Final Manuscript for submission to the

International Journal of Digital Libraries, Special edition: Research Data Publishing

We would be very grateful if you would fill in this short survey on your evaluation of the manuscript. It will take about 3 minutes to complete the survey and you don't need a SurveyMonkey account to do it (just close the annoying pop-up window if there is one: https://www.surveymonkey.com/r/workflows_for_data_publication

Workflows for Research Data Publishing: Models and Key Components

Authors: Theodora Bloom[1], Sünje Dallmeier-Tiessen[2*§], Fiona Murphy[3*], Claire C. Austin[4,5], Angus Whyte[6], Jonathan Tedds[7], Amy Nurnberger[8], Lisa Raymond[9], Martina Stockhause[10], Mary Vardigan[11]

***Lead authors**

[§]**Corresponding author:** sunje.dallmeier-tiessen@cern.ch

Contributors: Tim Clarke[12], Eleni Castro[13], Elizabeth Newbold[14], Samuel Moore[15], Brian Hole[15]

Author affiliations: [1]BMJ, [2]CERN, [3]Wiley, [4]Environment Canada, [5]Research Data Canada, [6]Digital Curation Centre - Edinburgh, [7]University of Leicester, [8]Columbia University, [9]Woods Hole Oceanographic Institution, [10]German Climate Computing Centre (DKRZ) , [11]University of Michigan/ICPSR

Contributor affiliations: [12]Massachusetts General Hospital/Harvard Medical School, [13]Harvard University/IQSS, [14]The British Library, [15]Ubiquity Press

Author statement: All authors[1] affirm that they have no undeclared conflicts of interest. Opinions expressed in this paper are those of the authors and do not necessarily reflect the policies of the organizations with which they are affiliated.

ABSTRACT

Data publishing is a major cornerstone of open science, reliable research, and modern scholarly communication. It enables researchers to share their materials via dedicated workflows, services and infrastructures and ultimately is intended to ensure that data – and in particular datasets underlying published results — are well documented, curated, persistent, interoperable, reusable, citable, attributable, quality assured and

[1] Theodora Bloom is a member of the Board of Dryad Digital Repository, and works for BMJ, which publishes medical research and has policies around data sharing.

discoverable. Needless to say, data publishing workflows potentially have an enormous impact on researchers, research practices and publishing paradigms, as well as on funding strategies, and career and research evaluations.

It is crucial for all stakeholders to understand the options for data publishing workflows and to be aware of emerging standards and best practices. To that end, the RDA-WDS Data Publishing Workflows group set out to survey the current data publishing workflow landscape across disciplines while at the same time paying attention to discipline-specific characteristics. We looked at a diverse set of workflows, including basic self-publishing services, institutional data repositories, curated data repositories, and joint data journal and repository arrangements to identify common components and standard practices. This permitted us to identify, analyze and categorize the main building blocks comprising data publishing workflows. We wanted to understand how workflows differ based on the desired outputs and how community needs play a role in workflows. Interestingly, we found that core concepts are congruent across disciplines and data publishing workflows.

The present paper describes our findings and presents them as components of a data publishing reference model. Based on the assessment of the current data publishing landscape, we highlight important gaps and challenges to consider, especially when dealing with more complex workflows and their integration into the wider community frameworks. We conclude the paper with recommendations to advance data publishing in line with the identified standards. It is our hope that as more research communities seek to publish data associated with their research, they will build on one or more of the components identified in creating their own workflows and thus accelerate uptake.

CONTENTS

INTRODUCTION

While some disciplines such as the social sciences, genomics, astronomy and Antarctic science have established cultures of sharing research data via repositories[2], it has generally not been common practice to deposit data for discovery and reuse by others – the barriers against doing so have simply been too high. Typically, it has only taken place when a community has committed itself toward open sharing (e.g., Bermuda Principles and Fort Lauderdale meeting agreements[3]), or there is a legal[4] requirement to do so, or where large research communities have access to discipline-specific facilities, instrumentation or archives. Among the major disincentives to sharing data through repositories is the amount of time required to prepare data for publishing, time that is perceived as being better spent on activities for which researchers receive credit (such as traditional research publications, obtaining funding, etc.). Unfortunately, when data are sequestered by researchers and their institutions, the likelihood of retrieval declines rapidly over time (Vines et al. 2014).

The advent of publisher and funding agency mandates to make data underlying scientific publications accessible is shifting the conversation from “Should researchers publish their data?” to “How can we publish data in a reliable manner?” We now see requirements for research reproducibility, openness and transparency and a new emphasis on data publication as a first-class research output. While there is still a prevailing sense that data carry less

[2] A repository (aka Data Repository or Digital Data Repository) is a searchable and queryable interfacing entity that is able to store, manage, maintain and curate Data/Digital Objects. Repository is a managed location (destination, directory or "bucket ") where digital data objects are registered, permanently stored, made accessible and retrievable, and curated (RDA, 2014). Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis (RDC, 2014). Similarly, ‘data services’ assist organizations in the capture, storage, curation, long-term preservation, discovery, access, retrieval, aggregation, analysis, and/or visualization of scientific data, as well as in the associated legal frameworks, to support disciplinary and multidisciplinary scientific research' (WDS, n.d.).

[3] http://www.genome.gov/10506376

[4] For example, the Antarctic Treaty Article III states that "scientific observations and results from Antarctica shall be exchanged and made freely available.” http://www.ats.aq/e/ats_science.htm

weight than published journal articles in the context of tenure and promotion decisions, recent studies demonstrate that when data are publicly available, a higher number of publications results (Piwowar and Vision 2013; Pienta et al. 2010).

Data publishing plays a key role in this new environment, sparking a community conversation around standards, workflows, and data quality assurance practices used by data repositories and data journals. A great deal of the activity around research data management is concerned with how best to handle the vast amounts of data and associated metadata in all their various formats. Standards are being developed by stakeholder groups such as the Research Data Alliance (RDA) and the World Data System of the International Council for Science (ICSU-WDS). In astronomy there has been a long process of developing metadata standards through the International Virtual Observatory Alliance (IVOA)[5]. Even in highly diverse fields such as the life sciences the BioSharing[6] initiative is attempting to coordinate community use of standards. There is a new understanding that data must be published and preserved for the long term to produce reliable scholarship, demonstrate reproducible research, facilitate new findings, enable repurposing, and hence realise benefits and maximise returns on research investments.

