RDA-WDS Publishing Data Interest Group

Data Publishing 2020: Proposal for a Coordinated Approach

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].

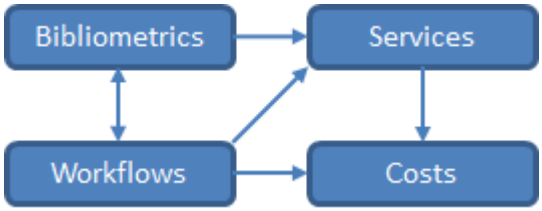


Figure 1: Relationship between WGs

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.

RDA-WDS Publishing Data Interest Group

Bibliometrics Working Group
Case Statement

Working Group Charter

Objectives

Bibliometric indicators are essential to obtain quantitative measures for the assessment of the quality of research and researchers and the impact of research products. Systems and services such as the ISI's Science Citation Index, the h-index (or Hirsch number), or the impact factor of scientific journals have been developed to track and record access and citation of scientific publications. These indicators are widely used by **investigators**, academic departments and administration, funding agencies, and professional societies across all disciplines to assess performance of individuals or organizations within the research endeavour, and inform and influence the advancement of academic careers and investments of research funding, and thus play a powerful role in the overall scientific endeavour.

The basic idea of bibliometrics is to evaluate the attention scientific publications receive within the scientific community. The classical approach is based on counting formal citations in the literature, and despite various critical aspects—ambiguity of authorship, self-citations etc.—these indicators have become widely adopted across all of science. Similar indicators for the value and impact of data publications are needed to raise the value and appreciation of data and data sharing as the missing recognition for data publication in science is seen as the major cause for the reluctance of data producers to share their data. The overall objective of this working group therefore is to

conceptualize data metrics and corresponding services that are suitable to overcome existing barriers and thus likely to initiate a cultural change among scientists, encouraging more and better data citations, augmenting the overall availability and quality of research data, increasing data discoverability and reuse, and facilitating the reproducibility of research results.

Principally, existing metrics for scientific papers could also be applied to data publications. However, an extrapolation of the classical bibliometric approach to research data are difficult to realize because:

- Citing data are not a standard practice in the scientific community. At present, references to data in the literature are rare and do not follow a generally agreed schema. No recommended Best Practices for data citation exist. This is also true for data products compiled in general from already published data.
- There is a large variety in the structure and practices of data repositories. Many repositories are not prepared for the data publishing concept, and have not implemented formal data publication procedures. Granularity, versioning, persistent identification, metadata, and review of data entities are among the unresolved issues.

Besides the classical approach, various alternative metrics for data evolved during the last years. These so-called 'altmetrics' are based on data usage analysis (except citations as indicators of usage) and content evaluation quantified e.g. through dataset downloads or analysis of annotations of datasets by users (social tagging). However, applications of existing solutions are isolated and scarcely comparable, thus are currently not usable as a basis for representative indicators. Nevertheless, seeing the potential and dynamics behind developments, altmetrics need to be considered as serious concepts beside the classical approach.

Any approach to data metrics needs to address the challenge of a cultural change in science toward full appreciation and recognition of data as an essential part of the scholarly record. Metrics for data need to be designed and conceived in a way that all stakeholders will embrace them as credible, valuable, and meaningful.

Deliverables

As a summary, one may say that at present there is no generally acknowledged metric for data. This Working Group will bring together the essential stakeholders in this field, will investigate the requirements and recommend necessary steps to be taken. Activities will address different levels:

- Organizational: What are the overall changes in the scholarly publishing system needed to foster proper attribution of datasets? Which are the building blocks for an optimal system? Which changes are needed from funders, data centres, science publishers, and science service providers? What is the optimal way of interaction between stakeholders? Do we need commonly operated services?
- Technical: Which are the technical components, interfaces, and standards that need to be developed and used? What current capabilities can be adopted as solutions, what is missing?
- Methodological: What methodologies for data metrics need to be developed? What are the costs and benefits of altmetrics versus traditional processes? What research into indicators is needed and what are the strengths and weaknesses of individual indicators?
- Financial: What are the costs for data metrics (seen as a cost component of data publication)? Who will pay for it?

This Bibliometrics WG is part of the overarching RDA-WDS IG on Data Publishing and as such covers a particular thematic field. On the one hand, the group relies in part on the results from the other groups—in particular the workflows WG— and on the other hand, it delivers results to the other groups—here in particular to the data publishing services WG.

Value Proposition

Good and practicable bibliometrics are fundamental for establishing data publication and data sharing as a recognized contribution to science. This is a prerequisite for realizing the vision of an open, comprehensive, global knowledge base of scientific data as the new paradigm for scientific discovery in the 21st century.

Who will benefit

Data bibliometrics will allow data producers, data centres/publishers, data managers, research facilities and academic institutions, science publishers, and funding agencies to demonstrate quantitatively and formally the significance and viability of data to the advancement of science.

Impact

We anticipate that bibliometrics for data will have a profound impact on the willingness of researchers to make their data openly accessible, on the availability of sustained funding for data centres, and on the institutional changes in academic institutions to acknowledge formally data contributions as part of the scholarly record that is used in tenure and promotion evaluations. These anticipated cultural changes will likely lead to a rapid growth of available data.

Engagement with existing work in the area

An overview of relevant initiatives, projects, and platforms will be developed and maintained at the level of the RDA-WDS Publishing Data Interest Group, and may be found in the survey[1].This WG will focus on bibliometrics but will also keep in touch with the RDA data citation WG which has a more technical scope. Collaboration will be sought in particular with the following groups:

- DataCite—who are the main minters for data DOIs and are providing some statistics on data DOI resolutions
- Data Citation Index Thomson Reuters—who are planning on tracking data citation metrics
- ICSU CODATA WG Data Citation—who are providing guidance on definitions and syntax for data citation
- CrossRef—who will be exchanging metadata profiles with DataCite
- Altmetrics, Mendeley—for their tools in tracking the impact of non-article research outputs
- ImpactStory—who assess broad impacts of diverse products including papers, datasets, software, etc.
- Force11—who are working to synthesise and refine principles for data citation
- Scopus - whose abstract and citation database of peer-reviewed literature features smart tools to track, analyse and visualize research

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

Work Plan

The Bibliometrics WG envisions completing four tasks that will feed into and coordinate with the other work done by the other working groups in the Publishing Data Interest Groups:

1. **Compilation of existing work.** To date considerable work has already been done to make the case for citing data. Both the general 'why' and the general 'how' are well served by universities, data centres, and international initiatives such as DataCite, as well as various specialized groups and facilities, such as DCC, AGU and so forth. There has been far less—or less well-known—work performed which analyses the specific requirements for particular subject areas or research communities. Current barriers and potential solutions are not well covered, and while there has been much consensus that 'this is a big issue that someone needs to tackle', little practical progress has so far been made.
2. **Summarise current practice and policies** of data centres, funders, journals, learned societies and publishers (examples: PANGAEA, IEDA, EBI, Dryad, GBIF, Pensoft data journals, Nature, Elsevier, Wiley, Scopus, BioMedCentral, PLOS. eLife, PeerJ, F1000, EMBO journals, RCUK, NSF, EU, ANDS, AGU, RMetS). What recommended practices exist, who is citing whom, how are they citing, how are metrics used, who is evaluating what, what criteria are considered meaningful and valuable? Are there consequences for non-compliance? If so, how are they policed/enforced? What are the commonalities across the stakeholder community?
3. **Evaluation of possible approaches [potentially as a survey]** (impact & feasibility), including altmetrics, Data Citation Index, Mendeley, and other usage measures which are emerging as a result of Open Access in general. The WG will evaluate recent projects such as PREPARDE, JoRD, pilots such as F1000R with Figshare, Altmetric with various publishers, Elsevier's linking with some data centres, as well as future possibilities with CrossRef, DataCite, Thomson Reuters Industry Forum, STM Association, EarthCube, and others.

