and FTP, and proprietary protocols, such as specific-service REST APIs, or local file system directories; standard protocols expect metadata records to be provided as individual files, while proprietary

protocols may handle very specific data provision formats, e.g., zip files, CSV, HTTP responses, etc. The Collector Service manages a number of plug-ins (to be selected in the setting up of aggregation workflows of a data source) and can therefore be extended to include custom collection scenarios

- *Metadata building*: D-NET includes a Metadata Builder Service capable of mediating between external formats and the internal D-NET XML format for metadata records; the service handles files collected by the Collector Service and maps them one-to-one or one-to-many into corresponding D-NET metadata records; similarly to the Collection Service, the Metadata Builder Service manages a set of plug-ins, hence can be extended with new ones to handle custom scenarios, and the proper plug-in is to be selected in the setting up of the data source aggregation workflow;
- *Harmonization*: D-NET includes a Transformation Service that handles the transformation of a set of D-NET metadata records onto a corresponding set by means of transformation rules, expressed in a D-NET transformation language; the rules express a mapping from a metadata scheme onto another scheme, both structurally and semantically, e.g., conversions between vocabularies and value formats [2].

Data sources can specify whether or not they are also publisher data sources, in order to instruct the harmonization on how to assign the proper *publishedBy_datasource* and *providedBy_datasource*. In the case of provider data sources, such as "aggregators" of content (e.g., DataCite), Service curators will introduce specific workflows for the integration of the publisher data sources "behind" the provider data source (e.g., data archives nourishing DataCite. Harmonization mapping for such provider data sources will be in charge of identifying for each incoming metadata package the respective publisher data source, in order to keep an exhaustive provenance record.

For the generic provider data source, the relative harmonization workflow makes use of D-NET's MetadataStore Service, Transformation Service, and HBASE Service (see Figure 7). Initially, metadata packages are cached in their native format (e.g., XML, CSV, Excel), then transformed, given a set of transformation rules, from such format onto an internal XML format called "DLI" shown in Table 1.

Table 1 - DLI record structure

```
DLI_ID: % obtained as <PID_type>::<PID>
PID
PIDType: % from a vocabulary doi, PMCID, ncbin, pdb, etc.
authors
title
date
type: {publication, datasets, unknown}
provenance*
    providedBy_datasource
    publishedBy_datasource
    provision_mode: {resolved, collected, pushed, system_deduced}
    ingestion_date
    completion_status: {incomplete, complete, failed_to_resolve}
         % incomplete => type, authors, title, and date fields
         % are empty
relationship*
    target_object_type: {publication, dataset, unknown}
```

```
      target_object_title  % to be used as anchor label
      target_object_PID:
      target_object_PIDType % doi, PMCID, others
      target_object_DLI_ID
      provenance*
           publishedBy_datasource
           providedBy_datasource
           provision_mode:  {resolved,  collected,  pushed,  system_de-
duced}
           completion_status: {incomplete, complete, failed_to_resolve}
           ingestion_date
           relationship_completion_status: {incomplete, complete}
                %  incomplete =>  type and title fields are empty
           semantics
                % from DataCite relationships vocabulary or "unknown"
```

The DLI exchange format includes all information described in the data model, but also introduces some redundancy in order to become self-explanatory (e.g., enabling interpretation of target objects without necessarily accessing them).

Population of the information space graph Independently from the data source aggregation workflows, at given time intervals a population workflow will transform DLI records of all data sources into objects and relationships of the graph, which are encoded as rows of an HBASE column store. The choice of HBASE is due to its write-mode efficiency, since the graph will be often rebuilt from scratch, its scalability, as it is expected the graph will grow up to tens of millions of objects, an processing performance, needed to implement de-duplication and resolution phases.

The graph thus built may feature duplicated objects and objects whose *completion_status* is "*incomplete"*.

Deduplication of the graph is implemented as a sequence of MapReduce jobs, capable of 1) identifying groups of objects whose DOI, titles, authors and acceptance date have a high similarity distance, 2) merge groups of similar objects to produce a representative object (i.e., creation of new representative object, virtual deletion of objects it merges, update of relationships to provenance data sources to point the representative object). As things stand today, identification of similar objects is only based on equivalence of PIDs.