Traditionally, independent replication[7] of published research findings has been a cornerstone of scientific validation. However, there is increasing concern surrounding the reproducibility of published research, i.e., that a researcher's published results can be replicated using the data, code, and methods employed by the researcher (Peng 2011; Thayer et al. 2014; George et al. 2015). Here too, a profound culture change is needed to integrate reproducibility into the research process (Boulton et al. 2012, Stodden et al. 2013, Whyte & Tedds 2011). Data publishing is key to reproducible research and essential to safeguarding trust in science.

But what exactly is data publishing? Parsons and Fox (2013) question whether publishing is the correct metaphor when dealing with digital data. They suggest that the notion of data publishing can be limiting and simplistic and they recommend that we explore alternative paradigms such as the models for software release and refinement, rather than one-time publication (Parsons and Fox 2012). Certainly, version control[8] does need to be an integral part of data publishing, and this can distinguish it from the traditional journal article. Dynamic data citation is an important feature of many research datasets which will evolve over time,

[5] http://ivoa.org
[6] http://biosharing.org
[7] Replication is the evaluation of scientific claims by independent investigators using independent methods, data, equipment, and protocols (Peng, 2011).
[8] Version control (also known as 'revision control' or 'versioning') is control over time of data, computer code, software, and documents that allows for the ability to revert to a previous revision, which is critical for data traceability, tracking edits, and correcting errors (RDC, 2014).

e.g., monitoring data and longitudinal studies[9]. This challenge has already been covered by a data journal, e.g., the solution by Earth System Science Data and its approach to 'living data'[10].

As part of working toward a definition of data publishing, the Standards & Interoperability Committee of Research Data Canada[11] published a glossary of terms and definitions, and recently reviewed 32 international online data platforms for storage, data transfer, curation activities, preservation, access, and sharing features. A checklist was developed to compare criteria and features between platforms[12] (RDC 2014; RDC-SINCa 2015; Austin et al. 2015). The authors concluded that there is still a great deal of work to be done to ensure that online data platforms meet minimum standards for reliable curation and sharing of data. Guidelines were subsequently developed for the deposit and preservation aspects of publishing research data (RDC-SINC 2015b).

In the present paper we provide the results of a project undertaken to examine the role of repositories and data journals in publishing data and to characterize the resulting workflows. Our analysis involved the identification and description of a diverse set of workflows including basic self-publishing services, curated data repositories, and joint data journal and repository arrangements. This work enabled us to identify common components and standard practices as part of a reference model[13] for data publishing. While data publishing could be viewed as being synonymous with the processes of ensuring the quality and longevity of the published item (Lawrence 2011), the landscape is varied and during our analysis we uncovered important differences, gaps and challenges to consider. We describe these gaps and close the paper with recommendations for future action related to workflows for data. Based on our results, we also propose refined definitions for data publishing[14].

[9] https://www.rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html

[10] http://www.earth-system-science-data.net/living_data_process.html

[11] Research Data Canada (RDC) is moving to become Research Data Alliance - Canada (RDA-Canada).

[12] At the time of writing, the Research Data Alliance (RDA) Repository Platforms for Research Data Interest Group was also analyzing research data use cases in the context of repository platform requirements. The primary deliverable will be a matrix relating use cases with functional requirements for repository platforms. https://rd-alliance.org/groups/repository-platforms-research-data.html .

[13] In this case, we use the following understanding of a reference model: [...] is an abstract framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist. A reference model is not directly tied to any standards, technologies or other concrete implementation details, but it does seek to provide a common semantics that can be used unambiguously across and between different implementations. Source: OASIS, *https://www.oasis-open.org/committees/soa-rm/faq.php*

[14] The definitions will be submitted for adoption by the Research Data Alliance via the RDA Data Foundation and Terminology Interest Group, the RDA being a reference for a common language in this area (RDA, 2014a,b,c).

Terminology

When we use the term 'research data' we mean data[15,16] that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital outputs of a research project have the potential to become research data. Research data may be experimental, observational, operational, data from a third party, from the public sector, monitoring data, processed data, or repurposed data (RDC 2014).

Lawrence et al. (2011) define 'to Publish (sic) data,' as: "*To make data as permanently available as possible on the Internet*." Published data will have been through a process guaranteeing easily digestible information as to its trustworthiness, reliability, format and content. Callaghan et al. (2013) elaborate on this idea, arguing that formal publication of data provides a service over and above the simple act of posting a dataset on a website, in that it includes a series of checks on the dataset of either a technical (format, metadata) or a more scientific nature (is the data scientifically accurate?). Formal data publication also provides the data user with associated metadata, assurances about data persistence, and a platform for the dataset to be found and evaluated – all essential to data reuse.

The RDA Data Foundation and Terminology group has taken a repository-based approach to this issue. The group has defined data publication as a process whereby data are subjected to an assessment process to determine whether they should be acquired by a repository, followed by a rigorous acquisition and ingest process that results in products being made publicly available and supported for the long term by that repository (RDA 2014). Some authors make a distinction between metadata publication and data publication. However, we would argue that data and their associated metadata must at least be bi-directionally linked in a persistent manner, and that they need to be published together and viewed as a package since metadata are essential to the correct use, understanding, and interpretation of the data.

It is worth noting that there is continued discussion about these definitions (Parsons and Fox, 2012). Some views of data publishing overlap with those of 'research data management' (e.g.,

[15] Data are facts, measurements, recordings, records, or observations about the world collected by scientists and others, with a minimum of contextual interpretation. Data may be any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records (RDC 2014; Landry et al. 1970; Zin et al., 2007).

[16] See Zin et al. (2007) for an analysis of 130 definitions of data, information and knowledge provided by an expert panel of 45 leading scholars in information science, and the development of 5 models for defining data, information, and knowledge.

DCC, n.d.), or the 'digital curation lifecycle' (Higgins, 2008). However for the individuals involved in the process – whether as authors, researchers, or data managers - 'publication' also implies making data publicly discoverable[17]. While curation and data management include activities such as data appraisal, ingest and preservation, these do not necessarily result in data being made public (e.g., they may be held privately, and/or in a dark archive). In practice, we find that a focus on discoverability is indeed a key aspect of data publishing in journals and repositories.