   The survey should target anyone who will be using bibliometrics for data, for example researchers, research project administrators, librarians/repository managers, research funders, tenure committees/institution administrators, journal publishers, research infrastructure managers, etc.
   a. Survey task breakdown
      i. Determine audience: which questions to target which groups?
      ii. Develop questions to evaluate current approaches outlined in our summary.
      iii. Develop questions to ask for feedback for our user requirements deliverables—"Do the current practices meet your needs?", "What are your bibliometrics needs?", "If you are aware of bibliometric tools, but do not use them, why? And what would encourage you to use them?"—and needs for information/communication of bibliometrics tools. The questions will be a combination of ranking on a scale and free text, as appropriate.
      iv. Keep the survey short, and phrase questions and preamble to encourage responses from broad range of stakeholders. We might need to adapt the survey according to the interest of the different stakeholders.
4. **Develop recommendations [based on the survey]** for what is required and what steps need to be taken. These will address different levels of granularity as required:
   a. Organizational: What are the overall changes in the scholarly publishing system needed to foster proper attribution of datasets, and how can they be successfully

achieved? Is there a distinction to be made between formal and informal metrics? What are the building blocks for an optimal system? What changes are needed from funders, policy makers, data centres, science publishers, learned societies, and science service providers? What are the optimal channels of sustained interaction between stakeholders? Is any group unrepresented by the current WG? Do we need commonly operated services? How community-specific do recommendations need to be in order to support change?

b. Technical: What technical components, interfaces, and standards are needed? What is currently available, usable, and appropriate, what is missing? How can necessary technical changes be implemented most efficiently and effectively?
c. Methodological: What methodologies for data metrics need to be developed? What are the costs and benefits of altmetrics versus traditional processes? What research is needed into indicators and what are the strengths and weaknesses of individual indicators? Survey.
d. Financial: What are the costs for data metrics (seen as a cost component of data publication)? Can current cost models for bibliometrics be adapted for data?

Adoption Plan

Deliverables of the WG

The WG will generate recommendations for all the key stakeholders: funders, learned societies, data centres, publishers and researchers (end 2014). The recommendations will comprise:

1. Case studies: Based on compilation of existing work (see above).
    - Identify potential partners by name/organization. Ideally involve them and achieve buy-in from these entities ahead of publication of the deliverable itself to ensure forward momentum.
    - Clear list of bibliometric examples and potential use cases. Include a variety of data types, subject areas and bibliometric types (citation, usage, social media as appropriate)
2. General requirements for citability of scientific data (granularity, citation information and persistent identification) - rely on CODATA Report and the response to survey (described in work plan above) where possible.
3. Use cases and requirements: to provide guidance on concrete, practical next steps. Plans for implementation of a bibliometrics system (as input to the Publishing Services WG) and user consultation with stakeholders on those plans.

Milestones and intermediate products

1. See the bullet points above in the Work Plan
2. IDCC in February 2014: Survey ready to distribute and broadcast at this meeting. Produce flyers to be distributed with conference materials with QR code.
3. Dublin plenary in March 2014:
    3.1. Present the summary of current practices
    3.2. Have survey ready for deployment, along with preliminary results from the first month of survey collection.

Project Management

Mode and frequency of operation

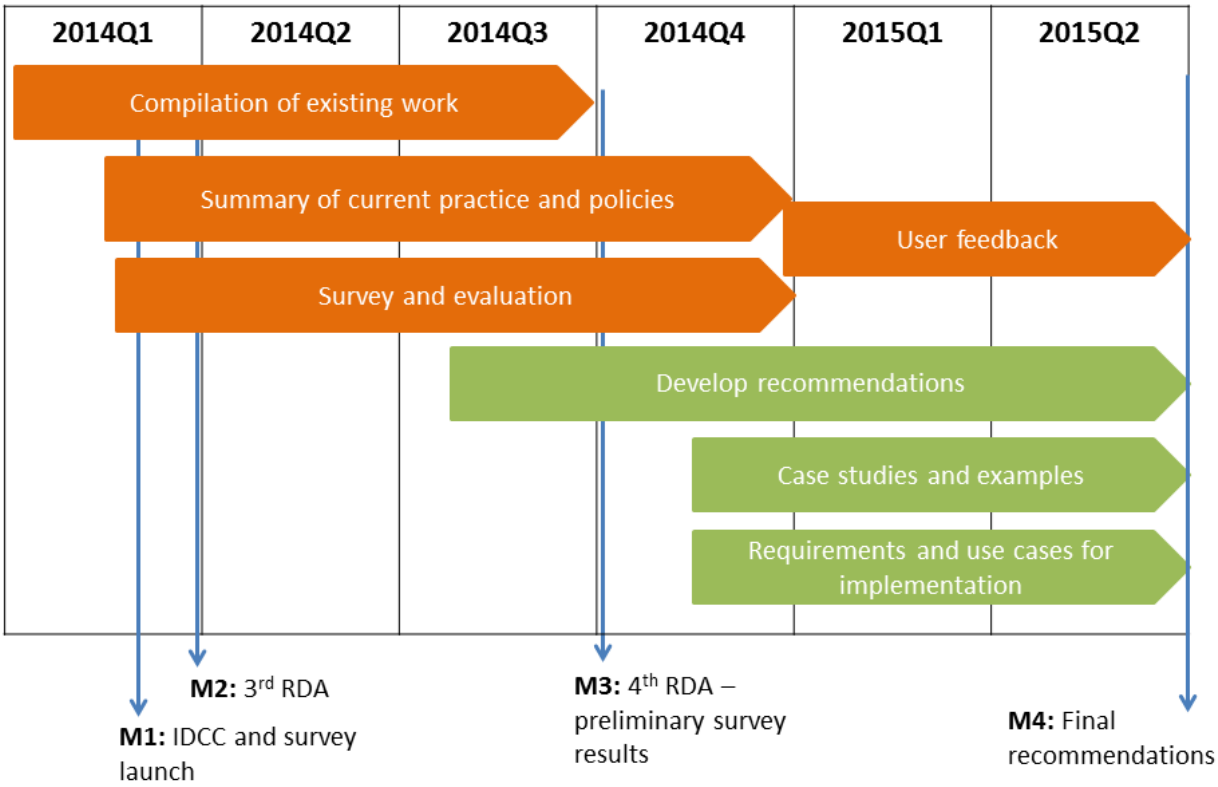
1. Every 6 weeks: Teleconference between Co-chairs of the Publishing Data Interest Group and Chairs of the Working Group
2. Open Webinars for dissemination and exchange
3. Face to face meetings during RDA plenary meetings

Consensus, conflicts, staying on track and within scope, and moving forward

The WG will hold teleconferences approx. every 6 weeks to discuss assignments and progress toward the deliverables. Sub groups to take responsibility for specific sections and/or projects. Bring in ad-hoc members to test and refine areas where additional expertise is required, or if there is a diversity of opinion or other complication.

Planned approach to broader community engagement and participation

Via webinars, conferences, and intermediate reports, findings and step-by-step deliverables will be communicated to members of the Data Publishing IG, WDS members, and additional stakeholders



and initiatives addressing the same topic.

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.

Membership

- Kerstin Lehnert (US, IEDA, WDS) [**CO-CHAIR**]
- Sarah Callaghan (UK, BADC) [**CO-CHAIR**]
- Jan Brase (Germany, DataCite)
- Ross Cameron (The Netherlands, Scopus)
- Cyndy Chandler (US, Woods Hole Oceanographic Institution)
- Ingeborg Meijer (The Netherlands, University of Leiden)
- Fiona Murphy (UK, Wiley-Blackwell)
- Lyubomir Penev (Bulgaria, Pensoft Publishers)
- Fiona Nielsen (UK, DNAdigest.org)
- Nigel Robinson (UK, Thomson Reuters)
- Mary Vardigan (USA, ICPSR)
- Jochen Schirrwagen (Germany, Universität Bielefeld)

References

All of the working groups in the Publishing Data Interest Group have a common bibliography[2] in which publications relevant for this particular group are marked correspondingly.