Resolution of PIDs is implemented as a Map-only job that finds incomplete objects, identifies the respective resolver service based on the object PID namespace, and tries to fetch the missing metadata fields. PID namespaces are often derived as default values from the provider data source (e.g., Elsevier data source today provides only links from publication DOIs and PDBs) and in some cases derived by the PID format. The result of such operation, be it successful or not, is tracked by the system and ends up enriching the provenance information of the given objects. A resolver is identified by the PID namespace it can handle and it is a library that offers a method of the form: *resolve(PID, PID_type)*. The Service maintains a registry of resolver libraries, implemented as plugins, and exposes APIs for internal use that resolve PIDs of a given PID type by identifying the best resolver available. Table 2 lists the resolvers currently in use by the service.

Table 2 - *Resolvers currently integrated by the Service*

Name	Web Site	PID type
ANDS	*http://www.ands.org.au/*	*ANDS URLs*
CrossRef	*http://www.crossref.org/*	*DOIs*
DataCite	*https://www.datacite.org/*	*DOIs*
OpenAIRE	*http://www.openaire.eu*	*OpenAIRE identifiers*
PubMed	*http://www.ncbi.nlm.nih.gov/pubmed*	*PMC identifiers*
RCSB	*http://www.rcsb.org/*	*PDBs*

Access to the information space graph. Once the information graph has been refreshed, a further workflow will be fired to ensure the graph is "published" according to all expected formats and back-ends. The workflow executes a MapReduce job over the HBASE graph representation to generate DLI exchange format records (post duplicate identification and object resolution) corresponding to the disambiguated graph, i.e., discarding objects and relationships whose status is not "active". The resulting DLI records as stored in an HDFS file system, which allows an efficient (parallel) reading of the records in order to send them to the different back-ends: Solr Index Service serving portal search and browse functionality, OAI-PMH Publisher Service, LOD Service, and the ElasticSearch index used to implement the resolution functionality. Today, the status of the expected functionalities is as follows:

Search and browse access to the graph: the Service provides a web portal (http://dliservice.research-infrastructures.eu) for end users to (full-text) search and browse relationships between datasets and publications and to visualize statistics on the distribution of such relationships (e.g., per data source, per type, etc.);

Bulk access to the graph: the Service supports OAI-PMH APIs to export the DLI information space in the shape of DLI records towards interested third-party services (http://dliservice.prototype.research-infrastructures.eu/oai)

Data source resolution: the Service implements workflows capable of aggregating a data source willing to benefit from "resolution" functionality; such workflows collect object PIDs from a data source and ingest them in the graph with status "intersect", without resolving them. After de-duplication, the workflows identify the set of "intersect" objects for the data source that have been merged with others in the graph, hence have inherited links to other objects from equivalent objects in the graph. Such subset is then used to construct an OAI-PMH set of DLI XML records exposed via the Service APIs, from which third-party services (including the requesting data source) can collect the enriched records;

Resolution of PIDs: The PANGAEA data center team is working to extrapolate the current PANGAEA linking service[23] into a generally usable linking service that will enhance the current Service content provision system. The service will offer PID-

[23] *Elsevier and PANGAEA Take Next Step in Connecting Research Articles to Data,* http://www.prnewswire.com/news-releases/elsevier-and-pangaea-take-next-step-in-connecting-research-articles-to-data-99533624.html. See also [7].

resolution APIs and be optimized for high-volume read access by science publishers and bibliometrics service providers. It will be based on Elasticsearch[24], hosted in the Amazon EC2 cloud, and will provide linking information and render metadata badges that can be embedded into article publisher's web pages to show linked data sets (see [7]). Based on this service, a new section of the DLI portal will display linking statistics based on Elasticsearch aggregations using visualization features of Kibana[25]. The REST APIs will accept PIDs and relative PID_Type and return the list of relationships relative to the given PID, i.e., one entry for each relationships outgoing the object, as described by the format in Table 3. The entries contain minimal information, enough to detect the nature of the target object (i.e., publication, dataset, unknown) and display it via user interfaces with a title and list of authors.

