ICSU-WDS & RDA
Publishing Data Services WG

Interoperability Framework Recommendations

Prepared by Adrian Burton & Hylke Koers, March 2016

These recommendations are a sub-set of the recommendations and reporting materials of the RDA/WDS Publishing Data Services Working Group (PDS WG). This document focuses on the first steps to a broader interoperability framework and enabling infrastructure proposed to underpin a global data-literature information eco-system. Other documents provide an overview of the activities, other outputs and services emerging from the PDS WG.

If you would like further information, please contact the co-chairs of this working group: adrian.burton {at} ands.org.au or hylke.koers {at} elsevier.com or visit https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html.

Summary

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

Frame

The working group re-affirms its commitment to the principles outlined at the beginning of the working group and proposes these recommendations consistent with the spirit of those principles enhanced with the experience of the 18 months work of the group.

As future perspective, the working group envisages an *enabling infrastructure* that allows links between research products (notably literature publications and data sets) to be shared and aggregated. This infrastructure should be:

- Cross-disciplinary and global (built for, and aspiring to, comprehensiveness)
- Transparent with provenance allowing users to make trust and quality decisions
- Open and non-discriminatory in terms of content coverage and user access (this also means ranging from formal to informal, and from structured to non-structured content)
- Standards-based (content standards and exchange protocols)
- Participatory and adopted, including community buy-in
- Sustainable
- An enabling infrastructure, on top of which services can be built (as opposed to a monolithic “one-stop-shop” solution).

Scope

These recommendations are a step in a journey towards a broader open information ecosystem which facilitates the generating, sharing, aggregating and exploitation of the links between research data and the literature. By creating an efficient information environment for collecting this information, the working group anticipates a potential groundshift in practices and attitudes toward data citation. Therefore many of the recommendations are phrased dually:

1. taking into account current heterogenous practice and information environment
2. anticipating greater common practice and simpler standard information workflows

The recommendations do *not* include

- value-adding services, but rather focuses on **enabling infrastructure** that supports higher-level services
- specific technical rules for publishers or data centres, but rather **high-level information concepts for localisation in various communities**.
- discipline specific considerations
- links to people, grants, software etc but rather **focuses on literature-data and data-data links** (but notes that the infrastructure should allow for interoperability with such information)
- value judgements on publications or data - but rather the Infrastructure will support different levels of quality, capture provenance and allow filtering on assertion source and provider
- community building or attitude change, while nevertheless acknowledging that the community is crucial in implementation.

Analysis of Information Flow

The recommendations are based on the typical information flows that result in a link between publications and datasets:

Where does information about links come from?

Some typical sources of data-literature link information include:

- Data centers
- Publishers
- Repositories (also grey literature)

Current practices are in flux. Currently it is the case that linking information mostly goes “from data center to publisher”, but it is anticipated the balance may shift towards “from publisher to data center” as data citation becomes more prevalent and as the technical recommendations from this working group become broadly adopted.

