D-Lib Magazine

The Magazine of Digital Library Research

D-Lib Magazine

January/February 2017
Volume 23, Number 1/2
Table of Contents

The Scholix Framework for Interoperability in Data-Literature Information Exchange

Adrian Burton
Australian National Data Service, Melbourne, Australia
http://orcid.org/0000-0002-8099-7538
adrian.burton@ands.org.au

Hylke Koers
Elsevier, Amsterdam, The Netherlands
http://orcid.org/0000-0001-6538-7590
h.koers@elsevier.com

Paolo Manghi
Institute of Information Science and Technology — CNR, Pisa, Italy
http://orcid.org/0000-0001-7291-3210
paolo.manghi@isti.cnr.it

Markus Stocker
PANGAEA/MARUM, University of Bremen, Germany
http://orcid.org/0000-0001-5492-3212
mstocker@marum.de

Martin Fenner
DataCite, Hannover, Germany
http://orcid.org/0000-0003-1419-2405
martin.fenner@datacite.org

Amir Aryani
Australian National Data Service, Melbourne, Australia
http://orcid.org/0000-0002-4259-9774
amir.aryani@ands.org.au

Sandro La Bruzzo
Institute of Information Science and Technology - CNR, Pisa, Italy
http://orcid.org/0000-0003-2855-1245
sandro.labruzzo@isti.cnr.it

Michael Diepenbroek
PANGAEA/MARUM, University of Bremen, Germany
http://orcid.org/0000-0003-3096-6829
mdiepenbroek@pangaea.de

Uwe Schindler
PANGAEA/MARUM, University of Bremen, Germany
http://orcid.org/0000-0002-1900-4162
uschindler@pangaea.de

Corresponding Authr: Paolo Manghi, paolo.manghi@isti.cnr.it

https://doi.org/10.1045/january2017-burton

Abstract

The Scholix Framework (SCHOlarly LInk eXchange) is a high level interoperability framework for exchanging information about the links between scholarly literature and data, as well as between datasets. Over the past decade, publishers, data centers, and indexing services have agreed on and implemented numerous bilateral agreements to establish bidirectional links between research data and the scholarly literature. However, because of the considerable differences inherent to these many agreements, there is very limited interoperability between the various solutions. This situation is fueling systemic inefficiencies and limiting the value of these, separated, sets of links. Scholix, a framework proposed by the RDA/WDS Publishing Data Services working group, envisions a universal interlinking service and proposes the technical guidelines of a multi-hub interoperability framework. Hubs are natural collection and aggregation points for data-literature information from their respective communities. Relevant hubs for the communities of data centers, repositories, and journals include DataCite, OpenAIRE, and Crossref, respectively. The framework respects existing community-specific practices while enabling interoperability among the hubs through a common conceptual model, an information model and open exchange protocols. The proposed framework will make research data, and the related literature, easier to find and easier to interpret and reuse, and will provide additional incentives for researchers to share their data.

Keywords: Scholarly Communication, Data Article Interlinking, Interoperability, Open Science, Research Data

1 Introduction

With the increasing role of data in the research enterprise, there is increasing interest in getting and exposing the links between publications and the underlying data [3, 4, 5].

Links between research data and literature are considered desirable since they increase the visibility, discovery, and retrieval of both the data and the literature [9]. An ideal scholarly communications system has been proposed where a reader of a journal article would be able to follow such a link to the data that supports the findings of the article. Conversely users of a dataset would use such links to find previous research literature based on that dataset [10]. The comprehensive global propagation of such bi-lateral links throughout the scholarly communications system (i.e. throughout publishers of data and literature) is an aspiration since it would significantly aid the scientific method by improving discovery of and access to related knowledge and underpinning observations.

Similarly, linking literature and data supports credit attribution mechanisms. Better attribution of credit for published data has been proposed as an important motivator for researchers to publish data in the first place [12]. The impact of a dataset may be in part deduced from the references to it in the literature or indeed from the number of other datasets derived from it [11, 15, 16]. Here again the comprehensive global view of these links is an aspiration since it would provide the scaled-up information pool needed for research impact analysis.

In many areas of science a crisis of reproducibility has been declared [14]. The integrity of conclusions reported in scholarly literature can be improved in part by links between literature and its underpinning data. Recent surveys report that lack of data routinely contributes to irreproducibility of research findings [13]. Vice versa, literature provides valuable contextual information needed for proper data reuse. In her recent book, Borgman argues that to "interpret a digital dataset, much must be known about the hardware used to generate the data, whether sensor networks or laboratory machines; the software used to encode or analyze them, whether image processing or statistical tools; and the protocols and expertise necessary to bring them together" [8]. Literature provides such contextual information needed for proper data interpretation, reproducibility, as well as repurposability. A standardised universal approach to linking data and literature is again an aspiration since sector-wide interoperable information systems to capture the links will encourage data citation as default behaviour, which would go some way to addressing the reproducibility of science.