Table 3 – Resolution of PIDs: response format

```
Source_object: <PID_type>::<PID> %PID to be resolved
Target)object: <Target_object_PID>:<Target_object_PIDType>
Target_object_title
Target_object_authors
Target_object_type: {publication, datasets, unknown}
```

5 Service operation

Service operation incudes all the activities needed to grow an up-to-date and high-quality information graph in order to serve a number of consumers. In the following sections, the current status of both lines of activities is presented. Finally, the forthcoming upgrades of the software in order to meet the overall DLI Service functional requirements defined in section 3 will be described.

5.1 Data curation

Data curation activities are carried out by administrators in charge of the following actions:

- *Registration of data sources*: admins set up the relative harmonization rules and, for pull data sources, configure the time-schedule of the harvesting and the specific access protocol handler.
- *Data sources workflow management*: Admins are also in charge of setting up the workflows importing the list of publisher data sources that are indirectly providing content through a common provider data source (e.g., DataCite). As mentioned above, the information (Metadata) relative to such data sources is necessary to generate complete provenance information for the objects and relationships.
- *Quality control*: admins are also in charge of ensuring the quality of the generated information space. Controls are implemented via D-NET Monitoring Services

[24] Elasticsearch, https://www.elastic.co/products/elasticsearch
[25] Kibana, https://www.elastic.co/products/kibana

(DataQ [14]), designed to (i) collect observations sent by the D-NET workflow engine during the execution of the aggregation workflow steps; and (ii) perform controls to verify the consistency of such observations over time. Examples of controls can be:

- Data source consistency: the number of links collected from a given data source should increase over time;
- Information space consistency: the number of links in the graph should increase when moving from a graph to its new version;
- Information space access alignment: since the information space is made available across different back-ends to support a variety of protocols (OAI-PMH, REST access, Web portal), when a new graph is generated and casted according to different physical representations via MapReduce jobs the system must control the alignment of numbers across the back-ends, e.g., the same number of links, the same number of publications, the same number of datasets.

Quality control is currently not active, and its operation is envisaged once the DLI service will be deployed as a production system in early 2017.

Currently, the prototype includes relationships and objects from the data sources reported in **Table 4**.

Table 4. Objects and relationships contributed by data source at the moment of writing (1 January 2017). At the time of writing, the Service holds 2.8M objects with 9,8M links between them (4,9M bi-directional relationships), out of which 1,4M reach a publication and 7M reach a dataset.

Content provider	**Contributed links**	**Referred publications**	**Referred datasets**	**Referred unknown type objects**
Datasets in DataCite	7,273,251	22,517	798,897	479,463
OpenAIRE Resolver	0	6,625	0	0
Cambridge Crystallographic Data Centre	1,067,074	235,480	526,405	442
IEDA	1,498	398	491	27
OpenAIRE	17,064	6,625	2,669	479
IEEE	94	16	47	0
Elsevier	138,976	7,383	65,849	1
3TU Datacentrum	432	40	174	137
Thomson Reuters	48,466	4,206	24,326	8
PubMed Resolver	0	7,582	0	0

Australian National Data Service	19,078	473	3,546	2,176
EuropePMC	1,032,868	102,180	0	406,308
PANGAEA	894,598	12,603	271,361	33,724
Mendeley Data and published articles	36	15	18	0
DataCite Resolver	0	0	17,907	0
Springer Nature	60,392	7,363	28,295	14
RCSB	175,648	44,862	88,702	79
CrossRef	0	313,462	163	0
ICPSR	16,120	3,791	1,765	51

5.2 Real use-case consumers

Web portal The web portal (see Figure 8) interacts with the Solr full-text index in order to support web users with search, browse, and navigation of objects in the information space graph. The portal also allows web users to navigate relationships between data sources and objects, be them datasets or publications, and between objects themselves.