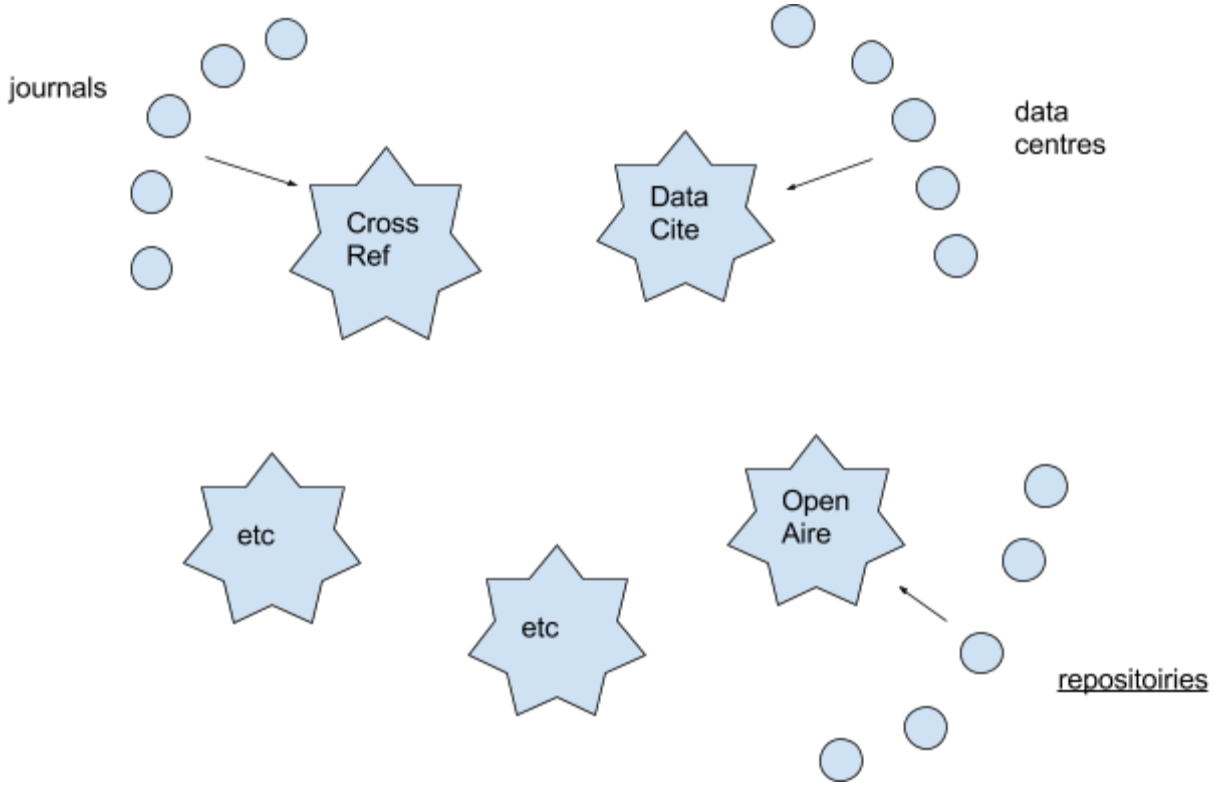
Are there aggregation points for this information?

There would seem to be natural aggregation points for the information about links between literature and datasets based on the established and expected behaviours of different communities. For example:

- CrossRef is a natural aggregation hub for information from many journal publishers;
- DataCite is a natural aggregation hub for many data centers;
- OpenAIRE is a natural aggregation hub for institutional (data, preprint, or publication) repositories.

In addition, there are natural aggregation points for specific scientific domains that could be leveraged - for example:

- PubMed and Europe PubMed Central (EBI) for life sciences;
- ADS for astronomy;
- INSPIRE for high-energy physics;
- PermaCC for law.



These aggregators form natural hubs for information about the links between publications and repositories.

At what point in the publication lifecycle is link information recorded?

Information about the links between literature and data can be recorded at first publication of the object or at a later stage in the scholarly communications lifecycle[1].

"T=1": At the publication of a research object (which could be a data set, or a literature publication), its links to datasets or literature can be recorded as part of the publication process - when a journal article is published with references to data or when datasets are deposited in a datacentre with links to literature in the accompanying metadata.

[1] One could also imagine this happening in an e-lab notebook, so, prior to "formal publication" (T=0?)

"T=2": At various times after publication of a research object, information about the links to data or literature are sometimes added post-facto through a number of processes:

- Journal publishers processing backfiles to identify informal references
- Curators or librarians for example scouring literature for references to data from their data centre
- Grassroots crowd-sourcing
- Systems establishing links though inference

As mentioned above, the status quo sees most links being *formally* recorded at *T=2* or only by some data repositories at *T=1*. Of course many researchers refer to a data set one way or another in a manuscript, so links are being recorded but not in a structured way. Moreover any links that become subsequently apparent are often updated at T=2.