Although linking literature and data (and having a joined-up view of those links) is widely recognized as being of great value, current solutions are clearly not realizing the potential. Publishers, data centers, and indexing services have for some time now established bidirectional links between research data and the scholarly literature. However, these links are bilateral agreements between two (or at most few) organizations. The data-literature linking between PANGAEA, a Data Publisher for Earth and Environmental Science, and selected Elsevier journals is an example for a Publisher-Data Center linking initiative. In addition to being bilateral, such linking lacks an industrial standard.

Unfortunately, the plethora of bilateral agreements and implementations between organizations are sub-par and the current situation in literature-data linking is overall unsatisfactory. Bilateral agreements come with a number of undesired side-effects. First, most publishers, data centers, repositories, and infrastructure providers remain disconnected. Second, the inherent heterogeneity resulting from the considerably different individual agreements and practices hinders global interoperability. Examples of the heterogeneity of practices include different PID systems, e.g. HTTP URIs such as Digital Object Identifiers (DOI) and accession numbers such as Life Science Identifiers (LSID); different ways of referencing data, e.g. with formal citations or in-text references; and different timepoints at which data is cited, e.g. pre-/at-/post-publication.

Scholix (SCHOlarly LInk eXchange) is a community and multi-stakeholder driven effort to address the current unsatisfactory situation in data-literature information exchange. Scholix is a high level interoperability framework aimed at increasing and facilitating

exchange of information about the links between data and scholarly literature, as well as between data. The framework proposes an overall cohesive vision and approach to bring together the existing initiatives. At present such aspirations are hampered by the lack of agreed ways of expressing links between data and literature and common ways of exposing those links to a potentially comprehensive global information system, known as literature to literature links. The Scholix framework targets the latter by establishing an agreed interoperability framework and thus creating a pull factor to incentivise the former.

2 Scholix Framework

Scholix is an overarching framework for existing technical initiatives that individually address parts of the overall problem that is hindering better linking between data and the literature. Figure 1 presents some of the currently existing technical and social initiatives. The framework proposed by Scholix is a technical solution building from and coordinating current technical initiatives. Though grounded in technical challenges, Scholix is informed by and contributes to social change needed to improve literature-data linking, for instance by enabling the realization of third-party services.

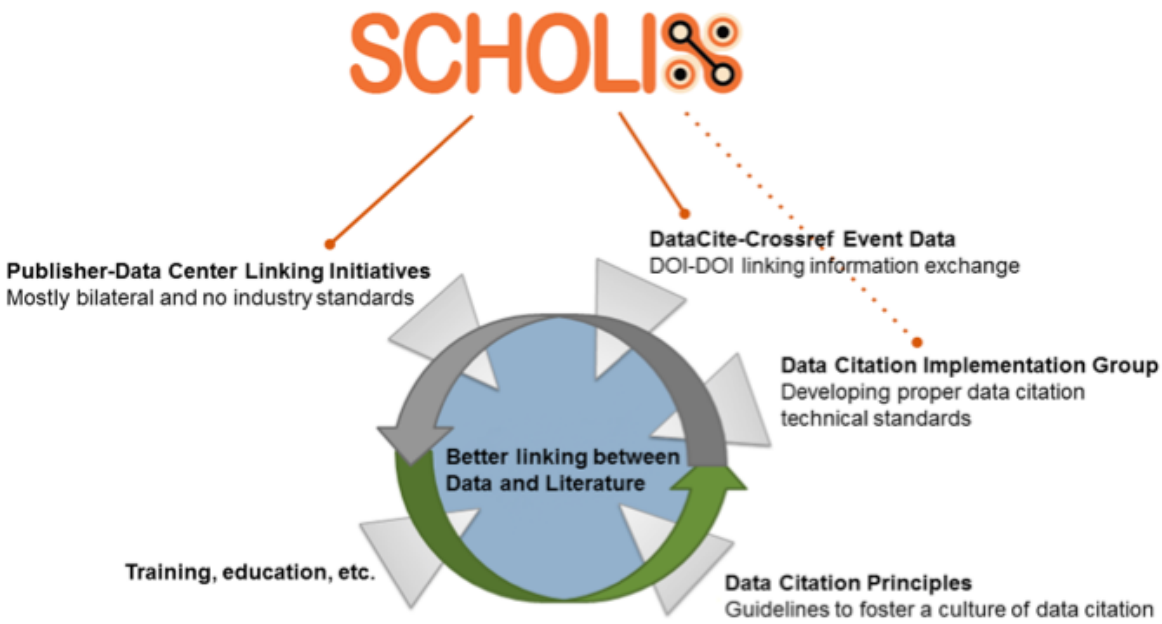

Figure 1: Scholix: An overarching framework for existing technical initiatives that individually address parts of the overall problem that is hindering better linking between data and the literature.