Other references specific to this case statement are:

- Jason Priem, Dario Taraborelli, Paul Groth, Cameron Neylon, “altmetrics: a manifesto”, v1.01, Sept 2011[3]
- NISO to Develop Standards and Recommended Practices for Altmetrics, 20 June 2013[4]

[2] Bibliography: http://goo.gl/wA1G27
[3] “altmetrics: a manifesto” http://altmetrics.org/manifesto/
[4] NISO to Develop Standards and Recommended Practices for Altmetrics: http://www.niso.org/news/pr/view?item_key=72efc1097d4caf7b7b5bdf9c54a165818399ec86

RDA-WDS Publishing Data Interest Group

Cost Recovery for Data Centres Working Group Case Statement

Working Group Charter

Objectives

Basic funding of data infrastructure may not keep pace with increasing costs[1]. Therefore, there is a need to consider alternative cost recovery options and a diversification of revenue streams. In short: who will pay for public access to research data?[2]

This Working Group proposes to make a contribution to strategic thinking on cost recovery by conducting research to understand current and possible cost recovery strategies for data centres. The Working Group will pay particular attention to data centres' involvement in data publishing activities and examine such initiatives as a potential source of alternative revenue.

The Working Group will produce a report providing conclusions and recommendations about the potential appropriateness of different cost recovery models to different situations and the potential of data publication initiatives fitting into a cost recovery strategy. The Working Group will also

[1] Sustaining Domain Repositories for Digital Data: A White Paper http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

[2] F. Berman, V. Cerf, 'Who will pay for public access to research data?' in Science Magazine, vol.341, p.616-617

contribute its findings to the combined testing of the various models/scenarios/mechanisms developed in the four Data Publishing Working Groups.

These deliverables will build on five areas of work:

1. A summary of current work on cost models[3];
2. A survey of funding policies specifically relating to how the costs of data availability/publication may be recovered;
3. A survey, by means of a questionnaire and case studies, of various existing approaches to cost recovery/business models;
4. A survey of other stakeholders (publishers, researchers) to understand their position and policy in relation to charging models and their role in the publishing process;
5. The outcomes of the Working Group on Workflows.

Value proposition: Cost recovery for Data Centres

The challenge: cost recovery and sustainability of public data products.

A number of initiatives examining data citation and data publication aim to integrate data more effectively with the process of scholarly communication and the 'record of science'. A vision of the future in which 'data papers' have scholarly currency, in which data can be accessed and visualized directly from the online literature requires partnerships between publishing platforms and data centres (e.g. [4]). This vision also demands that data centres providing access to published datasets should have sustainable business models.

A lot of work is going on to understand the costs of maintaining long-term accessibility to digital resources, to identify different cost components and based on this to develop cost models[5]. However, in a broader context—which considers data as part of research communication—the identification of costs and development of cost models addresses only part of the problem. In times of tightening budgets, it is important to address the challenge of ensuring the sustainability of data centres - and considering this in the context of the broader processes for data publication. Many established national and international data centres have reliable sources of income from research funders. However, these sources of income are generally inelastic and may be vulnerable. There is concern that basic funding of data infrastructure may not keep pace with increasing costs. Therefore, there is a need to consider alternative cost recovery options and a diversification of revenue streams.

This Working Group proposes to make a significant—but achievable—contribution to strategic thinking in this area by conducting research to understand current and possible cost recovery strategies for data centres. We will pay particular—but not exclusive—attention to data centres'

[3] In this document, we will differentiate between two aspects: 'cost models', i.e., a description of what different aspects of data curation and storage cost, and 'cost recovery', i.e., models and ways in which data centres can charge for their services. We consider both components to be within scope of this working group, but the focus of the questionnaire and the testing is to provide a practical overview and advice on various cost recovery models.

[4] Science as an open enterprise, The Royal Society Science Policy Centre report 02/12, Issued: June 2012 DES24782, ISBN: 978-0-85403-962-3, The Royal Society, 2012

[5] www.life.ac.uk, www.costmodelfordigitalpreservation.dk, www.beagrie.com/jisc, http://brtf.sdsc.edu/, http://www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2, http://www.alliancepermanentaccess.org/index.php/aparsen/, http://4cproject.net/

involvement in data publishing activities and examine such initiatives as a potential source of alternative revenue.

By means of a questionnaire and a set of case studies, this working group will shed light on data centres' current practice of cost recovery and identify possible opportunities for data centres looking to diversify income streams. A number of important questions will be considered, including:

- What cost recovery models are currently being employed by data centres?
- What trends are perceived by data centres with regard to the vulnerability of funding and what are the possible responses to diversify income streams?
- What cost recovery models are available within current, largely grant based funding of research?
- What cost recovery strategies are available while maintaining a commitment to open access to research data?

The principal activity of the WG will be to survey a set of data centres and provide a group of case studies addressing these questions in detail, for use within the test environment developed by other RDA Working Groups. This work will build upon existing work on cost models and funder's policies, as they relate to cost recovery of data curation costs from research projects. These aspects of the work will help the WG analyses how the costs of particular activities may be covered, and to what extent it is possible to hypothecate charges and encourage clarity around who pays for what. However, the principal focus will be on understanding the alternative options available for cost recovery and diversification of revenue streams for data centres. There are various options available and the involvement of data centres with 'data publication' initiatives is a significant new development that will be considered in this study.

Summary of the cost components of a data publication:

- Ensuring a publishable data product—annotation, metadata, codebooks etc.—is argely the responsibility of the researcher.
- Quality assurance and review process is conducted in some instances by the publisher but also, largely, by the data repository.
- Long-term preservation—archive and services for access— is largely the responsibility of the data repository.

Scope and Terminology

The working group has a realistic objective of conducting a survey among a limited set of data centres in addition to a small number of carefully selected case studies. In order to be as informative as possible, this research will be narrow and deep. Additional funding will be sought to allow the expansion of this activity, the involvement of a greater number of data centres in the survey and as case studies.

Deliverables

The working group will produce a report providing conclusions and recommendations about the potential appropriateness of different cost recovery models to different situations and the potential of data publication initiatives fitting into a cost recovery strategy.

With respect to the implementation of the recommendations from the final report, we choose the following approach. The recommendations will produce a number of scenarios for cost recovery. These scenarios will contain criteria to determine their appropriateness to different situations. For each scenario a use case will be tested in practice by one of the stakeholders involved (digital repository/publisher/research institution/etc.). The cost recovery model will be embedded in their services and their administrative and financial workflows.

Next to this, the Working Group will contribute its findings to the combined testing of the various models/scenarios/mechanisms developed in the four Data Publishing Working Groups. We will match their workflow reference models with our cost recovery scenarios: what are the practical implications of the different scenarios on the workflows?

These deliverables will build on five areas of work:

1. A summary of current work on cost models—the WG will simply provide a summary of existing work (4C, ICPSR, APARSEN, etc.)—will inform our own analysis.
2. A survey of funding policies specifically relating to how the costs of data availability/publication may be recovered—the WG will provide a brief summary of funder expectations or rules governing the funding of data deposit from research grants. This will build on some existing work conducted by the Knowledge Exchange, DCC and others. We will follow up with specific approaches to certain funders to understand the principles underpinning the policies and any possible changes in direction.
3. A survey—by means of questionnaire and case studies—of various existing approaches to cost recovery/business models. This will be the core new activity of the Working Group. The WG will survey members of the World Data System, holders of the Data Seal of Approval, members of ICPSR and other established data centres.
4. On the basis that possible cost recovery options include pay-to-deposit and pay-to-access, other stakeholders—funders, publishers and researchers—will be surveyed to understand their position and policy in relation to charging models and their role in the publishing process.
5. The outcomes of the Working Group on Workflows. This Working Group will produce a classification of a representative range of workflow models. In each case, the varying stakeholders and their different roles and responsibilities will be identified as well as the likely associated resource and cost implications.