METHODS

The RDA-WDS Publishing Data Workflows Working Group (WG)[18] was formed to provide an analysis of a representative range of existing and emerging workflows and standards for data publishing, including deposit and citation, and to provide components of reference models and implementations for application in new workflows. For the project, we compiled and analyzed a set of repository and journal workflows with a broad disciplinary spread, including major players in the data publishing world and organizations dealing with "long tail" data. The goal was to understand the key components of the workflows, the parties responsible for specific workflow steps, and the gaps or barriers to realising benefits and efficiencies. The project also analyzed quality assurance to determine where this was present in different workflows and the various mechanisms for improving quality.

The collection, analysis and standardization of data publishing workflows was (and is) an iterative process. Given the lack of a single source of reliable and comprehensive information, we chose a direct approach, contacting individuals in charge of key workflows. We consulted the registry of research data repositories maintained by re3data.org but concluded that we were interested in different types of information, including qualitative aspects. As the membership of the RDA-WDS Publishing Data Workflows WG was quite diverse in terms of disciplinary and stakeholder participation, we also drew upon the knowledge of that group.

During the period February 1-May 31, 2015, twenty-six data publishing platforms were surveyed, and the information was organized into a comparison matrix. Workflows were characterized across the following dimensions:

- Discipline

[17] The term "researcher" is used throughout this article to denote the person tasked with publishing the data. However, as explored in the Discussions section, a number of functions might be included here, for example: "data author", "technical support personnel", "laboratory chemist" or "project manager". Although not the focus of this paper, there is anecdotal evidence that primary research paper authors are not necessarily the same cohort as data products authors.

[18] https://www.rd-alliance.org/groups/rdawds-publishing-data-workflows-wg.html and https://www.icsu-wds.org/community/working-groups/data-publication/workflows

- Function of workflow
- The assignment of persistent identifiers (PIDs) to datasets
- The PID type used -- e.g., DOI, ARK, etc.
- Peer review of data (e.g., by researcher and by editorial review)
- Curatorial review of metadata (e.g., by institutional or subject repository)
- Technical review and checks (e.g., for data integrity at repository/data centre on ingest)
- Discoverability: Was there indexing of the data, and if so, where?
- Formats covered
- Persons/Roles involved, e.g., editor, publisher, data repository manager, etc.
- Links to additional data products (data paper; review; other journal articles) or "stand-alone" product
- Links to grants, usage of author PIDs
- Whether data citation was facilitated
- Whether the data life cycle was referred to
- Standards compliance

Terms were then standardized and the content normalized across the matrix to make it easier to compare and assess the various entries across platforms, and to identify basic and common elements versus "add-on" features. Several approaches were combined to produce an enhanced dataset. Some workflows were presented and discussed in videoconference and face-to-face meetings. In collaboration with the Force11 Implementation Group (Force11 2015), emphasis was also given to workflows facilitating data citation. Across the whole dataset, publicly available information was used to benchmark terms and ontologies. "Metrics" was added as an additional field. The resulting resource was then re-circulated to the group, whereupon a number of further annotations and corrections were made.

By this point in the analysis, the data workflows were grouped according to the following categories:

- Depositor/Initiator of the workflow
- Ingest/Curation/Metadata/Data administration
- Review/Quality assurance/Quality control
- Dissemination/Access
- Metrics/Additional services

The detailed information and categorization can be found in the analysis dataset (Murphy et al. 2015).

RESULTS AND ANALYSIS

Traditional research (Figure 1-A) is now evolving to include reproducible research workflows (Figure 1-B). In the present paper, we focus on and describe a basic reference model comprising elements of a data publishing workflow, represented by a generic repository workflow (Figure 1-1), and a generic data journal workflow (Figure 1-2). Both reference models are necessarily less detailed than the OAIS Reference Model (CCSDS, 2012) as they have been trimmed to highlight the core aspects of data publishing.

The workflow comparison demonstrates that it is usually the researcher who initiates the publication process once data have been collected and are in a suitable state for publication, or meet the repository requirements for submission. However, there are examples for which there is a direct “pipe” from a data production “machine” to a data repository (genome sequencing is one such example). Depending on the data repository, there are both human and automated quality assurance activities before data are archived for the long term. The typical data repository creates a repository entry (or database entry, or catalogue entry) for a specific dataset or a collection thereof. Most repositories invest in standardized dissemination for datasets, i.e., a landing page for each published item, as recommended by the Force11 Data Citation Implementation Group[19] (Starr et al. 2015). Additionally, some repositories facilitate third-party access for discoverability or metrics services.