Scholix is not a product, service, or infrastructure. It is a conceptual framework for interoperability. Scholix maintains an evolving lightweight set of guidelines rather than a normative standard. The framework is developed on consensus from various stakeholder groups in the research data landscape, including data centers, publishers, Crossref, DataCite, OpenAIRE, and many others. It is a result of the RDA/WDS Publishing Data Services working group, with endorsement by ICSU-WDS and pending endorsement by RDA.

Aiming at standardizing the exchange of data-literature link information between scholarly infrastructure providers, Scholix currently documents: (1) a multi-hub vision, (2) a shared conceptual model and information model for scholarly link representation, and (3) recommendations for standards, exchange formats, and protocols.

2.1 Multi-hub Vision

The Scholix framework acknowledges the existence of natural hubs for literature-data and data-data link information and proposes informal standardisation of information exchange between hubs at the wholesale level to enable a comprehensive joined view.

Hubs are (existing) services that collect and aggregate information about links from their respective communities. Example hubs for the communities of data centers, repositories, and journals include DataCite, OpenAIRE [1, 2], and Crossref, respectively. Figure 2 provides a schematic overview of the proposed multi-hub infrastructure. As we can see, natural hubs such as Crossref, DataCite, and OpenAIRE each aggregate information about links from a community. For instance, DataCite is an existing infrastructure that aggregates information from the community of data centers. Thus, community actors report information about links to their respective hub. Only hubs are connected and share information.

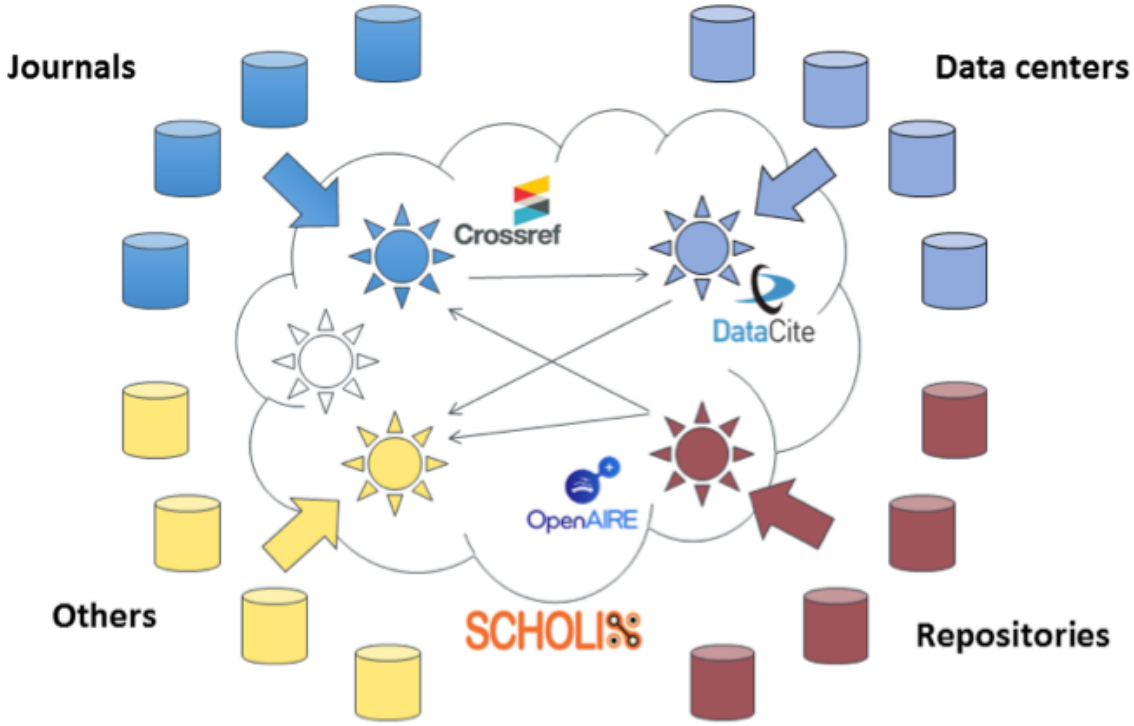

Figure 2: Schematic overview of the proposed multi-hub interoperability framework including the Crossref, DataCite and OpenAIRE hubs as collectors and aggregators of information about links between literature and data scholarly artefacts from respective communities.

The multi-hub network is interoperable among the hubs and thus facilitates interoperability among data centers, journal publishers, repositories, etc. via the hubs. Interoperability among the hubs is achieved by hub-commitment to the common information model and open exchange methods described earlier. A high level of conformity and uniformity is assumed in the information exchanged between the hubs.