Who will benefit

The principal beneficiary of this work will be data centre managers who will have an insight into alternative options for cost recovery, substantiated by case studies. Other stakeholders in data publication will benefit from a clearer insight into the relationship between policy, funding and cost

recovery. The ultimate contribution of the WG will be to present the community of stakeholders with examples of how sustainability of data infrastructure for publication may be achieved.

Engagement with existing work in the area

There is a significant existing body of work on cost models. The European 4C project has undertaken a significant task of synthesis in this area. Other initiatives on costs and cost models include Knowledge Exchange and APARSEN.[6] The Working Group will closely monitor the developments and outcomes of this work and it will feed into our summaries of cost models and the survey of policies.

We are not aware of other work in the specific area of cost recovery, hence the need for this Working Group. There are a number of practical initiatives for data publication exploring new business models for data repositories, for example, Dryad.[7] Similarly, a number of established data centres (DANS in the Netherlands, ADS in the UK) are exploring new or supplementary approaches to cost recovery and these will be included among our case studies.[8]

The work of other Working Groups within the RDA-WDS Data Publication IG will be taken into account, in particular the Data Publication Workflows WG. The outcomes of our Working Group will feed into the Workflows WG.

We will also pay attention to the RDA Data Foundations and Data Certification Groups and the work of these groups will inform the framework used. Nevertheless, the focus of this WG is very much on understanding existing and possible approaches to cost recovery and this is a niche that is not being explored by other groups.

Work Plan

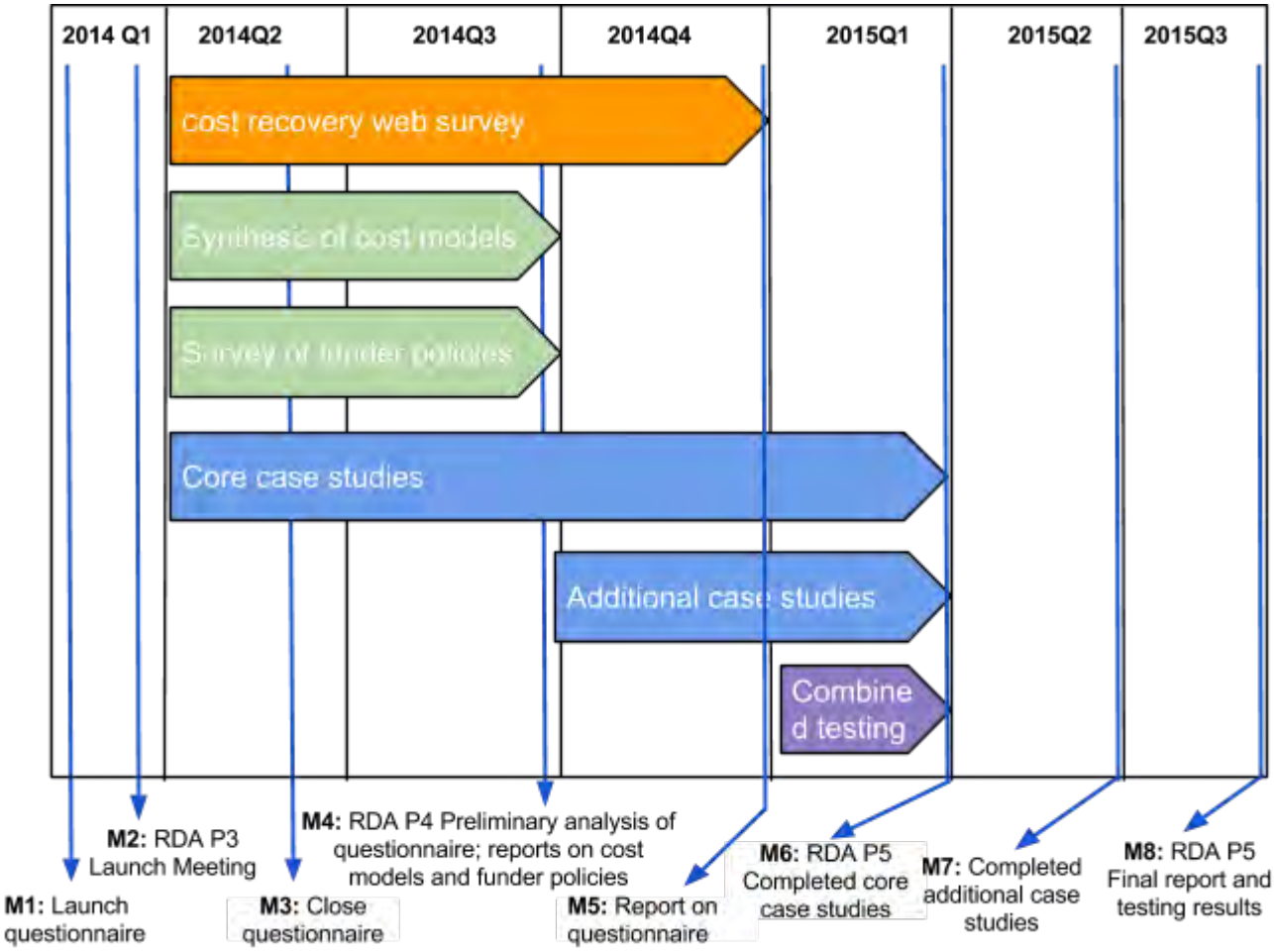
Deliverables and activities

1. Summary report of existing cost models to identify key cost points.
    - Summarize current cost models with specific consideration of how cost points identified might relate to cost recovery (or not).
2. Summary report of funder data policies insofar as they relate to cost recovery. This report will identify what funder policies say about the scope for cost recovery of data infrastructure from research grants (link with Knowledge Exchange work on funding).
    - Survey funder data policies insofar as they relate to cost recovery.
3. Questionnaire to survey cost recovery in a sample of data centres.
    - Design and set up questionnaire.
    - Compile list of target participants.
    - Conduct survey.
    - Analyse results.
4. Case studies from a selected set of data centres, providing more detailed analysis of cost recovery approaches.

[6] http://4cproject.net/, http://www.knowledge-exchange.info/, http://www.alliancepermanentaccess.org
[7] http://datadryad.org/
[8] www.dans.knaw.nl, http://archaeologydataservice.ac.uk/

- Set up initial set of case studies (e.g. ADS, Dryad, DANS, CCDC)[9]
- Results of questionnaire may be used to identify additional case studies. Case studies will provide detailed examples of cost recovery models; the rationale behind the choices made; the experience of the model and the current estimate of its appropriateness and likelihood of success.

5. Report on potential cost recovery models
   - On the basis of the questionnaire, case studies, and stakeholder engagement, the Working Group will prepare a report summarizing available cost recovery models and identifying the most likely ways in which providers of data infrastructure may diversify their income streams.
   - Which cost recovery strategies are most likely to be transferrable?
   - Are there patterns indicating that particular models may be more appropriate for certain institutions?
   - To what extent does 'data publication' offer an additional opportunity for cost recovery?
6. Combined testing of the various models/scenarios/mechanisms developed in the four Data Publishing Working Groups
   - match the workflow reference models with cost compensation models
   - identify the different cost components and the potential cost recovery options



[9] http://www.ccdc.cam.ac.uk

Milestones

Before 18 January 2014: submission of RDA Case Statement.

- **M1:** 1 March 2014: launch questionnaire (continual reminders)
- **M2:** 26 March 2014, WG Launch Meeting at RDA Plenary in Dublin.

1 month - April 2014: start synthesis of cost models and survey of funder policies; initiate core case studies.

- **M3:** 30 June 2014: close questionnaire

6 months - end Sept 2014: preliminary analysis of questionnaire; reports on cost models and funder policies.