[19] https://www.force11.org/datacitationimplementation

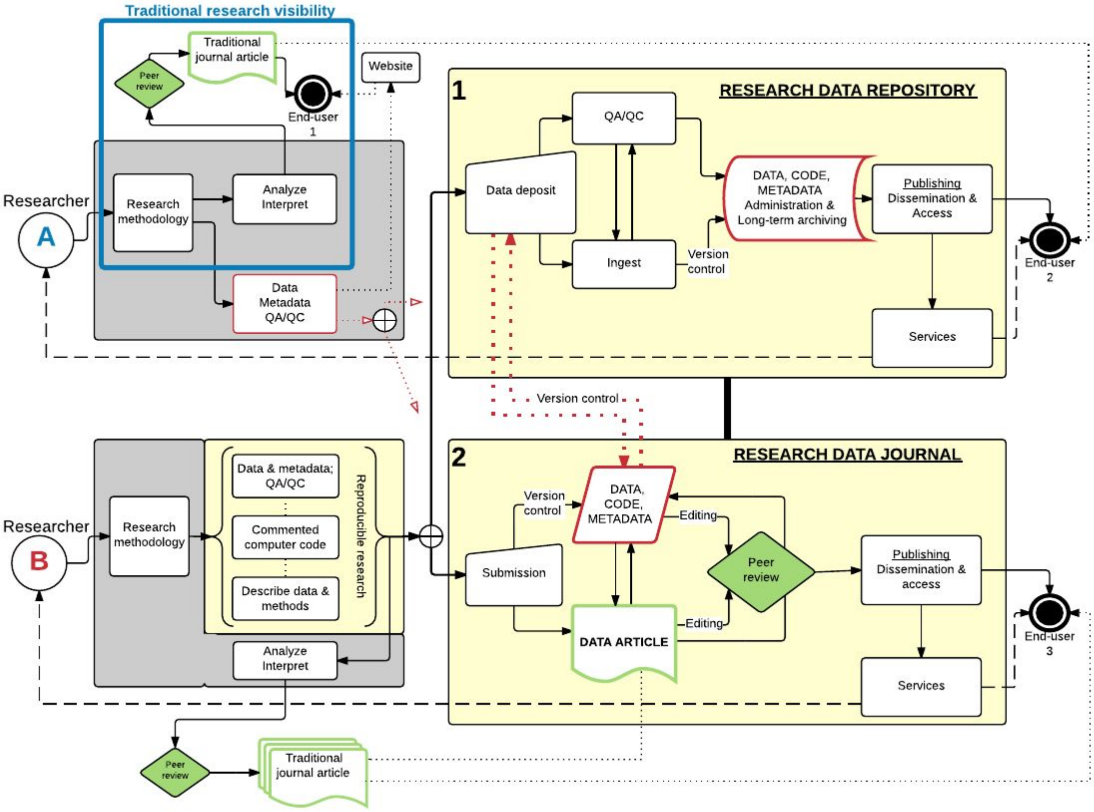

Figure 1. Research data publication workflows

As shown in Figure 1, researchers can and do follow a number of different pathways to communicate their work. Traditional, peer-reviewed journal publication is shown in the blue box (Figure 1-A). Emerging processes (highlighted in yellow) include reproducible research (Figure 1-B), depositing the data in research data repositories (Figure 1-1), and data journal articles (Figure 1-2). These are the two predominant workflows emerging from the analysis. Consequently, the various end-users have access to different slices of information: End-user #1 has access to the primary research article, and will have to search separately for underlying or related data (if it is available, which it often is not). End-user #2 has access to managed and curated data, code and supporting metadata, some of which may be reviewed and derive from reproducible research. End-user #2 may have to search for related journal articles. When data associated with a data journal article are uploaded to an approved repository[20], end-user #3 will have access to properly managed, curated, and peer reviewed data (including metadata and computer code), and a peer-reviewed data article deriving from reproducible research in addition to the traditional journal article that analyzes and interprets the results.

[20] Approved by the data journal

Data journals (Figure 1-2) also use the concept of the traditional journal (Figure 1-A). Although not completely analogous, similar terms and processes help situate the value of publishing in data journals alongside traditional research outputs. A researcher (A or B) might initiate related submissions to a data journal and a data repository either in parallel or sequentially, and before or after submission of an article to a traditional journal. There are some hard-wired automated workflows for this (e.g., with the Open Journal Systems-Dataverse integration (Castro and Garnett, 2014)), or there can be alternate automated or manual workflows in place to support the researcher (e.g., Dryad). Specific templates usually encourage researchers to provide metadata or documentation on important elements of the data package (e.g., methodology or code specific metadata). Data journal publishing's core processes consist of scientific peer review and dissemination of the datasets. Naturally, for the review team, this usually comprises pre-publication access to the dataset, which would be available in some data repository, and also demands explicit version control solutions for datasets and data papers. The final steps of data journal publishing are aligned with the traditional publishing process.

Diversity in Workflows

While the workflows in Figure 1 appear to be fairly straightforward, the underlying processes are, in fact, quite complex and diverse. These differences were most striking in the area of curation, which is not surprising given the broad spectrum of participants. Repositories that offered self-publishing options without curation had abridged procedures, requiring fewer resources but also potentially providing less contextual information and fewer assurances of quality. Disciplinary repositories that performed extensive curation and quality assurance had more complex workflows with additional steps, as one would expect. Such steps could be consecutive, facilitate more collaborative work on the data in the beginning, or anticipate standardized preservation steps in further steps.

Reviewing the metadata expectations across repositories revealed a different dimension of diversification. Highly specialized repositories frequently focused on specific metadata schemas and pursued curation accordingly. There was thus metadata heterogeneity across discipline-specific repositories. Some disciplines have established metadata standards, similar to the social sciences' use of DDI. In contrast, more general repositories tended to converge on domain-agnostic metadata schemas with fields common across disciplines, e.g., the mandatory DataCite fields (DataCite, 2015).

Meanwhile, the data journals also exhibited a notable degree of similarity in the overall workflow, but differed in terms of levels of support, review and curation. As with the repositories, the more specialized the journal (e.g., a discipline in the earth sciences with pre-established data sharing practices), the more prescriptive the author guidelines and the more specialized the review and quality assurance processes. With the rise of open, or post-publication, peer review, more journals are inviting the wider community to participate in the publication process.

As both the broader research community and some discipline-based communities are in the throes of developing criteria and practices for standardized release of research data, the services supporting these efforts, whether they resemble repositories or journals, also generally exhibited signs of being works in progress or proof-of-concept exercises rather than finished products. This is reflected in our analysis dataset (Murphy et al. 2015). Depending partly on their state of progress during our review period, and also on the specificity of the subject area, some workflow entries were rather vague, e.g., “not defined”, “it depends”, “in development”.

Data Deposit

We found that a majority of data deposit mechanisms, which launched the workflows, were initiated by researchers, but their involvement beyond the step of deposit varied across repositories and journals. For many repositories, engagement with researchers ended at deposit while for others, especially those with more complicated quality control, there were additional points of interaction. Platform purpose (e.g., data journal vs. repository) and the ultimate perceived purpose and motivation of the depositor of the data all affect the process. For example, a subject-specialist repository, such as is found at Science and Technology Facilities Council (STFC) or the National Snow and Ice Data Center (NSIDC), screened submissions and assessed the levels of metadata and support required. The data journals, however, typically adopted a “hands-off” approach for whilst the journal was considered to be the “publication” outlet, the data were housed elsewhere in some selected repository. Hence the journal publishing team often relied on external parties – repository managers and the research community in general[21] – to manage the data deposit and assess whether basic standards were met for data deposition or if quality standards were met for publishing (see details below).

[21] Post-publication peer review is becoming more prevalent and may ultimately strengthen the Parsons/Fox continual release paradigm. See, for instance, F1000 Research and Earth System Science Data - for instance, see the latter journal’s website: . http://www.earth-system-science-data.net/peer_review/interactive_review_process.html.

Ingest

We found that discipline-specific repositories had the most rigorous ingest and review processes and that more general repositories, e.g., institutional repositories (IRs) or Dryad, had a lighter touch. Some discipline-specific repositories had multiple-stage processes including several QA/QC processes and workflows based on OAIS. Many IRs had adopted a broader approach to ingest necessitated by their missions, which involved archiving research products generated across their campuses, especially those found in the long-tail of research data, and encompassing historical data that may have been managed in diverse ways. As data standards are developed and implemented and as researchers are provided with the tools, training, and incentives needed to engage in modern data management practices, ingest practices will no doubt evolve.

