RDA-WDS Publishing Data Interest Group

Bibliometrics Working Group

Case Statement

Working Group Charter
- Objectives
- Deliverables

Value Proposition
- Who will benefit
- Impact

Engagement with existing work in the area

Work Plan

Adoption Plan
- Deliverables of the WG
- Milestones and intermediate products

Project Management
- Mode and frequency of operation
- Consensus, conflicts, staying on track and within scope, and moving forward
- Planned approach to broader community engagement and participation

Membership

References

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. <u>Organizational</u>: 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

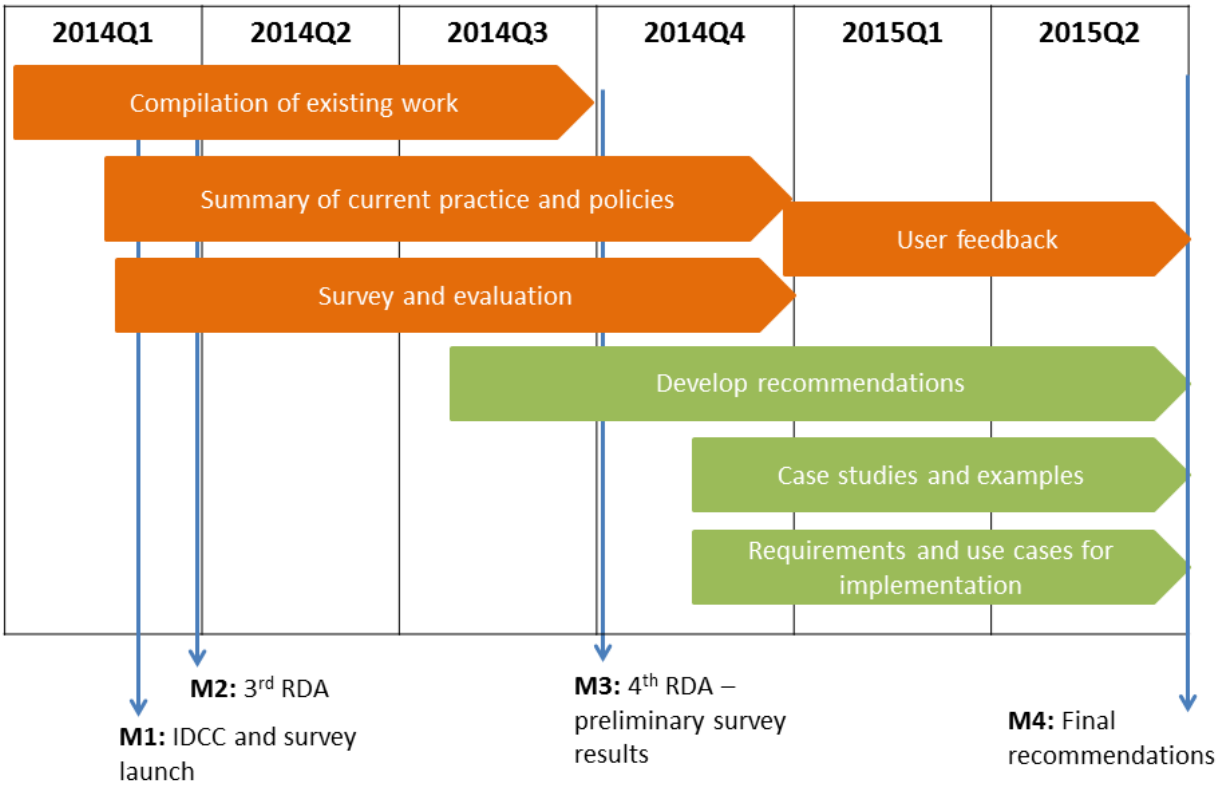
1. Every 6 weeks: Teleconference between Co-chairs of the Publishing Data Interest Group and Chairs of the Working Group
2. Open Webinars for dissemination and exchange
3. Face to face meetings during RDA plenary meetings

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