The proposed technical framework aims to make *T=1* recording of this link information the default practice (by making it easier, by making it standard, and by making it worthwhile). That means as any new object is published its links to data or literature are also declared in a standardised way. Such change to enterprise systems and community practice however will be measured in years. In the meantime a hybrid solution (including inference, manual retrospective linking, data cleaning) will be needed to carry this program forward within the existing information landscape.

How can literature-data references accrue over time?

As each new research object[2] (data or literature) is published, new links are recorded with objects previously published by different publishers. Such a system (distributed over both time and organisations) is greatly optimised by unique global identification of all research objects so that references recorded over many years by different players can be consolidated. Eg if a standard dataset were to be referenced by many journal articles in many different journals over many years, then to the extent that the reference is common (eg a DOI or a GenBank ID) the references can be aggregated by the proposed hub-interoperability model.

The proposed framework also provides incentives for the use of unique identifiers and standardised referencing. By creating a global information ecosystem for data-literature links, data centers and literature publishers have much to gain from being part of the greater network whose "entry fee" is standardised referencing and unique identification.

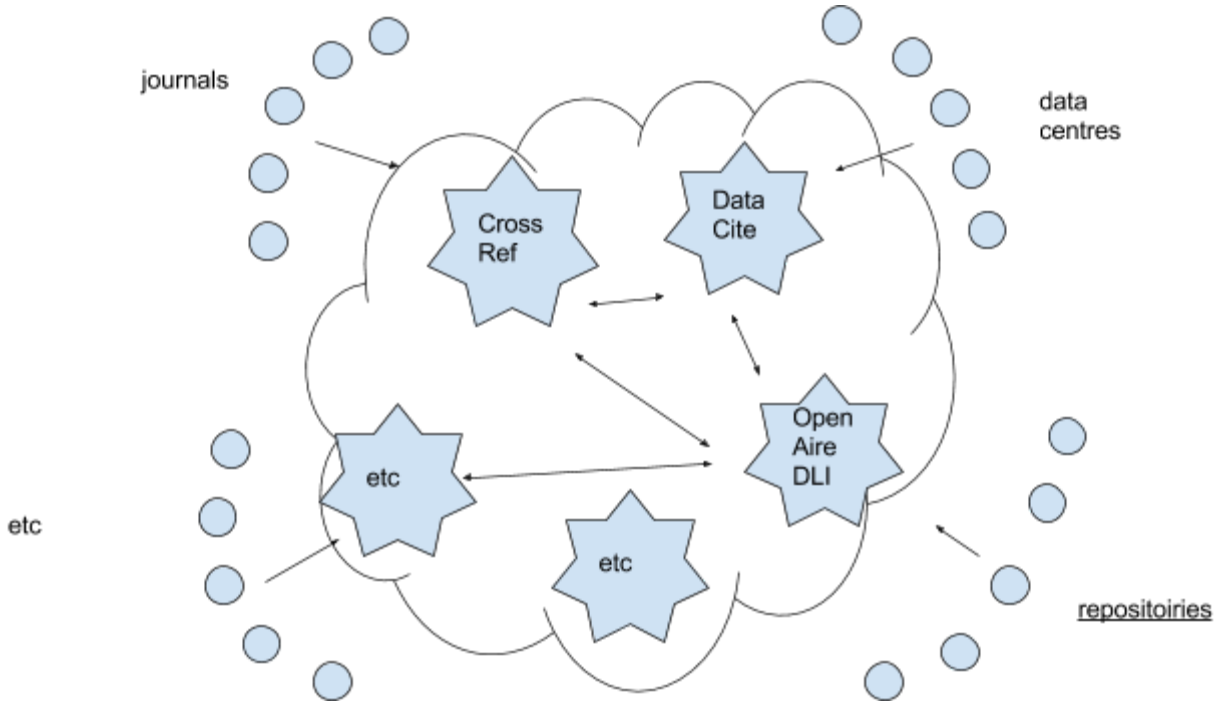
[2] Here the phrase "research object" is used in simple plain English terms: a published "object" from or for "research". In this context we expect them to be the output of research and often used as the input to further research. "Data" and "literature" are two types of research object within this model. Other types of research object have been mentioned (software, tweets, etc) but these are not the focus of the present interoperability framework. The phrase "research object" here is not used in the formal defined sense of Bechhofer et al 10.1038/npre.2010.4626.1.