As mentioned above, data journals often rely on external data repositories to handle the actual data management. This requires a strong collaboration between the journal and repository staffs and trust that the repository will pursue data management and ingestion according to acceptable standard procedures. Data journals and data repositories are encouraged to make such agreements (e.g., Service Level Agreements) public and transparent to users.

Quality Assurance/Quality Control

A range of quality assurance/quality control (QA/QC)[22] activities was in evidence across the varied types of organizations in our analysis. The first level of QA/QC typically occurs during data collection and data processing, prior to submission of the data to a repository. Once the data have been submitted, there are two more levels of QA/QC. We distinguish between peer review and the types of technical and metadata reviews that repositories and journals generally conduct. Overall, QA/QC in data publishing is considered a hot-button topic and is debated heavily and continuously within the community. Mayernik et al. (2015) describe a range of practice in technical and academic peer review for publishing data. Most disciplinary repositories and all of the data journals that we reviewed had some QA/QC workflows. The level and type of QA/QC services varied, however. For example, established data repositories (such as ICPSR or Dataverse, see Murphy et al. 2015) tended to have dedicated data curation personnel to help in standardising and reviewing data upon submission and ingestion, especially in the area of metadata. Some domain repositories like ICPSR go farther to conduct

[22] Quality assurance: The process or set of processes used to measure and assure the quality of a product. Quality control: The process of meeting products and services to consumer expectations. (RDC, 2014)

in-depth quality control checks on the data, revising the data if necessary in consultation with the original investigator.

Some data repositories involved researchers in their QA/QC workflows to validate the scientific aspects of data, metadata or documentation. Technical support, data validation or QA/QC was also done by various repositories, but the level of engagement varied with the service and the individual institutions. Whereas some services checked file integrity, others offered more complex preservation actions, such as on-the-fly data format conversions. On the other end of the spectrum were some multi-purpose repositories which might facilitate researcher integration into QA/QC workflows, but this was not a standard practice.

Data journal workflows typically involve researchers through an external and formalized standard peer review analogous to the traditional publishing process, e.g., by invited peer reviewers, or open peer review models. The latter model, as mentioned previously, does shift the QA emphasis to cover significantly more elements in the overall publication workflow. The journal workflows we examined typically straddled the dual processes of reviewing the dataset itself and the data papers, which were carried out separately and then checked to ensure that the relationship between the two was valid. Such QA/QC workflows for data journals demand a strong collaboration with the research community and their peer reviewers.

Given the wide spectrum of QA/QC services being offered, these reflections should be taken into account for future recommendations:

- Repositories which put significant effort into high levels of QA/QC benefit researchers whose materials match the repository’s portfolio by making sure their materials are fit for future reuse.
- General research data repositories, which must accommodate a larger variety of data, may inevitably have some limitations in QA/QC and these should be made explicit.
- There is currently insufficient information available to clearly rank or assess the efficacy of the QA/QC processes involved. Information about quality levels could be made visible or accessible to users more clearly (and also possibly exposed to third parties, such as aggregators or metric services).

Data Administration and Long-Term Archiving

Data administration and management can cover a range of activities for organizations involved in the data publishing process and facing the challenges of managing access, in both the near- and long-term. These activities may include dealing with file types and formats, creation of access level restrictions, the establishment and implementation of embargo procedures, and assignment of identifiers. The survey of data publication workflows demonstrated, again, the assortment of practices that may be found in each of these areas. These can vary from providing merely file format guidelines to active file conversions; from supporting access restrictions to supporting only open access; administering flexible or standardized embargo periods; and employing different types of identifiers. Each practice selected by a repository has an effect on successfully meeting the goals of long-term archiving.

Data publication, as defined for this paper, is associated with long-term archiving to assure perpetual access. While most repositories in this sample have indicated a commitment to persistence and the use of standards, the actual ability to carry this out may be uncertain in the long term. However, the adoption of best practices and standards will increase the likelihood that published data will be maintained over time. Several discipline-specific repositories already have a long track record of preserving data and are particularly concerned about detailed archival workflows. Other repositories are fairly new to this discussion and continue to explore potential solutions.

One example of a solution to the challenges of long-term archiving is repository certification systems. These have been gaining momentum in recent years and could help facilitate data publishing through collaboration with data publishing partners such as funders, publishers and data repositories. A range of certification schemes exists[23], including those being implemented by organizations such as the Data Seal of Approval (DSA)[24] and the World Data System (ICSU-WDS)[25]. Certification of data repositories provides a transparent and objective base for evaluating their trustworthiness in terms of authenticity, integrity, confidentiality and availability of data and services. The certification processes are based on catalogues of evaluation criteria and they vary from the more thorough and stringent ISO standard (ISO 16363, 2012), to the basic DSA and WDS certifications which strike a balance between simplicity and robustness of the work and the effort involved. However, the corresponding guidance for researchers and collaborating infrastructures is often not visible outside of the specific domains in which these certifications have been established. Journals and data

[23] Data Seal of Approval (DSA); Network of Expertise in long-term Storage and Accessibility of Digital Resources in Germany (NESTOR) seal / German Institute for Standardization (DIN) standard 31644; Trustworthy Repositories Audit and Certification (TRAC) criteria / International Organization for Standardization (ISO) standard 16363; and the International Council for Science World Data System (ICSU-WDS) certification.

[24] Data Seal of Approval: http://datasealofapproval.org/en/

[25] World Data System certification https://www.icsu-wds.org/files/wds-certification-summary-11-june-2012.pdf

journals rely on a range of self-selected repositories chosen based on their scope and the researcher's own criteria.

Dissemination, Access and Citation

Data packages in most repositories were summarized on a single landing page that generally offered some basic or enriched (if not quality assured) metadata. This usually included a DOI and sometimes another unique identifier as well or instead. We found widespread use of persistent identifiers and a distinct recognition that data must be citable if it is to be optimally useful.[26]

It should be noted that dissemination of data publishing products was, in some cases, enhanced through linking and exposure (e.g., embedded visualization) in traditional journals. This is important, especially given the needed cultural shift within research communities to make data publishing the norm.

Dissemination practices varied widely. Many repositories supported publicly accessible data, but diverged in how optimally they were indexed for discovery. As would be expected, data journals tended to be connected with search engines, and with abstracting and indexing services. However, these often (if not always) related to the data article rather than to the dataset per se. With the launch of the Data Citation Index[27] by Thomson Reuters, this picture has been expanded recently. Disseminating the content in an automated fashion more widely would also help enrich it (see discussion in the next section about additional services).

