The Mendeley Data Management Platform: Research Data Management From A Publisher's Perspective

Anita de Waard, a.dewaard@elsevier.com

Introduction

In its roles as both a publisher and an information analytics company, Elsevier is committed to contributing to ecosystems that support rigorous and reproducible research data creation throughout the research workflow, making data accessible, discoverable and reusable. Moving beyond content delivery, the overall goal of our Research Data[1] Management efforts is to develop knowledge management tools and solutions to support researchers, librarians, universities and government agencies in managing, sharing and measuring a rich variety of research outputs. Throughout our publishing and analytics tools and workflows, we aim to support researchers with tools and practices to improve exposure and access to their full research cycle, to augment the quality, transparency and impact of research and scholarship. This position paper describes some research underlying our current research and development efforts, and touches on the principles and practice of our tool suite.

1. Elsevier Research Data Principles

Elsevier actively participates in a number of community efforts, industry initiatives and standards and policy bodies to support more effective discovery, use and reuse of research data, including:

- Co-authorship of the Brussels Declaration[1], in collaboration with the STM Association[2];
- Cofounding, codevelopment and full implementation of[3] the ORCID[4] researcher ID System;
- Strong support for and co-authorship of the FAIR Data Principles[5], stating that data should be Findable, Accessible, Interoperable and Reusable;
- Co-chairmanship of and participation in Research Data Alliance[6] working groups (in particular the Data Publishing IG[7], the Data Discovery IG[8], the Metadata IG[9] and the Data Citation WG[10]),
- A fostering engagement with the Scholix initiative[11], a linked open database to link datasets to articles, which was recently fully integrated into[12] Scopus;
- A founding membership of Force11[13] which include co-authorship of the Force11 Manifesto[14] and implementation of the Force11 Data Citation Principles across our publishing program[15];
- A partnership on data archiving[16] with DANS, the Data Archiving and Network Services[17] group of the Dutch Royal Academy for long-tern preservation of data outputs in Mendeley Data;
- A partnership with DataCite[18], an international non-profit that focuses on data citation practices, to mint data DOI's for records in Mendeley Data.

Participation in these community efforts has lead us to define a number of principles[19] with regards to access to research data, the most of which is that ***research data should be made available free of charge*** to all researchers wherever possible, with ***minimal reuse restrictions***. Although expectations and practices around research data vary between disciplines and discipline-specific requirements need to be taken into account, in general we advocate and practice that ***researchers should remain in control*** of how and when their research data is accessed and used, and should be ***recognized and valued*** for the investments they make in creating and sharing their research data.

---

[1] Throughout this document, we define 'Research Data' as: the results of observations or experimentation that validate research findings, which often underlies, but exists outside of research articles; this includes raw data, processed data, software, algorithms, protocols, methods, and materials that are not a part of the published journal article.

Enabling effective reuse of research data is a shared aim, and all stakeholders (including publishers, libraries, funders, data repositories and other involved in the research data ecosystem) should ***work together to find efficiencies*** and avoid duplication of effort. Many different platforms, publications, tools and curation services can enhance research data by improving their discoverability, use, reuse, and citation: therefore, ***working as openly as possible*** **and** ***optimizing integration*** with workflow tools and processes developed by other parties is one of the key driving forces behind our development work. Lastly, where parties add value and/or incur significant cost or effort in enhancing research data to enable its reuse (by curation or other efforts), these ***contributions need to be recognized and valued***.

2. Open Data Report

To map the current landscape of data sharing practices, we conducted a year-long large-scale study, together with the CWTS (Centre for Science and Technology Studies) at the University of Leiden, available online[20], see also all data developed in the course of this work, at Mendeley Data[21]. Using a multi-faceted approach, we analyzed acknowledgement sections in scientific articles, conducted a large-scale (1200 people) survey on data sharing practices, and conducted three case studies in Soil Science, Human Genetics, and Digital Humanities.

From the bibliometric analysis, we found that, although the number of citations to data journals is growing, they are still a small portion of the overall publications, and their adoption is quite domain-specific. There is a lack of standards regarding data citation in regular journal articles, making it difficult to assess how widely data is shared and used across domains from the acknowledgement sections: addressing this could support improved data sharing, and therefore improving the perceived lack of credit for sharing data.