- **M4:** 22 September 2014 (RDA P4): present preliminary analysis of questionnaire; reports on cost models and funder policies

9 months - end December 2014: prepare report on questionnaire

- **M5:** 31 December 2014: Present report on questionnaire

12 months - end March 2015: (core) case studies;

- **M6:** 31 March 2015: present completed (core) case studies;

15 months - end June 2015: (additional) case studies and start combined testing within the four Data Publishing Working Groups

- **M7:** 30June 2015: Present completed (additional) case studies, start combined testing within the four Data Publishing Working Groups

18 months - end September 2015: prepare the final report and testing results

- **M8:** 30 September 2015 (RDA P5): Present final report and testing results

Project Management

The Working Group will have regular web meetings for working group members every six weeks.

Face-to-face meetings will be attached to the RDA Plenaries in Dublin in March 2014, in Amsterdam in September 2014, in March 2015 and in September 2015.

Other face-to-face meetings may be arranged if necessary, either for the whole group or for those working on specific work packages.

The Working Group leaders will participate in the regular meetings of the WDS/RDA Data Publishing Interest Group. These meeting will be used to keep the project on track, to monitor progress and resolve any differences of opinion.

Adoption Plan

Dissemination and Engagement

- Preliminary analysis of the questionnaire and the reports on cost models and funder policies will be disseminated at the RDA Plenary in Amsterdam in September 2014 and at SciDataCon in New Delhi in November 2014.
- The core case studies will be discussed at the RDA Plenary in Spring 2015.
- The final report will be delivered at the RDA Fall Plenary 2015.
- The publishers will be involved through the Data Working Group of STM, the ALPSP International Conference and the APE.

- The Working Group will also engage with generic meetings for data infrastructure providers, such as the WDS meetings, DataCite, etc.
- We will also target subject-focused conferences relating to the case studies: these are likely to cover social sciences and humanities (DANS), archaeology (ADS), crystallography (CCDC), life sciences/ecology (Dryad), European Geophysical Union/American Geophysical Union (WDCC).
- The initial work plan will focus on engagement with identified data centres, members of WDS, holders of the Data Seal of Approval, ICPSR members, and others. Should funding be available from other sources, the case studies will be expanded to include a larger number of data centres. In particular, a parallel activity would look at the cost recovery models for the European Research Infrastructure Consortia[10] (ERICs).

WG Members[11]

- Ingrid Dillo (DANS, NL) [**CHAIR**]
- Sarah Callaghan (BADC, UK)
- Simon Hodson (CODATA)
- Jared Lyle (ICPSR, US) tbc.
- Barbara Sierman (KB, NL)
- Frank Toussaint (DKRZ, Germany)
- Mark Thorley (NERC, UK, observer)
- Kim Finney (AAD, Australia)
- Anita de Waard (Elsevier, US)
- Eva Zanzerkia (NSF/GEO, US)
- Mikael Karstensen Elbæk (OpenAIREplus)

References

All of the working groups in the Publishing Data Interest Group have a common bibliography[12] in which publications relevant for this particular group are marked correspondingly.

[10] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric
[11] This list will be expanded, especially with representatives from data centers. There are a number of ideas of people who could be contacts (representatives of ADS, UKDA, ICPSR, CCDC, DCC, 4C).
[12] Bibliography: http://goo.gl/wA1G27

RDA-WDS Publishing Data Interest Group

Data Publishing Services Working Group Case Statement

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

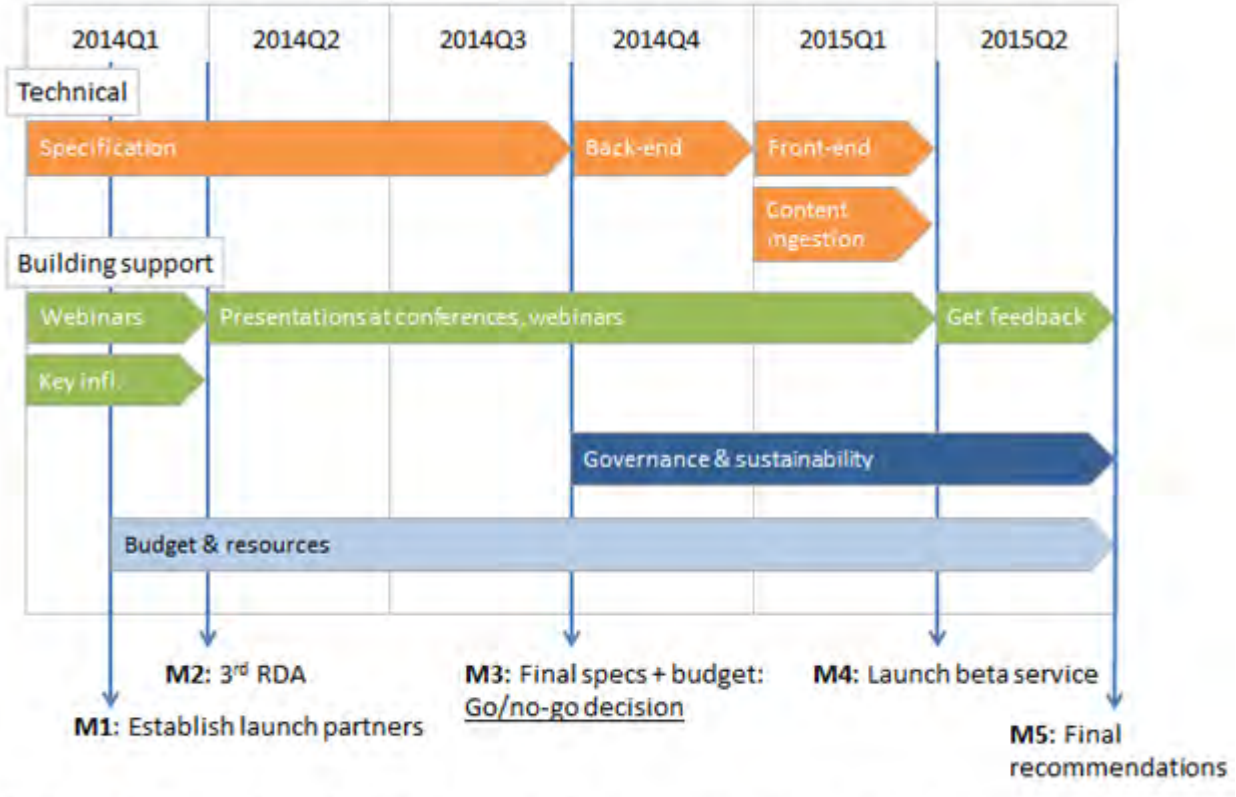
3. **Governance & sustainability.** This work stream will propose the long-term (beyond the beta phase) organizational structure to operate the cross-referencing service successfully and meets the needs of all stakeholder groups. This stream will also deliver a process through which new organizations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

RDA-WDS Publishing Data Interest Group

Workflows Working Group
Case Statement

Working Group Charter

Objectives

Researchers are increasingly encouraged or required to make their research data available for reuse but might often feel there are insufficient incentives for submitting and publishing data, resulting in low submission rates. Moreover, even when research data are preserved and submitted, it often happens with a bare minimum of metadata which inhibits reuse.

Why is this? There are established and/or emerging workflows for selected disciplines that enable the publishing of data and some provide credit via citation mechanisms. But in most disciplines researchers are simply not aware of such workflows and they may not be applicable without significant modification. Having information about workflows is therefore crucial for researchers—and the people/stakeholders supporting them—to understand the options available to practice open science. Workflows that enable persistence, quality control and access are all crucial to enhance the possibilities for greater discoverability as well as efficient and reliable reuse of research data.