What does this mean for the proposed framework?

Given these properties of the information flow around data-literature links, the proposed framework assumes:

- a number of natural hubs that aggregate data-literature links
- interoperability between those hubs and NOT between all the publishers of research objects
- the need for inference and retrospective linking (at least for several years to come)
- that some hubs will actively aggregate information from other hubs
- that other services will leverage this enabling infrastructure
- the need for unique persistent identification and standard referencing of research objects

Multi hub model



The proposed model assumes that natural hubs aggregate information from communities. A global information ecosystem emerges as a result of a high level of “few to few” interoperability between the hubs based on common information models and negotiated exchange protocols. Note that “many to many” interoperability is not required by data centres, journals and repositories; heterogeneity is a natural part of the outer layer and is addressed by the hubs with their communities. A high level of conformity/uniformity/interoperability is assumed in the information exchanged between the hubs.

It is expected that some hubs will take advantage of that inter-hub interoperability to aggregate this information globally, eg the current DLI service (operating on OpenAIRE infrastructure) which provides services over that aggregation; some hubs may query other hubs only for information on a specific community or discipline, eg astronomy or life sciences.

This interoperability between participating hubs and the resultant aggregate pool of information about literature-data links can potentially support a number of third party services (which could be both commercial and noncommercial) such as research impact measures, integrity measures, discovery services, and research information services.

Interoperability framework

The proposed interoperability framework focuses on exchanging high volumes of information between hubs (natural aggregators of data-literature link information). The framework represents an informal arrangement to optimise this exchange, taking into account existing practice and realities. It is not a fully specified, normative, open standard but rather an attempt to increase interoperability between a set of willing hubs.

Interoperability in this sense is a spectrum which may span from on the one hand a set of shared concepts to on the other a fully automated exchange of mutually intelligible information. This framework may be used by different players in the system to achieve greater or lesser degrees of interoperability. A shared set of concepts may allow parties to start a dialog and start to work together in a shared vision. A shared schema of elements with commonly defined values and pre-negotiated API may allow others to exchange high volumes of instantly actionable and intelligible information.