The FAIR principles[28] and other policy documents (e.g. Boulton G. et al., 2012) explicitly mention that data should be accessible. Data publishing solutions ensure that this is the case. However, there are some workflows that allow only specific users to access sensitive data. An example is survey data containing information that could lead to the identification of individual survey respondents. In these cases, a prospective data user can typically access the detailed metadata for the survey to determine if it will meet the required research needs, but a data use agreement must be signed to grant access to the data. This potentially gives rise to a situation where the data article or descriptor can be published openly, perhaps with a Creative Commons license, but its counterpart dataset may be unavailable except via registration or other authorization processes. In such cases the data paper may still play a role

[26] Among the analyzed workflows it was generally understood, that data citation which properly attributes datasets to originating researchers can be an incentive for deposit of data in a form that makes it accessible and reusable, a key to changing the culture around scholarly credit for research data.

[27] http://wokinfo.com/products_tools/multidisciplinary/dci/

[28] https://www.force11.org/group/fairgroup/fairprinciples

in facilitating data discovery and facilitating reuse, where appropriate, and enabling contributing researchers to gain due credit[29].

Citation policies and practice also vary by community and culture. Increasingly, journals and scientific publishers are including data citation guidelines in their author support services. In terms of a best practice or standard, the *Joint Declaration of Data Citation Principles* (Data Citation Synthesis Group, 2014) is gathering critical mass, and becoming generally recognized and endorsed. Discussions concerning more detailed community practices are emerging; for example, whether or not publishing datasets and data papers – which can then be cited separately from related primary research papers – is a fair practice in a system that rewards higher citation rates. However, sensible practices can be formulated.[30]

Other Potential Value-Added Services/Metrics

Data publishing services are not the only places to store information about data. Many repository or journal providers look beyond the workflows that gather the information about the research data, and also want to make this information visible to other information providers in the field. This can add additional value to the data being published. If the information is exposed in a standardized fashion, data can be indexed and be made discoverable by third-party providers, e.g., data aggregators (see "services" box in Fig. 1). Considering that such data aggregators often work beyond the original data provider's subject or institutional focus, some data providers enrich their metadata (e.g., with data-publication links, keywords or more granular subject matter) to enable better cross-disciplinary retrieval. Ideally, information about how others download or use the data would be fed back to the original data providers. In addition, services such as ORCID[31] are being integrated to allow researchers to connect their materials across platforms. This gives more visibility to the data through the different registries, and allows for global author disambiguation. The latter is particularly important for establishing author metrics. During the investigation process, many data repository and data journal providers expressed an interest in new metrics for datasets and related objects. Tracking usage, impact and reuse of the materials shared can enrich the content on the original platforms and help in engaging users in further data sharing or curation activities. Furthermore, such information is certainly of interest for funders[32] of the respective infrastructures, and also funders of the research itself.

[29] See e.g. Open Health Data journal http://openhealthdata.metajnl.com/

[30] See Sarah Callaghan's blogpost: Cite what you use, 24 January 2014. Accessed 24 June 2015: http://citingbytes.blogspot.co.uk/2014/01/cite-what-you-use.html

[31] http://orcid.org/

[32] Funders have an interest in tracking Return of Investment to assess which researchers/projects/fields are effective and whether proposed new projects consist of new or repeated work.

Workflows to expose data publishing content to other providers in a standardized and strategic manner are crucial to enable discoverability of data. Presently, data reuse is hampered by the limited discoverability of the datasets. Projects such as the *Data Discovery Index*[33] are working on addressing this important issue and could serve as an accelerator to a paradigm shift for establishing data publishing within the communities.

One example of such a paradigm shift occurred in 2014 when the Resource Identifier Initiative (RII) launched a new registry within the biomedical literature. The project covered antibodies, model organisms (mice, zebrafish, flies), and tools (i.e., software and databases), providing a rather comprehensive combination of data, metadata and platforms to work with. Eighteen months later the project was able to report both a cultural shift in behaviour and a significant increase in the potential reproducibility of relevant research.[34] As discussed in Bandrowski et al (2015), the critical factor in this initiative's success in gaining acceptance and uptake was the integrated way in which it was rolled out. A group of stakeholders - including researchers, journal editors, subject community leaders and publishers - within a specific discipline, neuroscience, worked together to ensure a consistent message with a compelling rationale, coherent journal policies which necessitated compliance in order for would-be authors to publish, and a specific workflow for the registration process (complete with skilled, human support if required). Further work is required to determine exactly how this use case can be leveraged across the wider gamut of subjects, communities and other players.

DISCUSSION and CONCLUSIONS

The results from our analysis of data publishing workflows underscore an immense diversity, although basics components were fairly similar across platforms. Based on these results we refine the definition of data publishing below. Given the rapid developments in this field over the past years, it is to be expected that the diversity might grow even further. Some evident gaps and challenges hinder global interoperability and adoption. However, the results of our survey suggest that new solutions (e.g., for underrepresented disciplines) should build on some established components such as QA/QC workflows that best match the targeted use cases.

[33] http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-14-031.html

[34] The Resource Identification Initiative: A cultural shift in publishing [v1; ref status: indexed, http://f1000r.es/5fj]. Anita Bandrowski, Matthew Brush, Jeffrey S. Grethe, Melissa A. Haendel, David N. Kennedy, Sean Hill, Patrick R. Hof, Maryann E. Martone, Maaike Pols, Serena Tan, Nicole Washington, Elena Zudilova-Seinstra, Nicole Vasilevsky.

Key Components of the Reference Model

The present analysis provides a set of components that contribute to the generic reference models for data publishing. We distinguish two sets of services: basic and add-ons. The basic set consists of entries in a trusted data repository, including a persistent identifier, standardized metadata, and basic curation services. All additional services might be considered add-ons (Figure 2). These include components such as contextualisation through additional embedding into data papers or links to traditional papers, QA/QC and peer review, and services to enhance the visibility of datasets.

Generally, trusted data publishing should be an integrated chain of actions from both boxes in Figure 2. Depending on the use case, however, it might be appropriate to select one element or another from Figure 2. In the light of future reuse, we would argue that the basic elements of curation, QA/QC, and referencing should be covered.

