RDA-WDS Publishing Data Interest Group

Data Publishing 2020:
Proposal for a Coordinated Approach

Preface
Overall objectives
Working Groups
- Workflows for publishing data
- Bibliometrics for published data
- Services for publishing data
- The costs for publishing data

Common approach
Outlook
Management

Preface

The following case statements build on a range of initiatives in publishing data. In late 2012 an ICSU-WDS initiative on data publication was started and endorsed in 2013 by RDA as an RDA-WDS Interest Group (IG) on Publishing Data. The aim of this proposal is to identify and define best practice for publishing data and to test its implementation across the core stakeholders involved: Researchers, Institutions, Data Centres, Scholarly Publishers, and Funders. Currently, publishing data faces certain core problem, which are best ironed out in its early days, when the habits and customs are still flexible. The original ICSU-WDS concept addresses essential problems in this area and implications for the different stakeholders. Moreover, it was clear from the outset that topics are interlinked and that it will be difficult to address them separately. For this reason, we have decided to bring 4 Working Groups (WGs) under one umbrella to work on these topics in close conjunction.

These 4 working groups cover the following aspects of data publishing:

- Workflows
- Bibliometrics
- Services and registry
- Cost recovery models

These 4 WGs are closely interlinked. For example, bibliometrics on published data depends on the way data are published and cited, which in turn strongly influences the way services supporting the publication of data can be set up. In addition, any conceived solution will raise financial issues and thus raising the question on how resources and costs for publishing data may be identified and addressed. Interaction and exchange of results between groups is

monitored and guided through a common management structure that ensures involvement of the main stakeholders at all levels.

The concept of this proposal supplies a holistic approach aiming at pushing and establishing data publishing amongst stakeholders. All this eventually led to the decision to submit WG Case Statements to RDA as a bundle with each WG being indispensable for the success of the planned work.

Overall objectives

In the empirical sciences, data has traditionally been an integral part of scholarly publishing. However, rapid technical developments, such as digital data and high-throughput techniques, dramatically changed the scholarly publishing paradigm in the last decades. This requires new approaches to ensure availability and usability of science data. However, existing approaches to address this issue are mostly technically dominated and lack success because they do not supply the necessary benefit for data producers and the wider community and society. Instead, the concept of *Data Publication* is undergoing a renaissance as part of scholarly communication and on the base of new and proven technologies. Publishing data is a new and strong incentive for scientist to share their data and has positive effects on the data quality. The impact on citation rates can be seen in recent bibliometric studies on science articles providing access to underlying data.

Publishing data can follow good practice of conventional publication of articles in journals that includes online submission, quality checks, peer-review, editorial decisions, and an equivalent of ‘page proofs’. In fact, storage of data in public repositories and the ability to reference datasets is getting increasingly important. It is already mandatory for the acceptance of peer-reviewed publications in specific fields of research such as molecular sciences or ecology. However, *Data Publication* as a generally accepted new publication type—self-standing or supplementary to literature—is not without controversy.

For data centres, science publishers, and service providers data publication is a challenge in terms of organization, technical developments, and funding. Compared to science articles the economic value of data is generally higher but they also need more resources for production, processing, long-term archiving and publication. If published data are to be usable and as reliable as peer-reviewed science articles they should not only meet scientific requirements, but also archival and longevity requirements. Archiving and publishing procedures for data must be transparent and accepted as part of the science culture. Moreover, published data should conform to generally accepted content and interoperability standards, thus allowing for efficient usage and integration of data from various sources.

The overall objectives of the WGs under the umbrella of the RDA-WDS Publishing Data IG are to incentivize and enable researchers to publish data by:

- Promoting and establishing the data publication concept amongst data centres
- Promoting and establishing the data publication concept amongst professional and academic publishers and bibliometric service providers

- Establishing data publishing services amongst researchers and institutions as part of scholarly publishing

Working Groups

Four WGs have been set up, each of them addressing essential and practical issues to help enable the publication of research data as part of the scholarly record.

Workflows for publishing data

A number of workflows for publishing data exist, all of them individually adapted to their specific environments, but many researchers are unaware of them or they may not suit their needs. This WG will classify a representative range of existing workflows and identify generic components, which might be reused or repurposed elsewhere. These might include the challenges of technical and research QA/QC and peer-review as well as the role of researchers, their institutions, data centres, academic publishers and funders in the data publication process. The WG will then identify use case(s) in which one or more generic workflow component may be tested so that we can evaluate the benefits for all stakeholders involved and enable improved data discovery and reuse.

Bibliometrics for published data

As one of the major means to measure research productivity, bibliometrics of research data practically does not exist. The way data are referenced is inconsistent and—aggravating the situation—citing data is not common practice in scholarly publishing. Instead, data centres are trying to keep track of literature using their data or by supplying other metrics such as download statistics. The WG will investigate approaches and develop solutions that allow proper analysis of content and citations.