However, the hubs do not affect existing community-specific practices. There is no requirement for "many to many" interoperability between data centres, journals and repositories. In contrast to the uniformity among the hubs, heterogeneity is a natural part of the outer layer and is addressed by the hubs with their communities.

The resulting interoperable network of hubs and their aggregated information is expected to enable new third-party services, such as for research impact measures, integrity measures, discovery, and research information. Such services build on top of the proposed multi-hub infrastructure and may be commercial or noncommercial.

2.2 Conceptual Model

A shared conceptual model is a fundamental enabler of interoperability. The Scholix conceptual model is a way for scholarly infrastructure providers to begin to talk about data-literature links. Indeed, when parties involved in scholarly link exchange agree on concepts and language, a crucial and foundational element of interoperability is established.

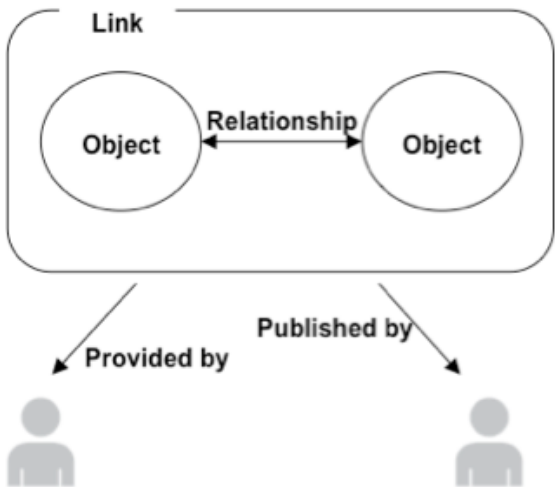

Figure 3: Scholix common conceptual model for information about links between literature and data scholarly objects.

At the core of the conceptual model is the *link* between two *objects*. The framework considers two types of research objects, namely literature and data. Theoretically the model extends to any type of research objects: software, algorithms, models, protocols, tweets, comments, and so on. In practice the Scholix initiative chooses to focus on data and literature in the first instance since it aspires to a comprehensive view of these, a non-trivial aspiration given the hundreds of millions of publications and datasets. Therefore other research objects are not for the time being the focus of the framework or its implementation projects.

Links are about single relationship between two objects. Complex relationships between many objects, or multiple relationships between two objects, are portrayed with multiple links.

Scholix is not primarily focused on properties of research objects themselves — neither data nor literature, e.g. authors, title, description — but rather on the properties of the assertion that two research objects are linked, e.g. relationship type.

In addition to the link, the conceptual model includes information about the parties that assert the relationship between two research objects and information about the parties that collect and/or make links available: data centers are an example of "link publishers", i.e. parties that assert relationships between research objects, while OpenAIRE is an example of "link providers", i.e. actors that collect, aggregate, and make links available.

2.3 Information Model

The information model of the Scholix framework specifies what information is needed about the elements of the Scholix conceptual model, appropriate for the purpose of exchanging data-literature links. Cardinality is also specified, identifying mandatory and optional information and how many times it may appear.

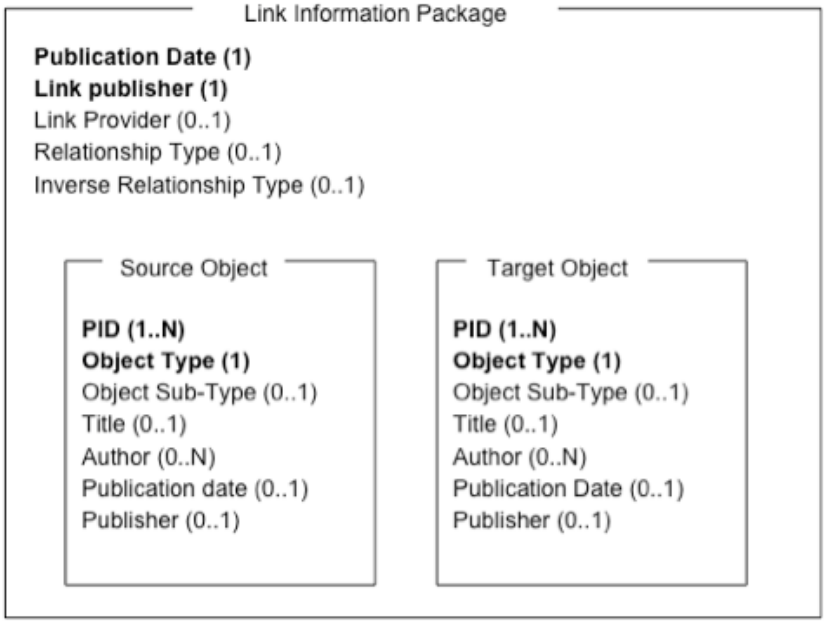

Figure 4: Information model entities and attributes.

