Research Data enters Scholarly Communication

Chaired by Mustapha Mokrane

Tuesday, 30 October 2012, 8:30–10:00

Publishing data has an overall positive impact on the quality and availability of scientific data. It is essential for the transparency of scientific work and provides the necessary incentives for data producers. Data publication services need to be integrated into the traditional science publication process. This requires a coherent system consisting of libraries, science publishers, and certified data repositories as, for example, supplied by the new ICSU World Data System (WDS). This session will address the background and initiatives related to data publishing, in particular, cross-linking journals and data archives, data citation and registration, peer review, and impact on citation rates.

A. Overview talk:

Michael Diepenbroek, PANGEA: Data Publisher for Earth & Environmental Science.

B. Data-rich publishing – the editors' view:

David Carlson, Editor in-chief of ESSD Journal.

C. Data-rich publishing – the publishers' view:

IJsbrand Jan Aalbersberg, STM Journals/Elsevier.

Fiona Murphy, Willey Blackwell.

D. Data-rich publishing – the data centres' view:

Wim Hugo, South African Environmental Observatory Network.

Presentations might address one or more of the following topics:

- Data publishing
- Linking scientific data and literature
- Journals and data centres

Abstracts:

Lessons in Data Publication from the journal Earth System Science Data

David Carlson and Hans Pfeiffenberger

Started during the International Polar Year 2007–2008, the Earth System Science Data journal (Copernicus) has gradually built a substantial collection of useful and relevant open access datasets, and a corresponding network of open access data centres. The ESSD publication opportunity attracts datasets from discrete events or projects, as well as systematic global-scale data compilations, and following a discriminating review of methods and quality, returns formal accreditation for the substantial data preparation efforts and enhanced visibility for the data products. Often, ESSD supports joint special issues with science journals, to publish datasets side-by-side with associated science papers. Because the journal intersects with the evolving worlds of data access and of scholarly publication, our experience offers valuable and continuing lessons for data providers, reviewers, data centres, publishers, and editors. The ESSD published products—accessible documented high-quality datasets—represent a very positive step forward for the data and science communities.

Research Data as integral part of Scholarly Publishing

Michael Diepenbroek

Data-intensive science is building on data sharing and integration of data from various sources. This—besides efficient usability of data—requires first and foremost a new transparency of scientific work. Recent surveys have shown the willingness of researchers to share not only their data but also their concerns. This clearly indicates that a cultural change in science is needed. This change is fostered through the idea of data publishing as a clear incentive for scientists to share their data. Only in the past few years have scientists began calling for data “citation” and referring to data “publication” rather than data “sharing” and “availability”. Data publication can be similar to the conventional publication of articles in journals that includes online submission, quality checks, peer-review, editorial decisions, and an equivalent of “page proofs”. In fact, data storage in authoritative open access databases is becoming increasingly important or is already mandatory for the acceptance of peer-reviewed publications in specific research fields such as, for example, molecular sciences or ecology. Data publication services need to be integrated into the traditional science publication process. This requires collaborations with science publishers and services to link articles and data on the editorial and for discovery of data. The impact on citation rates could be shown in bibliometric studies on science articles having supplementary data. Eventually, an index for science data would be needed similar to the Data Usage Index (DUI) or the newly propagated index for scientific data by Thomson Reuters. Data publishing is strongly supported by ICSU-WD, the Global Biodiversity Information Facility (GBIF), the European Commission, and further stakeholders.

Data Centre View: Data Publication

Wim Hugo

In recent years, the concept of data publication as a valid, measurable output of scientific endeavour has solidified to the point where it is clear that (1) it will provide heretofore unavailable incentives to data-focused scientists for peer recognition, (2) it will lead to increased availability of data, including quality assured data, (3) society in general and science as a whole will benefit from the compound effect of such availability, and (4) that it is now a question of when it will happen, and not if it will happen.

Data Centres will play a critical role in this new infrastructure in support of science. To some extent ISCU-WDS is pioneering the creation of such infrastructure since, based on its criteria for membership, the wider context of what constitutes viable data centres has been defined. In broad terms, it addresses technical, scientific, governance and financial feasibility for the longer term, underpinned by standards and policies that limit risk and engender trust.

To be effective, data curated by such centres should be

- Discoverable;
- Capable of being understood in respect of scope, quality, and usability, even if the data sets are large;
- Preserved and made available for the long term;
- Standardised in view of use by both humans and systems.