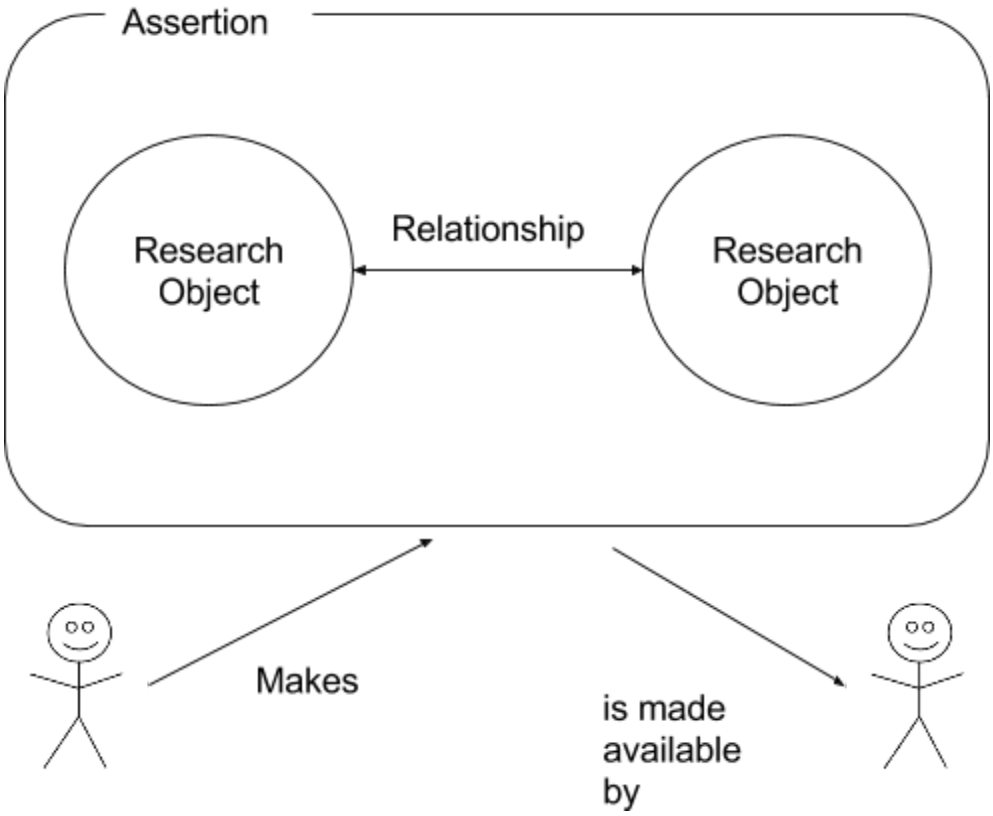
The proposed framework focuses on the exchange of information about the links between data and literature. It is not primarily focused on properties of either data and literature itself (eg the author, description,etc) but rather on the properties of the assertion that the two are linked (eg relationship type etc).

The framework includes guidelines on:

1. Shared conceptual model
2. Information model
3. Encoding options
4. Exchange options

Shared conceptual model

In exchanging information about the links between literature and data, the following concepts are foundational:



Here:

- Research objects include data and literature (and other types out of scope for this particular framework)
- Research objects have relationships with other research objects
- Parties in the research system assert that two research objects are related
- Parties in the research system collect and/or make those assertions available