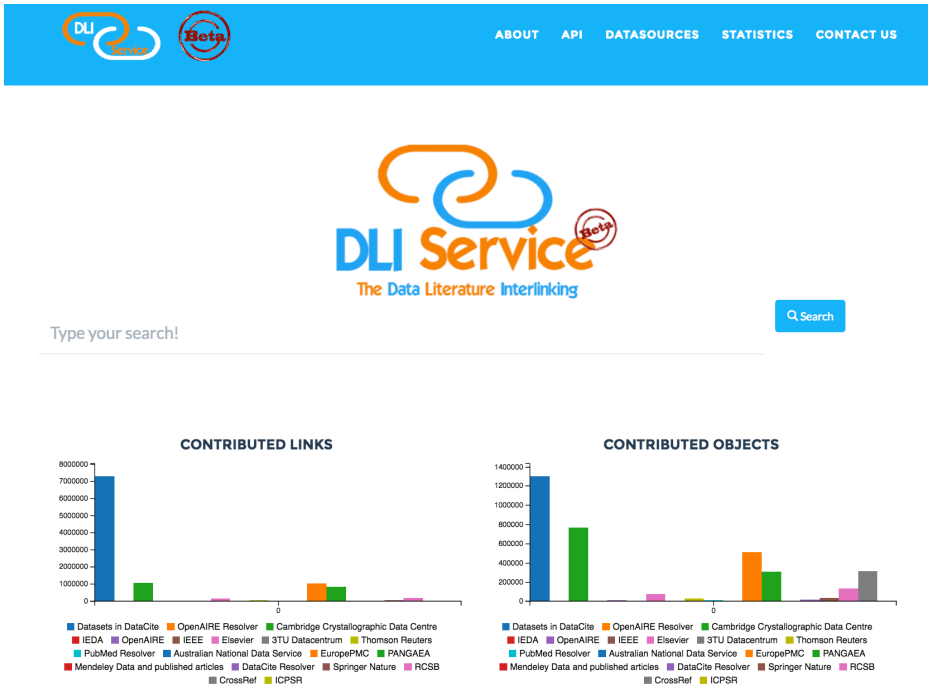

Figure 8 - DLI Service Portal

The portal also plays the function of entry point for general presentation of the service, for data source map and statistics (see Figure 9), and for public APIs.

Figure 9 - DLI Service: information space statistics

Data source resolution The DLI Service has today one data source registered and in the need of data source resolution functionality. The data source is CHORUS[26] and provides 102,000+ DOIs of publications via OAI-PMH APIs. The Service collected the publication DOIs, processed them through the workflows described above, and generated a corresponding OAI-PMH Set of 2,205 records, relative to CHORUS publication with relative links to datasets.

PID resolution for third-party services Even though PID resolution APIs are under construction, Scopus is currently exploring direct access to the DLI Service index in order to resolve publication DOIs and be able to show to end-users the list of datasets related to them. The Library of University of Illinois (Urbana-Champaign) has already implemented a working prototype of this functionality, today visible at: http://search.grainger.illinois.edu/searchaid2/mdfc/dataset.asp?typeofsearch=author&searcharg=Hunze+S.&OPERATE=GO.

Research Data Switchboard The RD-Switchboard has been described in Section 2.2. Integration between the Service and the RD-Switchboard takes is bidirectional: RD-

[26] CHORUS *Advancing Public Access to Research*, http://www.chorusaccess.org/

Switchboard system instances running world-wide can collect dataset-literature links from the DLI Service, providing added value to all the RD-Switchboard adopters; the DLI Service can collect dataset-literature links from running instances of RD-Switchboard, by registering them as provider data sources.

5.3 Forthcoming actions

The Service is under continuous development in order to achieve the objectives identified in the definition of the functional requirements in Section 3. By functional area, the following plan has been devised:

Aggregation of data sources. The prototype will be completed to allow "push" modality for data sources. In practice, push data sources can deposit on dedicated FTP folders, from which DLI aggregation workflows will collect metadata packages.