Data Centres face several challenges in providing the fabric of permanence and reliability that is required to make data publication a success. These include (1) funding for the indefinite[1] preservation of data; (2) universal access to prospective data providers, irrespective of data quality; (3) interoperability to the point of seamless integration with the journal publication industry; and (4) lack of capacity, know-how, and incentives amongst the producers and providers of data.

Positive trends need to be reinforced to support Data Centres in this endeavour:

1. Support for the provision of ring-fenced, grant-linked funding for the preservation of data;
2. Policies that support the publication of grant- and tax-funded datasets;
3. Inclusion of data management and informatics training as part of honours-level degree study;
4. Entrenchment of standardized data publication metrics and data management plans in proposal assessment processes;

[1] Which, for now, is still to be defined …

5. Due recognition in science and publication rankings.

We foresee some potentially negative outcomes from a more formalised data publication environment, including (1) an initial flood of low-quality submissions, (2) competition between Data Centres for funding of long-term preservation, and (3) competition from the established publication industry for what estimates show could be a doubling in size of the market.

Research questions and best practice development should at least include (1) scientific basis for the decision to terminate or alter the on-line availability of a dataset, (2) the extent to which derived works are unique, (3) minimum data and visualization services required to support interoperability with published articles and reports, and (4) integration of citation, quality, and traditional metadata records for a more comprehensive view of the characteristics of a dataset.

Integrating Research Data in the Article of the Future

IJsbrand Jan Aalbersberg

Driven by requests from institutions and funding bodies, the scientific community increasingly makes its research data available on the public internet. However, the awareness of that data, of its exact location, and of its relation to the enveloping research is still lagging behind—especially when compared to the amount of data being produced and already available. This is a problem that all parties responsible for research data and the communication of research are aware of and are trying to address. For instance, research data repositories improve the visibility of, and accessibility to, their data, and research data organizations like DataCite support the process of data registration and persistent linking through the assignment of DOI's.

As the scientific research article is a well-established and highly visible form of communication between researchers, the scientific publisher Elsevier is contributing to a solution to the data awareness problem through offering a variety of connections between the online scholarly article and the associated research data. In Elsevier's Article of the Future format, these connections help to increase the visibility of research data and add valuable context to the data to facilitate interpretation and reuse. Depending on the linking capabilities of the data repository hosting the research data, Elsevier's journal articles can link to datasets, entities in those articles can deep link to data elements, and data can even be integrated in research articles using discipline- and data-specific tools and applications.

The Geoscience Data Journal: A Locus for Stakeholder Collaboration in Data Publishing

Fiona Murphy, Sarah Callaghan, and Rob Allan

Due to multiple pressures—including funding resources, technical developments, and policy priorities—the scholarly communication ecosystem is changing. This is inevitably causing pressures to bear on not only researchers themselves but also their fellow stakeholders: funders, publishers, data centres, and libraries. Innovations in instrumentation, collection methods, software, and modelling are contributing to output, but we are not necessarily developing equal proficiency in managing said outputs. Accordingly, research data and datasets—their collection, curation, storage, citation, and re-use—have been identified as crucial to the progression of useful knowledge, so they need to be satisfactorily integrated into the research canon.

In addition, there are issues with the actual sharing of data. For example, the provenance of the dataset (and often the dataset itself) can be "changed" as it passes from one owner to another, thereby reducing any chances of using that data to test the reproducibility of results originally made from it. Also, the present mechanism for academic recognition revolves around the production and publication of peer-reviewed papers. The production of high-quality datasets takes time and effort, and is often insufficiently recognised as an activity worthy of prestige, even though the papers that result from that dataset may be considered of significant scientific importance. Simple sharing of data is unlikely to provide the data creators with the academic recognition they deserve.

This presentation examines the new title Geoscience Data Journal as a particularly relevant case study for this audience—in terms of subject matter and partnerships—and the on-going related research project PREPARDE (Peer REview for Publication & Accreditation of Research Data in the Earth sciences). The latter aims to develop the required workflows, guidelines, policies, and educative support to encourage and optimize best practice in all aspects of the research data life-cycle. These aspects include deposition in an accredited repository, peer review, publication, and the encouragement to reuse. Feedback from the audience on these aims would be most welcome as engagement and involvement with stakeholders are key PREPARDE objectives.