Although visibility services have been included in the add-ons, given the importance of discoverability to facilitate future reuse, we would argue they should be incorporated into the basic service set of modern and trusted data publishing, and indeed, our workflows in Figure 1 show a link from Data Publication to Services.

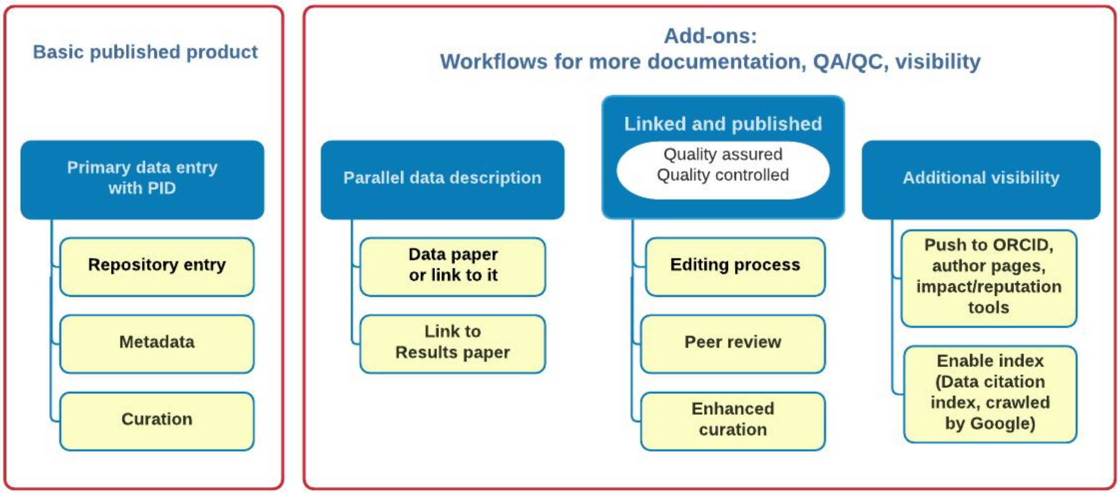
Basic published product
Primary data entry with PID
Repository entry
Metadata
Curation
Add-ons: Workflows for more documentation, QA/QC, visibility
Parallel data description
Data paper or link to it
Link to Results paper
Linked and published
Quality assured Quality controlled
Editing process
Peer review
Enhanced curation
Additional visibility
Push to ORCID, author pages, impact/reputation tools
Enable index (Data citation index, crawled by Google)

Figure 2. Data publishing reference model - Key components

Proposed New Definitions

Based on the analysis above, we propose definitions for the following terms[35]:

[35] These terms do not currently appear in the RDA Glossary of Terms.

Research data publishing

> *"Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable – all aspects of data publishing that are important for future reuse of data by third party end-users."*

The proposed definition applies also to the publication of confidential or sensitive data with the appropriate safeguards and accessible metadata. A practical application of this may be a published journal article that allows discoverability and citation of a dataset, while identifying access criteria for reuse, i.e., either not linking directly to the dataset or restricting access to the linked dataset. At the institutional level, for example, Harvard is developing a tool that will eventually be integrated with Dataverse to share and use confidential and sensitive data in a responsible way (Harvard, 2015).

Research data publishing workflows

> *"Research data publishing workflows are activities and processes in a digital environment that lead to the publication of research data, associated metadata and accompanying documentation and software code on the Web. In contrast to interim or final published products, workflows are the means to curate, document, and review, and thus ensure and enhance the value of the published product. Workflows can involve both humans and machines and often humans are supported by technology as they perform steps in the workflow. Similar workflows may vary in the details depending on the research discipline, data publishing product and/or the host institution of the workflow (e.g., individual publisher/journal, institutional repository, discipline-specific repository)."*

Gaps and Challenges

While there are still some disciplines for which no specific domain repositories exist, in general we are seeing a greater number of repositories of various types (re3data.org indexes over 1,200 repositories). In addition to the many disciplinary repositories, there are several new repositories designed to house broader collections, e.g., Zenodo, Figshare, Dryad, Dataverse, and the institutional repositories at colleges and universities. "Staging" repositories with a collaboration focus are also springing up, providing extensions to the more traditional workflows and reaching out into the collaborative working space -- e.g., Open Science Framework[36] which has a publishing workflow with Dataverse. Another example is the SEAD[37] (Sustainable Environment Actionable Data) project, which provides project spaces in which scientists manage, find, and share data, and connects researchers to repositories that will provide long-term access and preservation of data.

Despite this recent data publishing activity, gaps and challenges remain. Many of these come into play when considering more complex workflows. Some of the challenges that came to light during our analysis include:

- Bi-directional linking -- How do we link data and publications persistently in an automated way? Several organizations, including RDA and WDS[38], are now working on this problem. A related issue is the persistence of the links themselves.[39]
- Software management -- Solutions are needed to manage, preserve, publish and cite software. Basic workflows exist (involving code sharing platforms, repositories and aggregators), but much more work needs to be done to establish a wider framework, including community updating and initiatives involving linking to the data .
- Version control -- In general, we found that repositories handle version control in different ways, which is potentially confusing. While many version control solutions might be tailored to discipline-specific challenges, there is a need to address standardization as well. This concerns corresponding provenance information as well.
- Sharing restricted-use data -- There are different workflows at different repositories and journals, and most are not yet equipped to handle confidential data. It is important that the mechanism for data sharing be appropriate to the level of sensitivity of the data. The time is ripe for the exchange of expertize in this area so that workflows can be reused across data types.
- Role clarity -- Data publishing relies on a collaborative approach. For better user guidance and trustworthiness of the services involved, an improved understanding of roles, responsibilities, and collaboration is needed. A “who does what” in the current, mid- and long-term would ensure a smoother provision of service.
- Business models -- There is strong interest in establishing the value and sustainability of repositories. Beagrie & Houghton (2014)[40] produced a synthesis of data centre studies combining quantitative and qualitative approaches in order to quantify value

[36] https://osf.io/

[37] http://sead-data.net/

[38] RDA/WDS Publishing Data Services WG: https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html and https://www.icsu-wds.org/community/working-groups/data-publication/services

[39] See the hiberlink Project for information on this problem and work being done to solve it: http://hiberlink.org/dissemination.html

[40] http://blog.beagrie.com/2014/04/02/new-research-the-value-and-impact-of-data-curation-and-sharing/

in economic terms and present other, non-economic, impacts and benefits. A recent Sloan-funded meeting of 22 data repositories led to a white paper on Sustaining Domain Repositories for Digital Data[41]. However, much more work is needed in the context of publishing data[42], to understand viable financial models and distinguish trustworthy collaborations.