The information model specifies information structure for the link and the two research objects involved in the relationship.

Link. The mandatory fields of the link are *publishing date* and the *link publisher*. The *link publisher* is the party that states and publishes the assertion that two research objects are related (such as a journal publisher). As such, it represents important provenance information that enables consumers to make value or quality decisions based on their own criteria. Other information about the link is regarded as optional (at least in some circumstances). The *link provider* is the party from which the link was last collected, which may not necessarily coincide with the *publisher*. For example an aggregator of metadata, e.g., Crossref or OpenAIRE, may be a provider that redistributes links after having collected them from sources, i.e. the publishers. Typically, the link provider should be included if it is different from the publisher, to ensure equal visibility of all actors contributing to the dissemination of the information package. This information is used by hubs to deduplicate and trace provenance of aggregation procedures. Finally, the *Relationship Type* is strongly recommended, as it is indeed important for the interpretation of the link (e.g. `isCitedBy`, `isDerivedFrom`), but not made mandatory as in some contexts the relationship may not have been specified. Ultimately, Scholix regards the availability of a link between two research objects important and relevant for the community. It is expected that new Scholix compliant contributors will consider the relationship types as mandatory, and have a very broad relationship, such as `References`, to model the absence of any explicit semantics.

Objects. Research objects connected by a scholarly link are named `source object` and `target object` to enable determining the direction of the link. The mandatory fields of the two research objects related by a link are PID (persistent identifier) and *object type* (currently `dataset` and `publication`, but to be possibly extended to different classes of research

objects in the future). At least one PID for both research objects is required (although not specified in the picture, PIDs come with a PID type (e.g. DOI, PDB, URL), which instructs consumers on how to interpret and resolve the identifier), and the type of the source and target object must be specified. Other information is considered recommended but not mandatory, as it can in principle be recovered by means of resolution of the object PIDs. Provision of fields such as `Title`, `Author`, `Date`, `Publisher`, and `Relationship SubType` aids information consumers, and avoids the need for resolution, but on the other hand is subject to known issues of information integrity. Moreover some PIDs are more readily resolvable than others. For these reasons, such information is considered optional for the purposes of the Scholix information model.

2.4 Information Standards

The Link Information Package defined above provides the high level elements which would be serialised into a structure, schema or ontology of some kind. However, interoperability is not complete without application of encoding and semantic standards to the values that can be entered into those elements. At this point in the development of Scholix there is a commitment to apply standards, but firm decisions have not yet been taken. Table 1 below shows some of the potentially applicable standards for the elements in the Scholix information model.

Table 1: Information Standards

Information element	**Examples of potentially applicable standards**
(Link) Publication Date (Object) Publication Date	ISO8601, XML Schema dateTime W3CDTF, EDTF — to be explored
Relation Type	DataCite Metadata Schema: Relationship
Object Type	Resource Types from DataCite, COAR, CASRAI
Link Publisher Link Provider (Object) Publisher	ISNI, Ringgold, Digital Science GRID, PROV — to be explored further
Author	ORCID

A successor working group is planned with RDA and ICSU-WDS, tentatively called the Scholarly Link Exchange Working Group. That working group is tasked with making decisions on these encoding and semantic standards, constraining allowed value registers for each of the information elements.

2.5 Formats and Protocols

The Scholix interoperability framework does not mandate how to format and exchange a Link Information Package. Potentially, such information can be formatted and exchanged using a range of models and protocols — e.g. JSON, XML, or RDF (Resource Description Framework) formats and RESTful, OAI-PMH or SPARQL (SPARQL Protocol and RDF Query Language) protocols.

It is expected that all Scholix hubs will at least support a JSON/RESTful combination. Some will also allow the use of community-supported protocols such as OAI-PMH. The successor working group has committed to providing guidelines for serialising the information model via an XML and JSON schema and to documenting any corresponding exchange protocols.

The successor working group will also investigate additional formatting and exchange approaches that may only be supported in certain communities or certain services.

For example the Distributed Scholarly Compound Object (DISCO) [6] together with the RMap open APIs, show promise as a manageable yet extensible packaging format and exchange protocol which would allow flexible integration with other areas of the research graph (people, grants, software, etc).

The Scholix model would lend itself to semantic web approaches and the initiative is open to working with potential new hubs that would support and document the Scholix framework using e.g., the Web Ontology Language (OWL). The PROV model may also have application to inform the provenance information of the link [17].

The Scholix initiative will also investigate the use of Linkback approaches such as Webmention [18] to allow the information in a Scholix Link Information Package to be sent as a notification using Web-based approaches.

