RDA-WDS Publishing Data Interest Group

Data Publishing Services Working Group

Case Statement

Working Group Charter
- Objectives
- Deliverables

Value Proposition
- Who will benefit and Impact

Engagement with existing work in the area

Work Plan and Adoption Plan
- Milestones

Initial Membership

References

Working Group Charter

Objectives

The overarching scope of this Working Group (WG) is to address processes, workflows, and solutions that currently exist—mostly as bilateral agreements—between individual parties within the data publication landscape, and investigate how these can be lifted to one-for-all services that increase interoperability, decrease systemic inefficiencies, and power new tools and functionalities to the benefit of researchers. Such processes may lie at different moments in the data publication workflow, including the submission, editorial, review, and publication process.

As a primary point of focus, the WG will address the problem of limited interoperability between data repositories, scholarly journal publication platforms, and tools for bibliometric analysis. Currently, there is no common framework for cross-referencing datasets and published articles, which creates barriers and inefficiencies for the interlinking and contextualization of journal articles and datasets. This is a problem because better connections between articles and data will improve the visibility, discoverability, and usability of scientific content and serve to accelerate science in the 21st century.

To address this issue, the initial focus of this WG is to work towards a one-to-many cross-reference service for datasets and articles published in scientific journals, i.e. a service that at a minimum enables the identification of datasets associated with articles and vice versa. Additional features could include linking at different levels of granularity metadata to describe the nature of the relationship, relevant metadata for individual datasets, and articles. The WG appreciates that quality—both in terms of content and in terms of operational performance—will be a decisive factor

for successful adoption of the cross- referencing service. Equally, the WG appreciates the importance of the cross-linking service being inclusive and available to all stakeholders in the data publication landscape.

Deliverables

To achieve this, the WG has set several deliverables for 2014:

- Inventory of interlinking, cross-referencing, and other tools and processes relevant to data publication currently in place. An analysis of pros and cons, with an emphasis on scalability and doability.
- Gap analysis, including an analysis of needs & use cases for key stakeholders (data repositories, journal publishers, providers of bibliographic services, funding bodies, research institutions, researchers).
- Recommendations for a one-to-all cross-resolving service that benefits the stakeholders in data publishing. These recommendations will include technical, organizational, governance, and cost aspects.
- An operational and publicly available service for cross-referencing datasets and articles. Within the timeframe of this WG, such a service is expected to be in beta-release and with limited initial scope, but it should exemplify the recommendations of the WG in a way that is suitable for scaling up.

In addition to the cross-referencing service, which targets the later phases of the data publication workflow, the WG aims to identify other processes in the data publication landscape that are ripe for a common, one-for-all service approach. In particular, are there opportunities to create a standard, more streamlined workflow for researchers to find an appropriate data repository, to deposit their data, and to establish links to related journal publications? Are there opportunities to make it easier for editors and reviewers of scientific journals to find, access, and share relevant datasets?

Value Proposition

The outcome of this working group will benefit key stakeholders in the data publishing landscape including data repositories, journal publishers, providers of bibliographic services, funding bodies, research institutes, and ultimately researchers as data providers and data users. For all parties, access to and use of one-for-all services will enhance the discovery, availability and reliability of scientific content on the web. In the absence of accepted cross-referencing services, existing data partners adopt practical at-hand solutions to achieve system interoperability. Journal publishers work with authors and data repositories, often on an individual, mostly bilateral basis, in order to encourage and support the data providers while also establishing timely and efficient interconnections and processes. As cross-referencing services emerge, these partner-specific and discipline-specific solutions will become better interconnected, more efficient, and more dependable.

Who will benefit and Impact

In summary, the WG sees the key value elements for the proposed cross-referencing service for the key stakeholders as follows:

- For **data repositories** and **journal publishers**, interlinking journal articles and datasets will become a simpler, more scalable process with less overhead. This will help connect journal publications to underlying data, and help demonstrate how data are used in the scholarly literature.
- For **research institutes**, **libraries**, **bibliographic service providers**, and **funding bodies**, this service can power advanced bibliographic services and productivity assessment tools that track datasets and journal publications within a common and respected framework.
- For **researchers**, **data providers** and **data users**, a cross-referencing service will make the processes of sharing and of accessing relevant articles and data easier, more efficient, and more accurate. The cross-referencing service will be an enabling technology that results in better connections between different platforms for scientific content. From a reader's perspective, that means that relevant data and relevant articles will be easier to find and from an author's point of view it means greater impact for their work.

Engagement with existing work in the area

The WG appreciates the existence of a wide range of activities that are relevant to the aims and scope of this group. These activities include actual data publication models that are currently being employed or tested, as well as studies and other efforts conducted by working groups and interest groups. It is essential for the success of this WG to engage with relevant stakeholders in these activities to build a wide basis for support, to learn from their experiences, and to avoid duplication of efforts.