Services for publishing data

During the last century, services for publishing data have concentrated on registration of data entities. Services to cross-reference research data and literature or to publish data have only started recently and are limited in scope and functionality. The WG will investigate content and interoperability requirements for data centres and academic publishers. Building on existing components, the WG will concentrate on the conception and implementation of a one-for-all cross-referencing service.

The costs for publishing data

General services to publish data are not available and the necessary editorial process leading to quality assured and efficiently usable data requires resources to be quantified. At present, there is an imbalance between the capacities and functionality of existing data centres and data repositories and the global production of scientific data. Budgets of data centres generally cover a precise scope, mostly data production of the host institution. The

WG will supply cost estimates and elaborate a business model to compensate for additional costs of publishing data in an open access environment.

Common approach

The work programme for all WGs builds on results from previous or current initiatives and projects, in particular on current collaborations between ICSU-WDS data centres and services, academic publishers, and service providers. The topic of data citation has been treated exhaustively in the CODATA-ICSTI-WDS Task Group on Data Citation and the SCOR-IODE WG on Data Citation. Results from these groups will be helpful in particular for the bibliometrics WG. For more technical citation issues, the group will collaborate with the RDA WGs on data citation and persistent identifiers.

To supply an overview on previous work the group has compiled two documents, a bibliography of articles addressing one or more aspects of data publication[1] and a survey of related initiatives and projects[2].

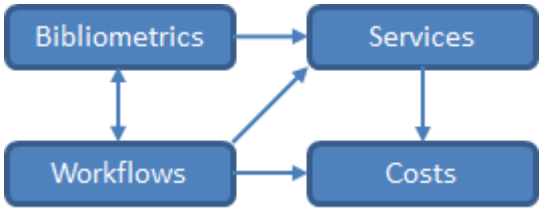

Figure 1: Relationship between WGs

As mentioned in the preface, coordination and combined testing of the various models, scenarios, and services developed in the four WGs is planned (Figure 1). In particular:

- Workflow reference models are matched with possible bibliometric scenarios (e.g. data references in journal articles, articles references in data publications, possibly also an altmetrics scenario reflecting usage of datasets)
- Workflow reference models and bibliometric scenarios are matched with capabilities of cross-linking service
- Workflow reference models are matched with cost compensation models
- Use cases are selected (and possibly bundled) to cover workflow reference models (e.g. standalone data publishing and data supplements to journal articles), cost compensation and bibliometric models, as well as different cross-linking mechanisms (e.g. data to article references, article to data references)
- Use cases are tested for the various combined aspects: workflow, bibliometrics, cost compensation, cross-linking.

[1] Bibliography: http://goo.gl/wA1G27
[2] Survey: http://goo.gl/0q2f8j

Outlook

The WGs aim at establishing data publishing among stakeholders. Therefore, towards the end of the 18-month period, part of the activities will concentrate on elaborating recommendations and concepts for testing the various models and prototypes in close collaboration between the different stakeholders from our groups. In particular, the WG workflows will deliver a concept for a test environment covering the different classes of workflows and involving all stakeholders with a representative number of data centres and science journals. The test environment will also include the operation of the cross-referencing service implemented by the WG publishing services and the corresponding bibliometrics system from the WG bibliometrics. Finally, the test environment will include monitoring the costs and the effectiveness of the potential cost recovery options.

Setup and operation of the test environment will need additional resources. The group therefore plans to apply for funding (e.g. from the EC Horizon 2020 program), whereby WG results and membership will supply the base for a corresponding proposal, which needs to be ready before the second half the 18-month period to ensure the seamless continuation of work.

Management

All WGs are managed collectively. The Project Board consists of the Chairs and Co-chairs of the IG and the four WGs plus the Executive Director of ICSU-WDS. The board is taking decisions on the general direction of the project, resourcing, planning, budgeting, membership, rescoping, and refocusing. Assignment of Chairs and Co-chairs is decided by the Co-chairs of the IG and the ICSU-WDS ED in consultation with the Chairs and Co-chairs of the WGs and approved by the members of the corresponding WG. Teleconferences are scheduled every six weeks between the Chairs of the WGs and the Co-chairs of the Publishing Data IG. In addition, the WGs have face-to-face meetings with all WG members. Thus, coordination of work between groups and timeliness of results is ensured. Meetings will also be used to discuss membership in the groups as well as input supplied within the wider context of the IG and other RDA WGs.

The WGs will also make use of the ICSU-WDS network including not only its members' data centres but also science publishers and other service providers. Some of the input needed for the different tasks will be received through the yearly reports of WDS Members, possibly by requesting specific information systematically (e.g. current practice of archiving and publishing data or relationship of archived data with literature). Other collaborative means are a set of mailing lists, individual for each group and for the overall group; teleconference system supplied by ICSU-WDS, and shared documents.