Information Model

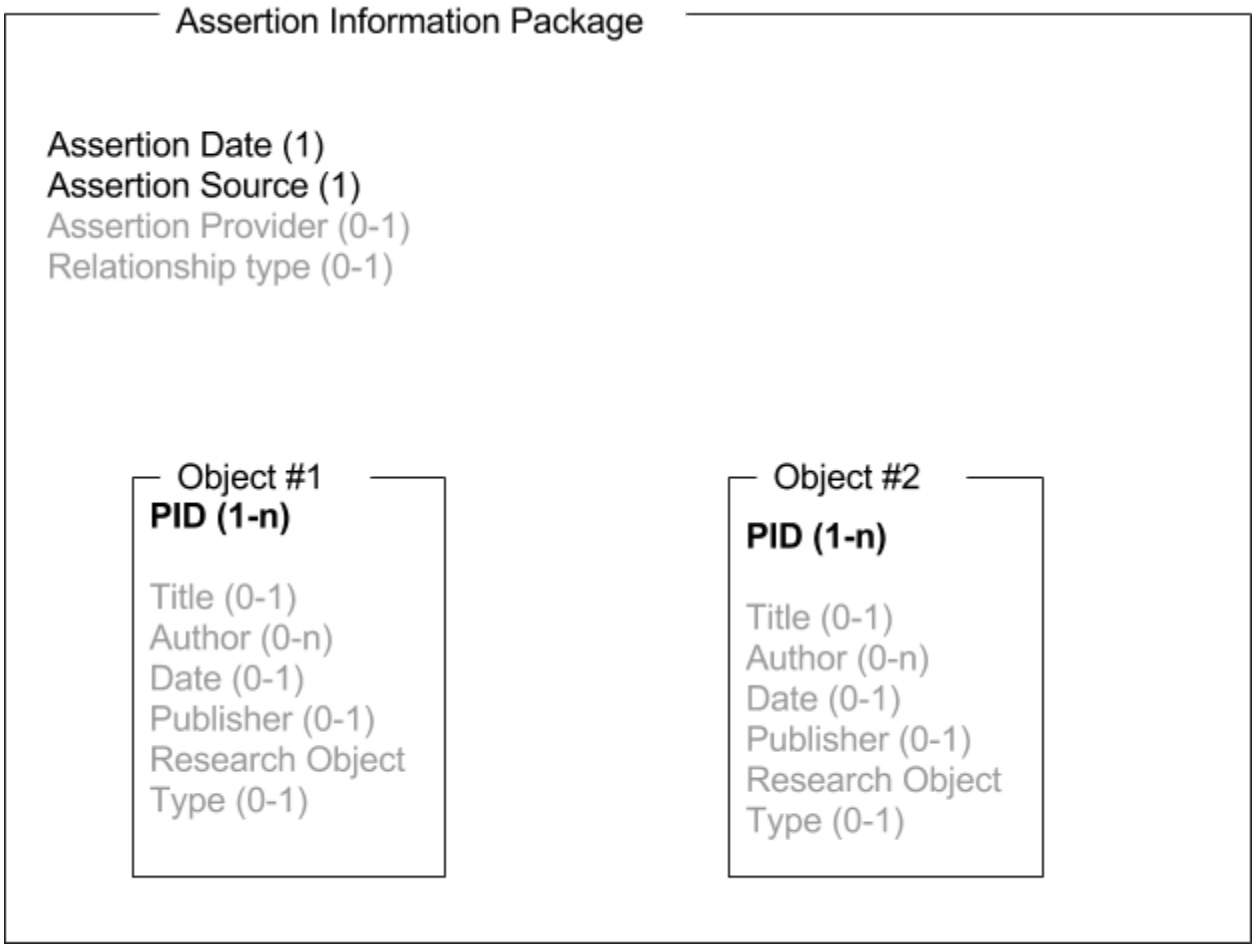


Figure: High level information modelling for an assertion information package.

The figure above represents the information requirements for exchanging data-literature link information between information systems within this framework.
Notes:

- Objects are numbered so that the direction of any relationship can be determined. Object #1 is the agent (subject in English) of any relationship statement. Other ways of determining the direction of relationships may be explored.
- Assertions are about single relationship between two objects. Complex inter-relationships between many objects or multiple relationships between the same objects are portrayed at least conceptually with multiple assertions.
- Assuming that PIDs can be resolved with bibliographic information, no other information is required about the objects. If the PID is not resolvable or globally unique, then extra information is highly desirable.

- The existence of two objects in an assertion is enough to assert “a relationship”. Relationship type is highly desirable.[3]
- Assertion Source is the party making the assertion that two research objects are related (eg a journal publisher). This provenance information is meant to provide end users the ability to make value or quality decisions based on their own criteria.
- Assertion Provider is a third party who aggregates such information (eg CrossRef) from the Assertion Source and makes it available. Only encoded if the assertor is different from the provider. This information is used by hubs to deduplicate and trace provenance of aggregation procedures.

Information Standards

A further level of interoperability is obtained by agreeing on the structure and meaning of the values for each of the things mentioned in the information model. This framework encourages best practice in the use of identifiers and controlled vocabularies and recommends further work be done in identifying appropriate standards applicable for each of the elements.

Practically all the elements in the information model would benefit from format or semantic standardisation and some potentially applicable standards are listed in the table below.

Information Element	Examples of potentially applicable standards
Assertion Date Research Object Date	ISO8601, XML Schema "dateTime" format, W3CDTF, EDTF - to be explored
Relationship Type	DataCite Metadata Kernel: Relationship
Research Object Type	Citation Styles - Types? - to be explored further.
Assertion Source Assertion Provider Research Object Publisher	ISNI, Ringgold, Digital Science GRID, PROV? - to be explored further.
Author	ORCID

[3] Given the centrality of relationship in this model, relationship type would/should be mandatory but it was thought that this may exclude legacy information where type has not been specified. It is expected that all new participants in the framework would consider relationship types as mandatory.