It is not expected that individual publishers, data centres or repositories use a "Scholix" schema for this information. Rather, the framework respects existing local community standards and supports the submission of such information to hubs using standards. It is expected that hubs translate from existing local community standards using the conceptual and information models, and expose the information to other hubs using standard Scholix approaches.

3 Scholix in Practice

The Scholix framework rests on a multi-hub network that collects and aggregates information about data-literature (and data-data) links. Each hub focuses on a particular community. Communities are for, e.g., the journal publishers, the data centers, etc. The hubs aggregate information from communities. Of particular interest are "natural" hubs, i.e. existing infrastructures that focus on community information aggregation.

The Scholix framework has been complemented by the development of a pathfinder aggregation service, i.e. the Data-Literature Interlinking Service (DLI Service) [7], as well as enabling infrastructure from global information services such as DataCite, Crossref, and OpenAIRE. Organizations are thus already starting to develop services that follow the Scholix framework. The following section describes these existing hubs individually before we introduce the vision of a connected and interoperable multi-hub network.

3.1 Existing Hubs

Data-Literature Interlinking Service

The Data Literature Interlinking service (DLI Service) populates and provides access to a graph of dataset-literature and dataset-dataset links aggregated from a variety of data sources. The links are collected from data sources managed by publishers (e.g. Elsevier, Springer Nature, IEEE), data centers (e.g. PANGAEA, CCDC, ICPSR), or other organizations providing services to manage links between datasets and publications (e.g. DataCite, OpenAIRE, Thomson Reuters), to be then harmonized and resolved (when only the PID is provided for the research objects involved). The DLI Service also offers an API to access the graph of links, specifically: (i) full-text search by field or free keywords returning publications and datasets matching the query (with a limit of 10,000 results); (ii) bulk access via OAI-PMH protocol, and (iii) PID resolution capabilities, returning the links whose source object has a given PID (pairs PID schema and PID). Such links will soon be exposed according to the Scholix Information Model and exchange formats. The DLI Service is accessible in BETA, features around 8M links, is scheduled to become a production service in early 2017.

Crossref and DataCite Event Data

Crossref and DataCite are collaboratively working on providing literature-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 links, and makes the links available to other Scholix Hub partners. Publishers and data centers submit literature-data links via the DOI metadata they deposit with Crossref and DataCite, respectively. Crossref and DataCite Event Data will become production services in 2017.

Research Data Switchboard

One of the third party tools that can implement and support the Scholix framework is the Research Data Switchboard (RD-Switchboard) and the research graph database created by the participants in the RDA Data Description working group.

RD-Switchboard is an open and collaborative software solution that addresses the problem of cross-platform discovery of research data. This 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 this system 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 main capability of RD-Switchboard is connecting datasets to grants, publications and researchers across multiple registries, and the outcome of the process is captured as a database using the Research Graph schema. The integration of Scholix framework into this system enables RD-Switchboard to ingest trusted information about the relations between datasets and publications, and enrich these relations using complementary information captured in the Research Graph database. The outcome can lead to a new discovery of links between datasets, publications and other scholarly works. At the time of writing this article the participants in the RD-Switchboard project are working on implementing a distributed research graph across institutional repositories, and the Scholix framework enables a well coordinated information exchange with data hubs and the sites who implemented and use the RD-Switchboard graph database such as University of Sydney and NCI (National Computing Infrastructure) in Australia.

3.2 Participation Options

The RDA/WDS working group has so far established a consensus and the first stages of a conceptual interoperability framework. There are a number of ways of implementing and furthering the Scholix framework.

Feed data-literature info to an existing Scholix hub. Journal publishers, repositories, and data centres can provide data-literature link information to existing Scholix hubs: DataCite, Crossref, OpenAIRE. This means simply enriching your existing metadata feed to these services with data-literature links *using the pre-existing community standard* (i.e. using the appropriate fields in the Crossref deposit schema, the DataCite schema or Dublin Core).

Become a hub. If you have data-literature link information that doesn't fit naturally into a feed to an existing hub then simply expose a new feed using the Scholix Guidelines. Then register as an implementor to be included in the DLI aggregation and other hub exchanges.

Develop a third-party service. Use data-literature link information from Scholix hubs in your own service. For example query the Scholix hubs to see who is referring to research objects in your service.

Help further develop the guidelines. The Scholix framework requires further extension, documentation, standardisation, serialisation schema and investigation of new applications and approaches. Support materials and community specific guidelines are also needed. Help expand and refine the Scholix guidelines and suggest new ways of application. In the first instance the framework has focused on the information flows needed to create a critical mass of data-literature link information. The second stage will include increased focus on query and access methods that services might use to get data-literature link information from hubs.

Scholix is an interoperability framework with an explicit technical focus. The overall goal of achieving better linking between data and literature must be complemented with advances in policy, community awareness, practice and capacity and capability which are being championed by a number of other initiatives.