The objectives of this Working Group are to provide an analysis of a representative range of existing and emerging workflows and standards for data publishing, including deposit and citation,

and provide reference models and implementations for application in new workflows. We will report on:

- Investigation and classification of current workflows for publishing data - including a brief gap analysis across disciplines for the identified use cases.
- Identification of a smaller set of reference models covering a range of such workflows to include:
    - when and where QA/QC and data peer-review fit into the publishing process (the broader subject of peer review itself is proposed as a future separate working group)
    - the role of researchers, institutions, data centers, publishers, funders, service providers and the wider community in the data publishing process
    - key barriers for identified use cases
- Selection of key use cases and organizations in which components of a reference model can be applied and promoted to the wider community, working closely with other WGs under the Publishing Data Interest Group.

We will build on the work of major past and current initiatives in which many of the working group members played leading roles. While these initiatives essentially focused on mature examples in particular in the Earth Sciences the work of this group will address a much more comprehensive and multi-disciplinary range of use cases, will classify workflow steps, components, and roles and eventually produce generic workflow models which towards the end of the project will be ready for testing and application in particular in the context of the ICSU World Data System and major science publishers.

Deliverables

- Provide a report summarizing the results of the investigation of current workflows including gap analysis
- A classification of a representative range of workflow models, in each case identifying the varying stakeholders and their different roles and responsibilities, to include where possible the likely associated resource and cost implications (working with relevant proposed RDA-WDS Costs of Publishing Data WG)
- Reference models summarizing key characteristics for each class of workflow
- Implementation of key components of a reference model to an existing use case(s) in order to illustrate the benefits to researchers and organizations of the reference model and the associated implications for the Working Groups on Costs, Publishing Services and Bibliometrics.

Value Proposition

Research communities and their institutions are considering—or in fewer cases are already implementing—workflows on their campuses or using platforms, such as discipline specific or national data centers, to allow their users to share and publish their research data. Many of them have to reinvent the wheel as there is no central resource or knowledge base to guide their efforts. Generic workflows, individual use cases or best practices for publishing data would aid them in establishing appropriate solutions that might include local systems enabling data deposit.

Research data are usually part of a network of scholarly objects, e.g. documentation, lab books or journal articles. It is expected that such clusters of information will continue to evolve and become more complex in the future as users expect to navigate seamlessly within them. They want to discover and access related information without major additional effort. This can only be facilitated by building a detailed understanding of the workflows and publishing outlets available right now. The main challenge will be identifying generic model elements while accounting for discipline specific features. We will cover different steps in the research lifecycle, as needed, e.g. from depositing data in repositories to dedicated data centers, data articles and journals. Identifying the steps in publishing workflows and who is responsible for various tasks can dispel some of the uneasiness for those encountering data publication for the first time as well as offer guidance across emerging and established tools being used in more advanced data sharing communities.

In classifying the current workflows, we will establish general models that allow for the individual imprints from various communities. The result will be reference models and components offering guidance for the wider community, from beginners to more advanced data publishers. This resource will be of use for any stakeholder group involved in publishing data. Repositories are often not aware of journal workflows and vice versa; understanding other parts of this complex endeavor helps each party see its role in the wider context. It is also useful in setting up mechanisms to link data and publications. Librarians have a role here in supporting researchers to find repositories to deposit in that are relevant both to their discipline and their publishing intentions, and offering guidance on the respective journal and repository workflows. The consortium of this working group comprises representatives of all these stakeholder groups to ensure the coverage of the wide range of use cases/best practices and viewpoints already emerging.

One important part of the work in the analysis of workflows will comprise the workflows for the usage of persistent identifiers, in particular Digital Object Identifiers (DOIs), which enable persistent links between digital objects, as well as accurate data citation. Data publication that enables data citation is a key incentive to make data accessible. Furthermore, such persistent identifiers allow an interoperable framework across platforms, publishers, repositories and others.

We plan to build on this in the second phase to test real implementation(s) of generic workflow model components in new scenarios. This offers a mutual benefit for both the provider who can test applicability and promote awareness of tools and for those implementing new workflows and who become able to offer their communities the benefits of publishing data.

Who will benefit

The main beneficiaries of the analysis and subsequent testing provided by this working group are the researchers and the main stakeholders involved in publishing and managing data, as well as in supporting scholarly communication. Better services and strategies for joint workflows will consequently influences the wider research communities. Discoverability and reuse of data will be enhanced, in particular through the unique collaboration between all relevant service groups participating in this group.

Authors will have clear channels for publishing data available to them and, crucially, will be able to derive credit from adhering to best practice in managing and sharing their data. Funders will be able to track the research data they have funded, measure its impact and guard it against repetition. Researchers will be able to work faster and achieve deeper insights outside their immediate subject domain. Librarians and data center experts become an integral part of the

ecosystem, e.g. through their expertise in cataloguing and metadata production and reference models for workflows are templates for ingest, QA, archiving and dissemination. Publishers and other service providers can use reference models for linking data with publications and provide innovative solutions to enhance access to and analysis of the published data. Workflows for data and metadata exchange between the stakeholders who hold it will help policy makers, funders and the public better ensure that the data underpinning published research is being made accessible cost-effectively. Policy makers and the public will be able to navigate the knowledge landscape with increased confidence in its veracity.

Impact

After the first phases, the identified use cases referenced to generic model elements where appropriate, will allow for a unique assessment of data publishing workflows today. This will directly inform and influence the work of all participating stakeholder groups, from repository providers to journal editors. The first steps enable an information exchange beyond the individual stakeholder groups and thus enable the adoption of best practices from other disciplines or joint workflows.

The proposed test implementation(s) of generic workflow model components in new scenarios benefit providers and users as explained previously and enables communities to meet key international and national government mandates to enable and incentivize data sharing for the benefit of all.

Working closely with the associated RDA-WDS Publishing Data Working Groups proposed on Bibliometrics, Costs and Publishing Services, we will identify the role and implementation of emerging metrics and impact/assessment tools in our test workflows and disseminate best practice. This will further the advancement of data aware incentive systems in research.

Engagement with existing work in the area

The WG builds on existing initiatives that have already contributed to a better understanding of workflows across disciplines or institutions for example. In addition, a number of use cases have been identified for the work planned in the forthcoming 1.5 years. Many of the use cases presented below are committed to this working group.

This WG will concentrate on workflows specifically, but linked to the other proposed WGs.
A general overview of relevant initiatives, projects, and platforms will be developed and maintained at the level of the RDA-WDS Publishing Data Interest Group, and may be found in the online survey[1].

In addition, we have been provided with agreements to offer materials by the following stakeholders and will seek further contributions via ICSU-WDS Members and RDA. It is envisioned to expand into a range of disciplines, including further partners in the Humanities and Social Sciences.

[1] Survey of related initiatives, projects platforms: http://goo.gl/0q2f8j

Ongoing initiatives addressing workflows for publishing data:

- The SCOR/IODE/MBLWHOI Library Data Publication Project has developed and executed projects related to two use cases: (1) data held by data centers are packaged and served in formats that can be cited and (2) data related to traditional journal articles are assigned persistent identifiers and stored in institutional repositories. The group has published the "Ocean Data Publication Cookbook"[2]
- The PREPARDE project has been on focused on the implementation of the Geoscience Data Journal (Wiley) workflow from author submission of datasets and papers, through technical and scientific review to publication, including specific data repository workflows at the British Atmospheric Data Centre (BADC) and US National Center for Atmospheric Research (NCAR). The latter incorporated ingestion of data, through data center technical review, to DOI assignment to dataset and bidirectional linking of data papers and datasets.
- KomFor[3] supplies a platform linking research and community based data facilities, libraries and journals. Funded by the German Science Foundation KomFor comprises a long standing consortium of ICSU World Data Centers and the TIB library in Germany collaborating to build sustainable and reliable ways for data publications in line with quality standards in scientific publishing. Part of the previous work of the consortium had been the implementation of DataCite[4].
- The ODIN[5] project studies workflows in two disciplines, the Humanities and Social Science and High Energy Physics, i.e. investigates commonalities and differences. The project has a particular focus on the implementation of persistent identifiers as an enabler in open science. The project is funded under the 7th Framework Program by the EC.