Moreover, the "synchronous" push option will be experimented with the PANGAEA data center. The Service will offer APIs to allow authorized data sources the direct ingestion of DLI metadata records in the information space graph. The action will ensure researchers immediate visibility of their depositions at local (community-recommended) archives to the wider community, via the DLI Service. Records thus "injected" will not be subject to de-duplication and resolution until they will be regularly collected from the data source via the respective aggregation workflow.

Population of the information graph. Population of the graph requires more sophisticated de-duplication algorithms, exploring object similarity beyond the equivalence of PIDs. D-NET offers de-duplication Services [4] already in use by the OpenAIRE infrastructure production system, which will be deployed and configured to adapt to the DLI data model and identifying equivalent objects based on properties such as titles, author names, and publication year.

Access to the information space graph. Several actions need to be undertaken to facilitate interoperability with other systems. Exposing content according to LOD is certainly crucial, in order to exploit at best the integration with the Linked Data cloud and enable LOD services to benefit from the DLI graph. The process will follow the methodology adopted by the OpenAIRE infrastructure production system, based on D-NET software, which supports MapReduce jobs for parallel reads of the de-duplicated graph and parallel writes onto a Virtuoso installation [20].

Moreover, the adherence to the Scholix link exchange format is in the plan, in order to establish data exchange interactions with the DataCite and CrossRef systems.

Finally, most importantly, by January 2017 the Service will undergo a migration to production level and be deployed at ICM data center, over the hardware infrastructure of the OpenAIRE infrastructure. The migration will ensure production-level availability and reliability of the Service.

5 Conclusions

This paper describes the work carried out by the joint ICSU-WDS and RDA Working Group "Publishing Data Services" (PDS-WG) that has enjoyed the support of OpenAIRE, CrossRef, DataCite, ANDS, PANGAEA, Elsevier, and many others. The key issue which the WG has addressed is the fragmentation of solutions and practices to link research data and the literature. Such links are beneficial for researchers in many ways, helping to increase visibility and discoverability of relevant research output, placing data in context to enable re-use, and supporting credit attribution mechanisms to incentivize researchers to share their data in the first place.

However, the current landscape is very fragmented, with many different organizations having knowledge about a small subspace of the "universe" of all links. That means that currently we cannot readily construct a full graph of links - and thus we are not utilizing to its full potential all the knowledge that exists about how research data and the literature are connected. Overcoming this problem of fragmentation, on both a technical and a social level, is the challenge which the WG has set out to address.

At the end of its predetermined 18-month lifetime, the main outputs of the WG are twofold. First, in a synergic effort with OpenAIRE, the WG has created an open, universal Data-Literature Interlinking (DLI) Service that aggregates, harmonizes, completes, and offers access to links between the scholarly literature and research data. While developed as a prototype, the DLI service is fully operational and can be queried to get access to a body of almost 9,8 million links aggregated from a variety of sources.

The technical development path of the DLI service reflects the WG's principles of openness, inclusivity, quality, provenance, and domain-agnosticism – as well as a pragmatic, "ground-up" approach to develop software in a test-and-learn approach that allows for continuous refinement of the system and the underlying data model. By establishing this service, the PDS-WG is demonstrating in a direct, hands-on way how the current situation of fragmented sets of links can be improved to realize a universal, one-for-all service architecture with common standards to the benefit of all stakeholders in the research data landscape.

The second main output of the WG are the "Scholix" (short for Scholarly Link Exchange) framework, which constitute an aspirational vision and a set of practical guidelines for a long-term infrastructure to support the sharing, exchange, and aggregation of links between research data and the literature. The framework will be reported on in more detail elsewhere, but it will be useful to underline in the present context that the development of the DLI has been a catalyst in engaging the right stakeholders to formulate such a vision, and served as a sand-box environment to learn more about the challenges regarding data acquisition, duplication, and information modeling. Also, the DLI Service will be further developed to become an integral part of the envisioned Scholix infrastructure.

Acknowledgements. The authors would like to thank the PDS-WG members and representatives from CrossRef, DataCite, The National Data Service, ORCID, The Research Data Alliance, ICSU World Data Systems, and the RMap project for many valuable discussions and constructive interactions. This work is partially funded by the EU projects RDA Europe (FP7-INFRASTRUCTURES-2013-2, grant agreement: 632756) and OpenAIRE2020 (H2020-EINFRA-2014-1, grant agreement: 643410).