4. Conclusion

Scholix is a proposed conceptual framework to drive interoperability between providers of links between research data and the literature. At its heart is an evolving lightweight set of guidelines that express and give shape to a joint vision in which natural hubs play a central role in aggregating and normalizing links contributed by various stakeholder groups such as data centers and journal publishers. Scholix is not a product, service, or infrastructure in itself. Rather, it aims at improving the incentive for data and literature publisher communities to capture and expose link information by making such information more easily globally visible and trackable, thereby increasing its value and utility for the benefit of all stakeholders.

It is recognized that the real value for Scholix is in the willingness of a broad group of publishers, data centers, and global service providers to work toward a more effective overall system. It also acknowledges that the envisioned future with robust, production-strength workflows for the aggregation of links by the various hubs from their constituencies will not materialize overnight. We argue that the proposed framework allows for the transition from the current state to the envisioned future state. A key element for managing this transition is the inclusion of the existing DLI Service as one of the proposed hubs to allow for organizations to contribute links in a flexible way while automated, standardized workflows are being implemented.

Currently, the framework is a set of informal guidelines and high level models. Nevertheless, it represents a significant first step toward a shared vision, standardisation, and the realisation of alluring benefits of a global information commons around data and literature. A further ICSU-WDS / RDA Working Group is being organised to elaborate the Scholix framework, coordinate the development of hubs, and support community adoption. The working group will also work actively with: service providers for benefit realisation; hubs and their communities for buy-in; and with international advocacy and peak bodies for culture change.

Acknowledgements

The authors would like to thank the Publishing Data Services-WG members and representatives from Crossref, DataCite, The National Data Service, ORCID, The Research Data Alliance, ICSU World Data System, 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), OpenAIRE2020 (H2020-EINFRA-2014-1, grant agreement: 643410), and THOR (H2020-EINFRA-2014-2, grant agreement: 654039). Some of this work is resourced by ANDS which is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy program.

References

[1] Manghi, P., Bolikowski, L., Manola, N., Schirrwagen, J., and Smith, T. (2012). OpenAIREplus: the european scholarly communication data infrastructure. D-Lib Magazine, 18(9). https://doi.org/10.1045/september2012-manghi

[2] Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela L., Castelli D, and 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. https://doi.org/10.1108/PROG-08-2013-0045

[3] Smit, E. (2011). Abelard and Héloise: Why Data and Publications Belong Together. D-Lib Magazine, Volume 17. https://doi.org/10.1045/january2011-smit

[4] Aalbersberg, I.J., Dunham, J., and Koers, H. (2013). Connecting Scientific Articles with Research Data: New Directions in Online Scholarly Publishing. Data Science Journal. 12, pp.WDS235-WDS242. https://doi.org/10.2481/dsj.wds-043

[5] Callaghan, S., Tedds, J., Lawrence, R., Murphy, F., Roberts, T., and Wilcox, W. (2014). Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples. International Journal of Digital Curation. https://doi.org/10.2218/ijdc.v9i1.310

[6] Hanson, K. L., DiLauro, T., and Donoghue, M. (2015). 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. https://doi.org/10.1145/2756406.2756952

[7] Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., and Schindler, U. (2015). On Bridging Data Centers and Publishers: The Data-Literature Interlinking Service. In Metadata and Semantics Research (pp. 324-335). Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_28

[8] Borgman, C.L. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press.

[9] Data Citation Synthesis Group (2014). Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11.

[10] Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R.R., Duerr, R., Haak, L.L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J.E., Lin, J., Nielsen, L.H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M., Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1:e1. https://doi.org/10.7717/peerj-cs.1

[11] Ball, A., Duke, M. (2015). How to Track the Impact of Research Data with Metrics. DCC How-to Guides. Edinburgh: Digital Curation Centre.

[12] Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., Wright, D. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. The International Journal of Digital Curation, 7(1):107-113. https://doi.org/10.2218/ijdc.v7i1.218

[13] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604):452-454. https://doi.org/10.1038/533452a

[14] Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. Nature. 515(7525): 9. https://doi.org/10.1038/515009a

[15] Cook, R.B., Vannan, S.K.S., McMurry, B.F., Wright, D.M., Wei, Y., Boyer, A.G., Kidder, J.H. (2016). Implementation of data citations and persistent identifiers at the ORNL DAAC. Ecological Informatics, 33:10-16. https://doi.org/10.1016/j.ecoinf.2016.03.003

[16] Lagerstrom, J. (2010). Measuring the Impact of the Hubble Space Telescope: open data as a catalyst for science. World Library and Information Congress: 76th IFLA.

[17] Gil, Y., Miles, S. (2013). PROV Model Primer. W3C Working Group Note, W3C.

[18] Parecki, A. (2016). Webmention. W3C First Public Working Draft, W3C.