Use cases (some of them already committed to the WG):

- ICPSR (committed): The Inter-university Consortium for Political and Social Research (ICPSR), a repository of social and behavioral science research data established in 1962, has a documented workflow[6] that tracks data from deposit through curation to publication when the data become discoverable with DOIs and accessible on the ICPSR Web site.
- PANGAEA[7] (committed) is a multidisciplinary data center archiving, publishing and distributing geo-referenced data from earth system research. All data sets in PANGAEA are machine readable, citable, fully documented, and can be referenced via DOI. PANGAEA has established workflows for standalone data publications and for data supplementary to a science article. For this purpose cooperation with Elsevier and further science publishers has been built up. A cross-linking service allows to reference supplementary data in PANGAEA directly from the abstract pages of Science Direct or Scopus.
- The MBLWHOI Library (committed) is assigning Digital Object Identifiers (DOIs) to appropriate datasets deposited in the Institutional Repository (IR), Woods Hole Open Access Server (WHOAS)[8]. The Library has also developed a system that automates the ingestion of metadata and datasets from the NSF funded BCO-DMO into WHOAS and

[2] Ocean Data Publication Cookbook: http://www.iode.org/mg64
[3] http://www.komfor.net/
[4] DataCite: www.datacite.org
[5] ODIN - ORCID DataCite Interoperability Network: http://odin-project.eu/
[6] ICPSR workflow http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html
[7] PANGAEA - Data Publisher in the Earth & Environmental Science: www.pangaea.de
[8] WHOAS: http://darchive.mblwhoilibrary.org/

returns a DOI to the data management office. WHOAS has a service with Elsevier to link from ScienceDirect to datasets in the repository and is indexed by Thomson Reuters Data Citation Index.

- The Published Data Library (PDL)[9] is a project of the British Oceanographic Data Centre that provides snapshots of specially chosen datasets that are archived using rigorous version management. The publication process exposes a fixed copy of an object and then manages that copy in such a way that it may be located and referred to over an indefinite period of time.
- INSPIRE physics (candidate): INSPIRE[10] is the digital library serving the global community of High-Energy Physics (HEP).Today, datasets on INSPIRE are treated as independent scholarly records and are assigned a DOI to facilitate their citation. For INSPIRE the next step includes data citations metrics, as well as stronger collaboration with publishers to identify, preserve and display datasets associated with publications.
- Geoscience Data Journal[11], Wiley (committed): Datasets published in this dedicated data journal undergo a peer review process. They are deposited in approved data centers, while being described in short data papers that give details on for example their collection or processing software that was use.
- PENSOFT biodiversity data journal[12] (committed): Community peer-reviewed, open-access, online platform for publishing, dissemination and sharing of biodiversity-related data of any kind. All structural elements of the articles – text, morphological descriptions, occurrences, data tables, etc. – are treated and stored in accordance with the Data Publishing Policies and Guidelines of Pensoft Publishers.[13]
- F1000Research[14] (committed): Life sciences journal - publishes all article types but ensures all articles include underlying data and software where relevant. Uses rapid publication followed by invited but completely transparent post-publication peer review.
- Publishing data in crystallography[15] (candidate, contact: Brian McMahon): Crystallography is a data-rich, software-intensive scientific discipline with a community that has undertaken direct responsibility for publishing its own scientific journals. That community has worked actively to develop information exchange standards allowing readers of structure reports to access directly, and interact with, the scientific content of the articles.
- EBI Genomics + Europe PubMedCentral, UK (candidate, contact: Jo McEntyre): The field of molecular biology has a long tradition of data sharing going back to the open data principles established with the Human Genome Project. Today, this is is a prominent discipline to observe data deluge. At EBI many databases to deposit data are developed and maintained. Researchers identify such datasets in publications by referencing the ID (e.g. accession numbers or DOIs).
- NPG's Scientific Data[16] (candidate - contacts: Ruth Wilson, Susana Sansone): This dedicated data journal published article types, called "Data Descriptors" which are quality assured through peer review process. The Datasets described are deposited in external, community approved repositories. Alongside the narrative Data Descriptor articles ISA-tab

[9] PDL: https://www.bodc.ac.uk/data/published_data_library/
[10] INSPIRE: www.inspirehep.net
[11] http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060
[12] http://biodiversitydatajournal.com/
[13] http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf
[14] http://f1000research.com/
[15] Publishing data in crystallography: http://dx.doi.org/10.1186/1758-2946-4-19
[16] Nature Scientific Data: http://www.nature.com/scientificdata/

metadata files are produced for each Data Descriptor to provide a machine readable format.

- Earth System Science Data Journal - ESSD[17] (candidate, contacts: Dave Carlson, Hans Pfeiffenberger): This OA journal focuses on the publication of datasets in the Earth Science. The publication process includes an open peer review process. Datasets are submitted to external approved data centers.
- GBIF data publishing[18] (candidate, contact: Vishwas Shavan): GBIF offers a workflow and a toolkit[19] for publishing biodiversity data comprising species occurrences data, species checklists, and corresponding metadata.
- Digital Curation Centre[20] (committed): Centre for advice on research data management to UK universities, including generic advice on support for data selection, deposit and publishing.
- Harvard Dataverse[21] (committed): includes the world's largest collection of social science research data. It supports a full data publishing workflow that can be reviewed and evaluated with a deposit API that integrates with journals to seamlessly deposit data to the research data repository as part of the article publishing workflow.

Work and Adoption Plan

The following tasks will be carried out in close cooperation with the parallel RDA-WDS Working Groups. The work will result in a consecutive set of reports, each of them open to external feedback.

The work is currently envisioned for a timeline of 1.5 years. The success of this working group depends on a close collaboration of all stakeholder groups from data centers to publishers to research community representatives and their institutions etc. We will pursue a wider dissemination and engagement of external initiatives that might emerge over the course of this work. This approach shall ensure an extensive coverage in terms of model components as well as use cases. We then intend to implement one or more model components in suitable use cases which will also require a strong and open engagement with the wider community.

Based on this approach, the work plan is split into four consecutive phases reaching out to mid-2015 on an 18-month timescale. However, given the practical implementation that is planned for the fourth phase in this working group, it is to be expected that work will continue far beyond that time horizon.

PHASE I (end March 2014): Understanding the current state

- Identify a representative range of workflows for publishing data across disciplines: analyze and consolidate them into several broad workflow classes
- Compose questions for use case representatives comparing and contrasting individually selected workflows against a broad workflow class
- Initial discussion of questions at IDCC14 workshop (end Feb 2014), coordination of survey questions with other WGs in the IG Publishing Data

[17] ESSD: http://www.earth-system-science-data.net/
[18] GBIF data publishing: http://www.gbif.org/publishingdata/summary
[19] http://www.gbif.org/ipt
[20] UK Digital Curation Centre http://www.dcc.ac.uk/resources/how-guides/cite-datasets
[21] Harvard Dataverse http://thedata.harvard.edu/

- Present initial synthesis of responses for presentation to RDA Plenary (end March 2014) [**Milestone 1**]
- Identification of groups, workshops and conferences to engage with the wider community and to establish further partnerships, i.e. with disciplinary interest groups.