References

1. *Publishing Data Services Working Group Case Statement,* https://www.rd-alliance.org/filedepot/folder/114?fid=239
2. Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela L, Castelli D, Pagano, P. (2014). The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. Program: electronic library and information systems, 48(4), 322-354.
3. Castelli, D., Manghi, P., & Thanos, C. (2013). A vision towards scientific communication infrastructures. International Journal on Digital Libraries, 13(3-4), 155-169.
4. Manghi, P., Mikulicic, M., Atzori, C. (2012). De-duplication of aggregation authority files. International Journal of Metadata, Semantics and Ontologies, 7(2), 114-130.
5. Smit, E. (2011). Abelard and Héloise: Why Data and Publications Belong Together. D-Lib Magazine, volume 17. DOI: 10.1045/january2011-smit
6. Pepe, A., Goodman, A., Muench, A., Crosas, M., Erdmann, E. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLOS One. DOI: 10.1371/journal.pone.0104798
7. Aalbersberg IJ. J., Dunham J., Koers H. (2011). Connecting Scientific Articles with Research Data: New Directions in Online Scholarly Publishing. Proceedings of the 1st ICSU World Data Systems Conference.
8. Callaghan, S., Tedds, J., Lawrence, R., Murphy, F., Roberts, T., Wilcox, W. (2014). Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples. International Journal of Digital Curation. DOI: 10.2218/ijdc.v9i1.310
9. Kobourov, Stephen G (2012). Spring embedders and force directed graph drawing algorithms. *arXiv preprint arXiv:1201.3011* (2012).
10. Manghi, P., Bolikowski, L., Manold, N., Schirrwagen, J., & Smith, T. (2012). Openaireplus: the european scholarly communication data infrastructure. D-Lib Magazine, 18(9), 1.
11. *The RMAP project, white paper, http://rmap-project.info/rmap/wp-content/uploads/RMap_Project_Overview_Revised_Final.pdf*
12. Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, et al. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE 9(12): e115253. doi:10.1371/journal.pone.0115253
13. Burton, A., and Koers, H. (2016). Interoperability Framework Recommendations. Available online https://sites.google.com/a/scholix.org/scholix/guidelines

14. Andrea Mannocci and Paolo Manghi, DataQ: A Data Flow Quality Monitoring System for Aggregative Data Infrastructures, 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016. Proceedings. Lecture Notes in Computer Science, Springer 2016.

15. Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., & Schindler, U. On Bridging Data Centers and Publishers: The Data-Literature Interlinking Service. In Metadata and Semantics Research (pp. 324-335). Springer International Publishing, 2015.

16. Hanson, K. L., DiLauro, T., Donoghue, M. The RMap Project: Capturing and Preserving Associations amongst Multi-Part Distributed Publications. In Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries (pp. 281-282). ACM. 2015

17. Artini, M., Atzori, C., Bardi, A., La Bruzzo, S., Manghi, P., & Mannocci, A. (2015). The OpenAIRE Literature Broker Service for Institutional Repositories.D-Lib Magazine, 21(11), 3

18. Bardi, A., Manghi, P., & Zoppi, F. Coping with interoperability and sustainability in cultural heritage aggregative data infrastructures. International Journal of Metadata, Semantics and Ontologies, 9(2), 138-154, 2014

19. Artini, M., Atzori C., and Manghi P.. Keeping your aggregative infrastructure under control. Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on. IEEE, 2014.

20. Vahdati, S., Karim, F., Huang, J. Y., Lange, C.. Mapping Large Scale Research Metadata to Linked Data: A Performance Comparison of HBase, CSV and XML. In Research Conference on Metadata and Semantics Research (pp. 261-273). Springer International Publishing, 2015

21. Atzori C., gDup: An Integrated and Scalable Graph De-duplication System. PhD Thesis. Department of Informatics and Engineering, University of Pisa. 2015 https://etd.adm.unipi.it/t/etd-05092016-090250