The data sharing survey (with 1167 respondents) showed that, although 69% of respondents found that sharing data was very important in their field and 73% wanted to have access to other people's data, only 37% believed there was credit attached to doing so, and only 25% felt they had adequate training to properly share their data with others. The main barriers for sharing data were privacy concerns, ethical issues, and intellectual property rights. Mandates from publishers or funding agencies were largely not seen as a driving force, and this combination of lack of training in how to share data, concern regarding reuse and privacy, as well as a lack of urgency in terms of mandates, drives the gap between desire and practice concerning data sharing.

In the case studies for three domains, we further found that national and regional differences in data sharing practice hampered widespread sharing and reuse because laws and customs differ in regions and countries. The good news is that in collaborative research projects naturally enable and enhance data sharing and storing practices, because of their distributed nature. In some fields, data sharing practices are engrained within the research practice: an example is Digital Humanities where e.g. sharing code through Github is endemic, and easily translates to data sharing. In other fields, such as human genetics, the fact that raw (sample) data and processed (analysed) data were used by different individuals at different moments made for an 'endemic' data sharing structure, which can be used to scale up sharing and publishing practice.

In summary, we see that data sharing is very much a practice in flux: there is a perceived need for better ways to share more data, but still a lack of standards, drivers, and training to do so. Addressing these issues is the main goal of our RDM Department, this report and other like it drive our community participation and the development of our tool suite.

3. Mendeley Data Management Suite

To address these matters while adhering to the data sharing principles outlined above, we are developing a suite of tools supporting the Research Data Management process throughout the research lifecycle. These tools are integrated into a common platform, the Mendeley Data Management Suite, which offers open APIs for integration with external standards and tools, as well as components developed by our partners at academic institutions and funding agencies.

See Figure 1 for the overall view of this tool suite, which consists of these components:

- Mendeley Data[22], a storage and preservation option for research data, which is freely available for researchers to store, collaborate, find, share and expose research data and annotations. Through the submission interface and user interface, datasets can be directly linked to articles or methods/protocols. Data can be embargo'ed and kept private for researchers or within a department, or shared openly and through a collaboration with DataCite[23], DOI's are assigned to each version of a dataset. On June 22, it was announced that Mendeley Data's open research data repository achieved the Data Seal of Approval certification[24]; which confirms that the repository complies with 16 guidelines, and is therefore a trusted digital repository.
- HiveBench[25], an online electronic lab notebook that supports protocol and data capture and improves rigor, reproducibility, and researcher efficiency. It allows for local or remote storage and offers an online WYSIWIG interface for authoring and editing protocols and outcomes.
- DataSearch[26], an open data search engine across a multitude of domain-specific and cross-domain repositories. In querying and enabling data previews of different data types and formats across heterogeneous outputs it supports making data 'Findable', as endorsed by FAIR.
- Research Elements[27], A collection of Open Access journals that make data, software, materials and methods available for discovery, reuse and citation. The journals are peer reviewed to increase the credibility of key research outputs normally excluded from traditional journals, making them more controllable and replicable.
- Data Lighthouse[28], a tool to support librarians in tracking RDM outputs and metrics in support of Data Management Plan compliance. Using Scopus searches, the service identifies papers that match to institutional outputs: then, using Scholix[29], associated datasets are identified.

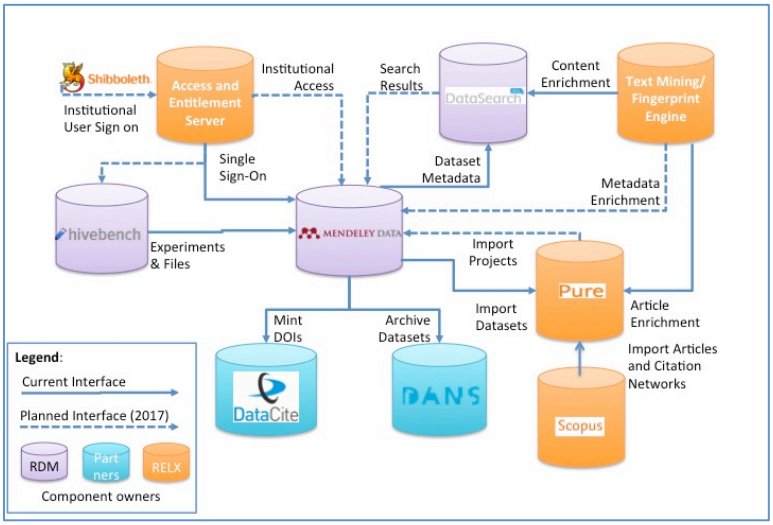

Figure 1: Integration between various components of the RDM Data Management Suite and various external standards.