Given that the information model regards three elements as essential (Assertion Date, Assertion Source, Research Object PID) it is recommended that immediate work be done to provide guidelines for recommended practice within this framework for these three elements.[4]

The global aspiration of the framework makes this work both pressing and complex. Agreeing on one (rather than zero or many) relevant standards across such a broad set of communities is challenging, and yet aggregating data-literature link information across such a broad set of communities will be greatly enabled by standardised information practice. Some of the work here may be identifying standards in different communities and enabling mappings.

Encoding Options

At this point in time the interoperability framework does not specify encoding of the information model in particular schemas or exchange formats. The same information could be formatted using json, xml, rdf triples, etc. At this stage in the formation of the community it is expected that hubs will nominate some of these commonly used standards formats as interchange formats that they support.

It is not expected that individual publishers, data centres or repositories are required to use a specific schema for this information. The information should be a small subset integrated into existing exchange schemas already in use in various communities (eg the DataCite Metadata Kernel). It is expected that hubs translate from existing local community standards using the common information model and expose the information to other hubs using standard formats.

The working group is keen to investigate further the use of DISCO's to encode the information in an assertion information package, as well as the concept of research objects bundles.

Together with well-defined open APIs, these DISCOs 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).

Exchange Options

Similarly, at this point in its development, this interoperability framework does not specify specific exchange protocols. It is expected that hubs in this distributed information system will support well documented open RESTful APIs as well as handful of community supported protocols such as OAI-PMH (alternative options to be explored further).

If the model progresses and scales to a truly global ecosystem it is anticipated that a more devolved distributed system will emerge. The working group is keen to investigate further the use of *LinkBack* or *webmention* methods to allow notifications to be sent using Web-based approaches.

[4] For example it is agreed that Research Object PID should be expressed as a URI

First steps toward interoperability

As a lightweight starting point all hubs will support the common information model formatted in JSON delivered through open documented APIs. With the same approach and effort applied to a typical data mashup, the hubs should be able to begin interoperation immediately. These participating hubs will be registered in the first instance on the initiative's website (with information about their interfaces and supported protocols). That register of participating hubs will in the future also be available in machine readable format with notifications.

From here to there: phasing and next steps

Phasing in of hubs

Current practices to collect and share links between data and the literature are diverse and often ad-hoc, whereas in the future we envision there will be robust, production-strength workflows to aggregate links by the various hubs from their constituencies. Moving from one situation to the other will not happen overnight, and necessitates a period of transition during which the contributing nodes can set up connections to one of the hubs. In the meantime, existing infrastructure and services for link consumers should remain available as much as possible, ideally even extending coverage and service over time.

The multi-hub model, as proposed in this recommendation, has an inherent flexibility that can be used to manage the transition from the current state to the envisioned future state. The existing DLI system, one of the proposed hubs, can play an important role in managing this transition as it allows for an organization to contribute links in a flexible way while automated, standardized workflows are still being set up.

The DLI may already be seen as a first step to bridge the gap between current practices and this future state. It contains over 1 million unique links contributed by a variety of sources across the different stakeholder groups and exposes them using a portal, OAI-PMH and an API service. Aggregation of links by the DLI has been very flexible, working closely with the various contributors to find solutions that fit with their existing systems and workflows. These solutions range from connecting the DLI to third-party API's, to ingesting links described in a CSV document delivered over email. On the ingestion side, we proposed to keep these flexible input options in place until standard, automated workflows between hubs and nodes are fully established. This way, any organization can still contribute links in a way that best meets their current capabilities as they develop the capabilities that will be needed for the long term. At the same time, on the output side, the DLI will build out its interface to start connecting with other hubs and allow for an easy exchange and aggregation of link information.

So, in short, we propose an approach where on the one hand we encourage the different contributors to set up standard, automated workflows to deliver link information to a hub; while at the same time we will allow them, during a transition time, to deposit their links into the DLI in a more ad-hoc, flexible way.