PHASE II (end of Q3, Sep 2014): Qualifying/Classifying the current state

- Classification and documentation of workflows
    - Define the nature and function of "components" and "workflows" in data publishing
    - identify associated resource and cost implications aligned with proposed Costs WG
    - Identify the varying stakeholders and their different roles and responsibilities: Researchers, Data Centers, Institutions, Libraries, Publishers, Funders & Service Providers
    - Identify where/when/how quality assurance and quality control and data peer-review takes place
    - Identify where/when/how research publishers and journals participate in the data publication process
    - Inclusion of domain coverage, legal and ethical aspects
- Preparation of a draft gap analysis of use cases (e.g. the coverage of components in an individual discipline). [**Milestone 2**]
    - Which components are commonly used, what is missing?
- Wider dissemination and feedback
    - Present draft gap analysis and classification of components: invite additional workflows through RDA community (and discipline specific outlets) and related feedback to refine the models
    - For presentation at RDA Plenary IV (end Sep 2014) and other selected workshops, webinars

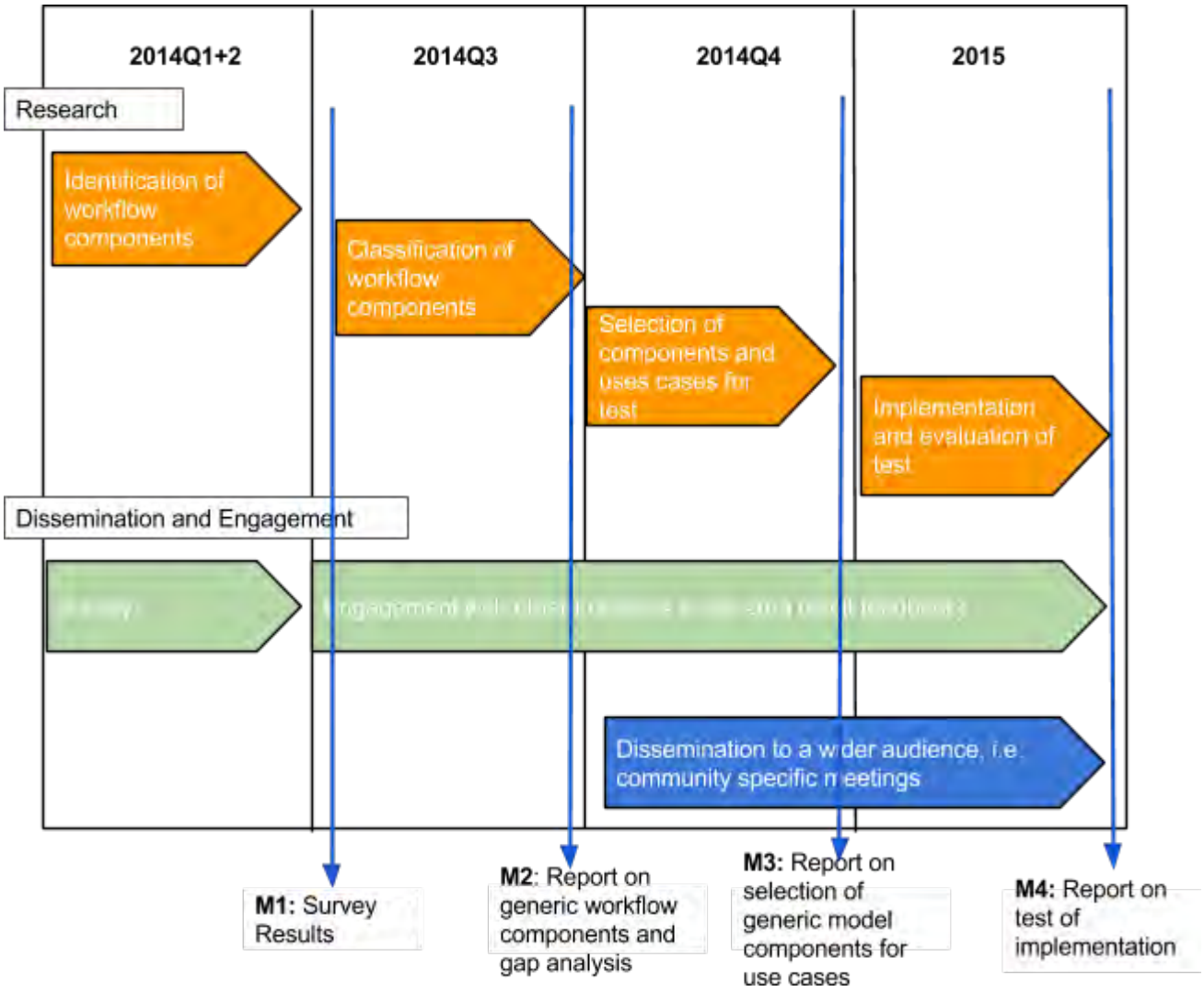
PHASE III (end of Q4, Dec 2014): Final analysis and use case selection for test implementations

- Preparation of final report on workflows, model components and gap analysis which includes feedback from the wider community [**Milestone 3**]
- Based on gap analysis identify and select use case(s) where reference model component(s) may be applied and tested/implemented
    - As part of this: selection of model components in the reference models which may be implemented in a test program for identified use case(s)
- Wide dissemination of results, i.e. also to community or discipline specific groups (possibly also depending on potential test environments)

PHASE IV (end of June 2015): Partnering to fill the gaps - test programme

- Conception of a program with all stakeholders to test one or more generic workflow component(s) in specific use case(s)
- Implement, evaluate and analyze benefits/challenges for use cases (end June 2015)
    - Select a group of potential partners/stakeholders for the implementation of test components in a use case. Questions to consider: What was missing in this use case before? How could the potential partners and model components facilitate data publishing in this institute/discipline/…?

    - Evaluation: how did the test implementation change data publishing in this use case and what were the benefits, including evidence and any available metrics?
- Dissemination within and beyond RDA/WDS



Milestones and intermediate products

- **M1** Survey Results
- **M2** Draft report on reference models, i.e. model components. Special attention is given to a gap analysis of current practices.
- **M3** Final report on components and identification of generic workflow components for test use case(s)
- **M4** Report on test implementation of components in selected use cases: suitability, pros and cons, as well as benefits for identified use case(s)

Project Management

Mode and frequency of operation

The RDA-WDS Working Groups hold face-to-face meetings at RDA Plenary Meetings, teleconference meetings every 6 weeks and builds on regular email communications.

Consensus, conflicts, staying on track and within scope, and moving forward

The RDA-WDS Publishing Data Interest Group will ensure coordination of the Working Groups through regular teleconference meetings of the Chairs of the Working Groups (every 6 weeks), mailing lists and through member involvement in other associated activities, i.e. discipline specific working groups which shall be of special relevance for this topic. The latter will be done through members of the working group as well as further partners who are identified during the work in the forthcoming 1.5 years.

Planned approach to broader community engagement and participation

This working group relies on a strong collaboration of different stakeholder groups. Therefore the engagement within the group and beyond is crucial for its success and will happen through the standard RDA and WDS channels (mailing lists, face-to-face meetings, webinars). In addition, it is envisioned to target community specific conferences, workshops and webinars to engage a broad spectrum of interested parties. This is particularly important when it comes to the discussion of the draft report, as well as the preparation of the test environments. The reports and any other outcome of the working group will be disseminated openly for future reuse/reference.

Membership

- Jonathan Tedds (UK, University of Leicester) [**CO-CHAIR**]
- Suenje Dallmeier-Tiessen (Switzerland, CERN) [**CO-CHAIR**]
- Merce Crosas (US, Harvard University)
- Michael Diepenbroek (PANGAEA)
- Kim Finney (Australia, AADC)
- John Helly (US, UCSD)
- Hylke Koers (The Netherlands, Elsevier)
- Rebecca Lawrence (UK, F1000 Research Ltd.)
- Fiona Murphy (UK, Wiley-Blackwell)
- Amy Nurnberger (Columbia University Libraries)
- Lisa Raymond (US, Library Woods Hole Oceanographic Institution)
- Johanna Schwarz (Germany, Springer)
- Mary Vardigan (US, ICPSR)
- Ruth Wilson (UK, Nature)
- Eva Zanzerkia (US, NSF)
- Angus Whyte (UK, DCC)

References

All of the working groups in the Data Publication Interest Group have a common bibliography[22] in which publications relevant for this particular group are marked correspondingly.

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