4. Future Outlook

We are continuing to work with partners in academia, industry and various government agencies to develop tools and standards that support proper data sharing practices and enhance researchers' efforts to create, share and access optimally usable research data, methods and software components.

Since a key driver for researchers to share their data is the ability to receive credit for doing so, one of our current focal areas is to contribute to ongoing efforts to develop a set of practical and easily implementable *data metrics*. In collaboration with the PlumX Tool Suite[30], Project COUNTER[31], the CItescore[32] Article Level Metrics team, and the Making Data Count[33] effort, are working on developing and sharing a 'Basket of Data Metrics', consisting of 10 different data-related metrics (see Table 1). Together with our academic development partners, we are exploring how we showcase these metrics today, and how these metrics might work in an institutional environment.

Goal:	**Metric:**
More data is **properly preserved**:	
1 **Stored,** i.e. safely available in sustainable repository	Nr of datasets stored in long-term storage
2. **Published**, i.e. long-term preserved, accessible via web, have a GUID, are citeable, with proper metadata	Nr of datasets published
3. **Linked**, to articles or other datasets	Nr of datasets linked to articles
4. **Validated**, by a reviewer/curated	Nr of datasets in curated databases/peer reviewed in data articles
More data is **seen and used**:	
5. **Discovered:** found by users	Nr of datasets viewed in databases/websites/search engines
6. **Identified:** Resolved through a Doi Broker	DOI is resolved
7. **Mentioned:** Social media and news	Social media and news mentions
8. **Cited:** Formal citations of data	Nr of datasets cited in articles
9. **Downloaded:** Distinct downloads	Downloaded from repositories
10. Reused: Dataset is used for new research	Mention of usage in article or other dataset

Table 1: Basket of Data Metrics

To develop these and other ideas in order to improve interconnected systems of research data management, Elsevier continues to actively seek collaborations with libraries, researchers, data centers and government agencies across the globe. We are actively seeking engagements in this domain, and through research collaborations and development, hope to continue to contribute to the development of distributed, robust and sustainable research data management ecosystems.

References:

[1] http://www.stm-assoc.org/public-affairs/resources/brussels-declaration/
[2] http://www.stm-assoc.org/about-stm/about-the-association/
[3] https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-joins-orcid-in-announcing-launch-of-orcid-registry
[4] https://orcid.org/
[5] https://www.force11.org/about/manifesto
[6] https://www.rd-alliance.org/
[7] https://www.rd-alliance.org/group/rdawds-publishing-data-workflows-wg/post/fwd-rda-wds-data-pub-workflows-data-pub-workflows
[8] https://rd-alliance.org/groups/new-paradigms-data-discovery-ig
[9] https://www.rd-alliance.org/groups/metadata-ig.html
[10] https://www.rd-alliance.org/groups/data-citation-wg.html
[11] http://scholix.org
[12] https://blog.scopus.com/posts/new-on-scopus-link-to-datasets-search-funding-acknowledgements-and-find-more-citescore
[13] https://www.force11.org/
[14] https://www.force11.org/about/manifesto
[15] https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-implements-data-citation-standards-to-encourage-authors-to-share-research-data
[16] https://www.knaw.nl/en/news/news/collaboration-dans-and-mendeley-on-archiving-datasets
[17] https://dans.knaw.nl/en
[18] https://www.datacite.org/
[19] https://www.elsevier.com/about/our-business/policies/research-data
[20] https://www.elsevier.com/about/open-science/research-data/open-data-report
[21] Berghmans, Stephane; Cousijn, Helena; Deakin, Gemma; Meijer, Ingeborg; Mulligan, Adrian; Plume, Andrew; de Rijcke, Sarah; Rushforth, Alex; Tatum, Clifford; van Leeuwen, Thed; Waltman, Ludo (2017), "Open Data: the researcher perspective - survey and case studies", Mendeley Data, v1 (http://dx.doi.org/10.17632/bwrnfb4bvh.1).
[22] http://data.mendeley.com
[23] https://www.datacite.org/
[24] https://assessment.datasealofapproval.org/assessment_244/seal/html/
[25] https://www.hivebench.com/
[26] https://datasearch.elsevier.com
[27] https://www.elsevier.com/authors/author-services/research-elements
[28] https://osf.io/nst8d/
[29] http://scholix.org
[30] http://plumanalytics.com/
[31] https://www.projectcounter.org/
[32] https://www.elsevier.com/editors-update/story/journal-metrics/citescore-a-new-metric-to-help-you-choose-the-right-journal
[33] http://mdc.lagotto.io/