In terms of operational solutions, a variety of data publishing models are employed or being tested today. These models include (1) data published as supplementary files with articles, (2) data that is deposited in data repositories—either in standard community-agreed formats, or as generic files with a certain level of metadata—and then linked to journal articles, and (3) data that is published in the form of marked up, structured, and machine-readable text (e.g. linked data). Notwithstanding this diversity, these models have some common problems to solve, in particular to adopt a unified approach to cross-reference data and articles in a standardized and dependable way, and to improve interoperability between different platforms though commonly accepted data and metadata standards. By engaging with the stakeholders in existing data publication models, the WG aims to deliver recommendations and services to address these challenges.

In addition to data publishing models in operation today, the WG recognizes the work that is done by other working groups and interest groups, and seeks to learn from them and coordinate activities wherever possible. An overview of relevant initiatives, projects, and platforms will be developed and maintained at the level of the Data Publication Interest Group[1]. Relevant groups

[1] Survey on relevant initiatives, projects, and platforms: http://goo.gl/0q2f8j

within the RDA include the Metadata Standards Directory Working Group[2], the Standardisation of Data Categories and Codes[3] WG, the Data Citation Working Group[4], and the Brokering Interest Group[5]. The WG will seek interaction with these groups to cross-pollinate and to learn how their findings can be relevant for the cross-referencing service, either at the pilot stage or thereafter.

It is worth noting that the WG is not aware of other groups working on a cross-referencing service at this moment, although there are potential synergies with services available, or under development, from DataCite, CrossRef, INSPIRE, ODIN, and OpenAire.

Finally, the present WG will work in close coordination with other Working Groups in the Data Publication Interest Group, most notably the Workflows and Bibliometrics Working Groups. The work carried out in these WG's will feed into this WG as requirements for the technical capabilities and organizational structure of a successful cross-referencing service. The cross-referencing service will also be an integral part of any test implementations carried out between the groups.

Work Plan and Adoption Plan

To meet the objectives that the WG has set out, the following four key work streams have been identified:

1. **Technical**: specification and development of cross-referencing service. This work stream will deliver the functional and technical specifications for the service. This includes metadata standards and the design for a back-end system, content ingestion system, and front-end services for both human and machine consumption. The work can be divided into two phases:
    - Specification, including an inventory of existing systems and tools, gap analysis, recommendations and functional specifications for the beta cross-referencing service.
    - Development of back-end and front-end systems, and content ingestion. Note that, where possible and desirable, the WG will seek to make use of existing systems.

    The work that is carried out by Workflows and Bibliometric working groups will feed into the technical specification and development process.

2. **Building support.** This work stream will reach out to the community to engage the key players and build a solid basis of support for the cross-referencing service. To achieve this, the following activities are planned:
    - Engage key influencers such as DataCite and CrossRef.
    - Organize targeted webinars for stakeholder groups: data repositories, publishers, and journal editors & researchers
    - Presentation at the 3rd plenary RDA meeting

[2] Metadata Standards Directory Working Group: https://rd-alliance.org/working-groups/metadata-standards-directory-working-group.html
[3] Standardisation of Data Categories and Codes: https://rd-alliance.org/working-groups/standardisation-data-categories-and-codes-wg.html
[4] Data Citation Working Group: https://rd-alliance.org/working-groups/data-citation-wg.html
[5] Brokering Interest Group: https://rd-alliance.org/internal-groups/brokering-ig.html

- Ongoing support-building activities through presentations at scientific conferences and webinars
- Wide announcement of the prototype service, ideally at a major conference, supported by marketing activities, press releases and publications