About the Authors

Adrian Burton is the Director of Services at the Australian National Data Service (ANDS). Adrian has provided strategic input into several national infrastructure initiatives, including Towards an Australian Research Data Commons, The National eResearch Architecture Taskforce, and the Australian Research Data Infrastructure Committee. Adrian is active in building national policy frameworks to unlock the value in the research data outputs of publicly funded research. Before being involved in research infrastructure, Dr Burton taught South Asian Linguistics and conducted research at the Australian National University and was responsible for liaison between academic staff and central information and technology services.

Hylke Koers is Director of Research Communities at Elsevier, based in Amsterdam. A physicist by training, Dr. Koers is passionate about improving the way researchers communicate and he enjoys the challenges of connecting innovative web technology with the social dynamics of scholarly publication. Before joining Elsevier, Hylke was Business Development Manager for MathJax, an open-source engine for rendering mathematics on the web. At Elsevier he was responsible for the Content Innovation program, which advances the format of the scholarly article through interactive data viewers and bi-directional linking with primary data at external data repositories. Together with Adrian Burton, Hylke co-chaired the ICSU-WDS / RDA Working Group on “Publishing Data Services” which delivered the Scholix framework.

Paolo Manghi is a (PhD) Researcher in computer science at Istituto di Scienza e Tecnologie dell'Informazione (ISTI) of Consiglio Nazionale delle Ricerche (CNR), in Pisa, Italy. He is acting as technical manager and researcher for the EU-H2020 infrastructure projects OpenAIRE2020, SoBigData.eu, PARTHENOS, EOSC, and RDA Europe, and he is the coordinator of the OpenAIRE-Connect project. He is active member of a number of Data Citation and Data Publishing Working groups of the Research Data Alliance; and invited member of the advisory boards of the Research Object initiative. His research areas of interest are today data e-infrastructures for science and scholarly communication infrastructures, with a focus on technologies supporting open science publishing, i.e. computational reproducibility and transparent evaluation of science.

Markus Stocker is a postdoctoral research associate with PANGAEA, the Data Publisher for Earth & Environmental Science, at the MARUM Center for Marine Environmental Sciences, University of Bremen, Germany. He holds a PhD in Environmental Sciences (Informatics) from the University of Eastern Finland and a MSc in Informatics from the University of Zürich, Switzerland. He is interested in environmental knowledge infrastructures and has several years of professional experience in software development with interests in semantic web technologies, environmental sensor networks, data management and data mining as well as formal representation of mined information and knowledge.

Martin Fenner is the DataCite Technical Director and manages the technical architecture for Datacite as well as DataCite's technical contributions for the EU-funded THOR project. From 2012 to 2015 he was technical lead for the PLOS Article-Level Metrics project. Martin has a medical degree from the Free University of Berlin and is a Board-certified medical oncologist..

Amir Aryani is working in the capacity of a project manager for Australian National Data Service (ANDS), and he is the co-chair of the Data Description Registry Interoperability WG in Research Data Alliance. Dr. Aryani has a PhD in computer science. His research is focused on the interoperability between research information systems, and he is leading the Research Graph project to build a large-scale distributed graph that enables connecting heterogeneous data infrastructures.

Sandro La Bruzzo is a research associate at Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy. He received his MSc in Information Technologies in the year 2010 at the University of Pisa, Italy. Today he is a member of the InfraScience research group, part of the Multimedia Networked Information System Laboratory (NeMIS). His current research interests are in the areas of Service-Oriented Infrastructures for Scholarly Communication, protocols for metadata exchange, interlinking of datasets and literature. He is currently working on the development and operation of the Aggregative Data infrastructures for the European Commission projects OpenAIRE2020, EFG1914, and EAGLE; he is also on charge of the developments and operation of the Data-Literature Interlinking Service.

Michael Diepenbroek has been Managing Director of PANGAEA, at MARUM, the University of Bremen since 1998. He holds a PhD in Geology from Free University of Berlin (1992). In early 1990, his work concentrated on relational and object oriented DBMSs. This work eventually led to the conception and implementation of PANGAEA, and its endorsement as ICSU World Data Center. Since 1998, he has been working on concepts for publishing data, thus contributing to the development of DataCite and services linking data and literature (Scholix). Between 2009 and 2015, Michael was the vice chair of the ICSU World Data System Scientific Committee.

Uwe Schindler is a software architect in the PANGAEA group. He studied Physics at the Friedrich-Alexander-Universität Erlangen-Nürnberg. He is specialist on full text search engines and chair of the project management committee of Apache Lucene. His responsibilities at PANGAEA are middleware development (all software that runs in the backend between user interface and data archive). He maintains the PANGAEA search engine and all its webservices, metadata management, Open Archives Initiative protocols, and data portals.