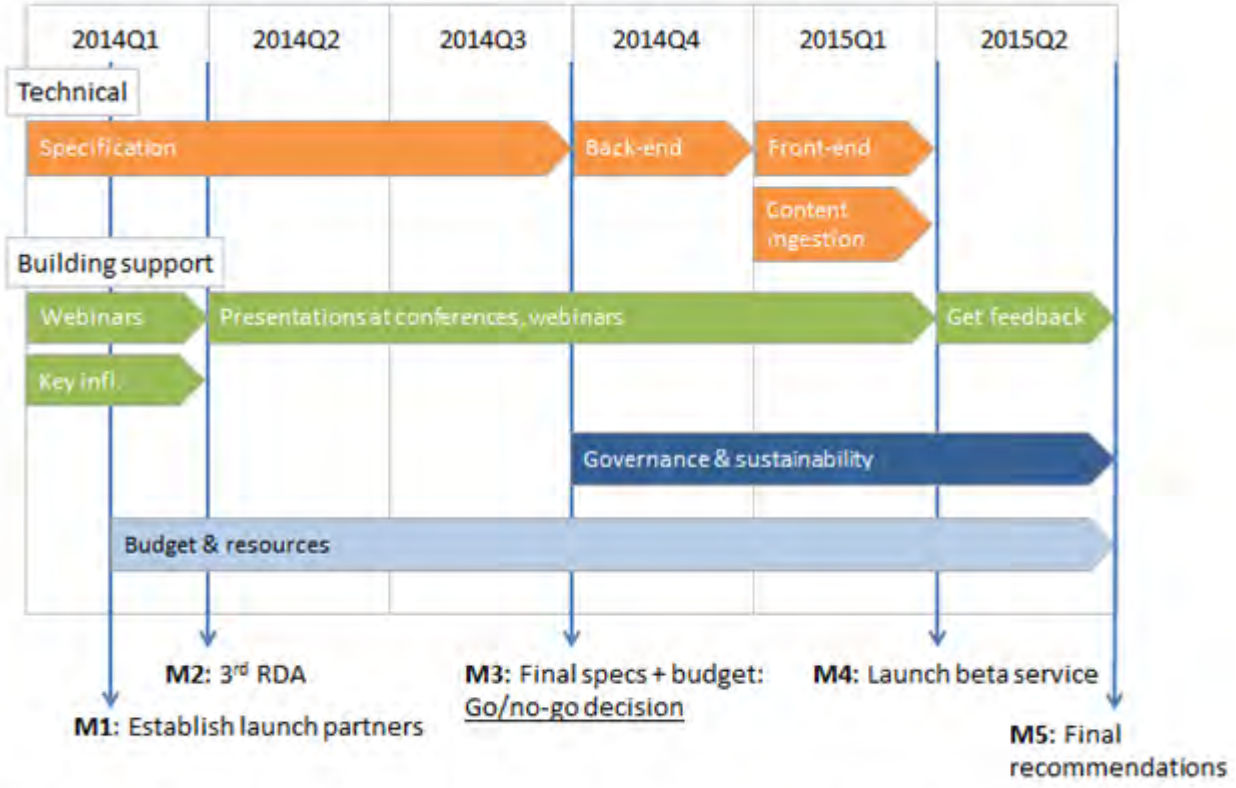
3. **Governance & sustainability.** This work stream will propose the long-term (beyond the beta phase) organizational structure to operate the cross-referencing service successfully and meets the needs of all stakeholder groups. This stream will also deliver a process through which new organizations can connect to the service.

4. **Resources and budget**. This work stream will manage budget and resourcing requirements, and seek external funding as needed to deliver the proposed service.

Note that the WG will not be conducting an initial survey, so as to be able to focus on the development of a functional prototype utilizing the expertise that is available within the WG and its network.

Work will be carried out in an 18-month time period starting from formal RDA endorsement. In terms of its internal operation, the WG will continue to meet through regular teleconferences, video calls, and face-to-face meeting.

The figure below provides a high-level summary of the activities carried out in the four work streams, assuming RDA endorsement at January 1st, 2014.



Outline of 18-month work plan with key milestones, assuming start on January 1st 2014

Milestones

The WG has identified five key milestones:

- **M1: Establish launch partners**. Launch partners are organizations that commit to contributing to the corpus of data-article associations that will be made accessible through the cross-referencing service. These organizations will also be asked to provide input on functional specifications and will be invited to provide early feedback on the system. Launch partners will initially be sought from within the working group, or from the extended network provided by the Data Publication Interest Group and its members. While working with a selected group of launch partners will be necessary to develop a pilot implementation, the full cross-referencing service will be inclusive and open to any organization that meets certain minimal quality criteria.
- **M2: 3rd RDA Plenary** (26-28 March 2014). This is an opportunity to build support, and also to share initial ideas on the specification of the cross-referencing service and gather feedback. Based on that feedback, requirements can be adjusted early on in the process as needed.
- **M3: Final specifications and budget known: go/no-go decision**. When the final specifications, together with budget and resource requirements, are known then the go/no-go decision will be made to build the cross- referencing service. This decision will be made by the Data Publication Interest Group Co-chairs, together with the Chairs of the Working Groups under the Interest Group. In the case of a no-go, the deliverables of the Working Group will be reviewed.
- **M4: Beta cross-referencing service go-live**. This will be announced widely, ideally at a major conference, and supported by marketing activities and possibly one or more scientific publications. The WG will endeavour to reach out to as many organizations as possible, asking them to test the service and provide feedback.
- **M5: Final recommendations**. The WG will deliver a final report on its work, including recommendations on the future of the cross-referencing service and on other components of the data publication system that are ripe to be lifted to a one-for-all service architecture.

Initial Membership

Membership of the WG aims to reflect the key stakeholders in Data Publication Services, including data repositories, publishers, journal editors, researchers, research centres, providers of bibliographical information, and funders. The WG members listed alphabetically by surname:

- Hylke Koers (Elsevier) **CHAIR**]
- David Anderson (NOAA)
- David Carlson (ESSD)
- Janine Felden (MARUM)
- John Helly (UCSD)
- Francisco Hernandez (Flanders Marine Data Center)
- Paolo Manghi (OpenAire)
- Caroline Martin (Sciences Eaux & Territoires Journal / IRSTEA)
- Jo McEntyre (EBI)

- Lyubomir Penev (Pensoft Publishers)
- Nigel Robinson (Thomson Reuters)
- Johanna Schwarz (Springer)
- Eefke Smit (STM)
- Juanle Wang (WDC for Renewable Resources and Environment)
- Eva Zanzerkia (NSF)

References

All of the working groups in the Data Publication Interest Group have a common bibliography[6] in which publications relevant for this particular group are marked correspondingly.

[6] Bibliography: http://goo.gl/wA1G27