- Incentives -- Data publishing offers clear incentives to researchers, e.g., a citable data product, persistent data documentation, and (potentially) information about the impact of a scholar's work. Also, many repositories offer support when submitting data. Such incentives should be communicated more explicitly and potentially collaboratively. In addition, we should be trumpeting the fact that formal data archiving results in greater numbers of papers and thus more science, as Piwowar and Vision (2013) and Pienta et al. (2010) have shown.
- Data citation support -- Broadly, there is only partial implementation of the practices and procedures recommended by the Data Citation Implementation Group (Starr, 2015). However, there is a widespread awareness. There are also a wide range of PIDs emerging: ORCID, DOI, FunderRef, RRID, IGSN, ARK and many more. Clarity and ease of use needs to be brought to this landscape.[43]

In addition to the above challenges, the challenges of more complex data – in particular, big data and dynamic data – need to be addressed. Data publishing needs to be 'future-proof' in these cases, also. Data publication processes from the past 10 years still focus on irrevocable, fully documented data for unrestricted (scientific) use (Brase et al., 2015), but there is a requirement from the research communities[44] to cite data before it has reached an overall irrevocable state and before it has been archived. This particularly holds true for scientific communities with high volume data (such as High-Energy Physics and Climate Sciences), and for data citation entities including multiple individual datasets for which the time needed to reach an overall stable data collection is increasing. This may result from an increasing number of variables requested and provided, and technical infrastructures reducing the effort for data collection. For these dynamic datasets, a citation is needed by the creators as well as by data users. A citation requires a citable published data product. This topic has not been included in the general data publication discussion as yet. However, version control and

[41] http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

[42] RDA/WDS Publishing Data Costs IG adresses this topic: https://rd-alliance.org/groups/rdawds-publishing-data-ig.html

[43] The THOR project is just starting at time of writing and will be operating in this space. http://project-thor.eu/

[44] For example, in genomics, there is the idea of numbered "releases" of, for example, a particular animal genome, so that while refinement is ongoing it is also possible to refer to a reference data set.

keeping a good provenance record[45] of the datasets is crucial for citations of such data collections and is an indispensable part of the data publishing workflow.

A proposal for scalable dynamic citation has been put forward by the Research Data Alliance data citation working group[46] (RDA, 2015). This solution uses version control for data collection (which ensures changes and deletions to data are assigned timestamps), assignment of Persistent Identifiers (PIDs) and timestamps to queries and expressions that identify a cited subset of data, and computation of hash keys for the selection result to ensure subsequent verification of identity.

With respect to the gaps and challenges discussed above, we realize that the set of workflows we examined is limited and thus our gap analysis may be incomplete. A more general challenge we encountered during the project is that it is difficult to find clear and consistent human-readable workflow representations for repositories. The trust standards (e.g., Data Seal of Approval, Nestor, ISO 16363 and World Data System) require that repositories document their processes, so this may change in the future, but we recommend that repositories publish their workflows in a standard way for greater transparency. It should be emphasized that this is not only a matter of technical trustworthiness, but also helps in user engagement and collaborations. Researchers are often not aware of the different workflows and data publishing solutions available. Thus, in the new field of data publishing this could be a crucial step to help researchers find their way through the diverse options available. The analysis has shown that a variety of workflows exists, with potentially many more emerging, so that researchers could or will be able to choose their best fit. It is necessary to improve the guidance which distinguishes relevant features, such as QA/QC and different service or support levels.

Conclusions and Best Practice Recommendations

In conclusion, we make the following suggestions to organizations establishing new workflows and to those seeking to transform existing procedures:

- Start small and build components one by one, with a good understanding of how each building block fits into the overall workflow and what the final objective is. These building blocks should be open source/shareable components.

[45] For scientific communities with high volume data, the storage of every dataset version is often too expensive. Versioning and keeping a good provenance record of the datasets is crucial for citations of such data collections. Technical solutions are developed e.g. by the European Persistent Identifier Consortium (EPIC).

[46] https://rd-alliance.org/groups/data-citation-wg.html

- Follow standards whenever available to facilitate interoperability and to permit extensions based on the work of others using the same standards. For example, Dublin Core is a widely used metadata standard, making it relatively easy to share metadata with other systems. Use disciplinary standards where/when applicable.
- It is especially important to implement and adhere to standards for data citation, including the use of persistent identifiers (PIDs). Linkages between data and publications can be automatically harvested if DOIs for data are used routinely in papers. The use of PIDs can also enable linked open data functionality.
- Document roles, workflows and services. A key difficulty we had in conducting the analysis of the workflows was due to the lack of complete, standardized and up-to-date information about the processes and services provided by the platforms themselves. This impacts potential users of the services as well. Part of the trusted repository reputation development should include a system to clarify ingest support levels, long-term sustainability guarantees, subject expertize resource, and so forth.

To close, against the backdrop of the gaps and challenges outlined above, we offer a description of the key components of a best practice scenario based on our workflow analysis and lessons learned. First, we would like to see a workflow that results in all scholarly objects being connected, linked, citable, and persistent to allow researchers to navigate smoothly across and to enable reproducible research. This includes linkages between documentation, code, data, and papers in an integrated environment to streamline and improve the research process. A big challenge here is that the repositories and higher education institutions (holding a critical mass of research data) have not yet fully engaged with the large scientific publishers (which host the critical mass of discoverable, published research) and vice versa. Although new journal formats, such as Elsevier’s “Article of the Future” and Wiley’s “Anywhere Article,” which link data to papers and enrich the reading experience are increasingly being developed, progress is still being impeded by cultural, technical and business model issues. Second, in the ideal workflow scenario, all of these objects need to be well documented to enable reproducible research and to ensure that other researchers can reuse the data for new discoveries. Third, we would like to see information standardized and exposed via APIs and other mechanisms so that metrics on data usage can be captured as part of the workflow. Related to this, we note that biases in the funding and academic reward systems tend to value cutting-edge high-tech projects above data-driven secondary analysis and reuse of existing data. More attention (i.e., more perceived value) from funders is a key component in changing this paradigm.

Finally, all the components of this system need to work seamlessly in some sort of an integrated environment. Therefore, we advocate for the implementation of standards, and the development of new standards where necessary, for repositories and all parts of the data publishing process. Trusted data publishing should be embedded in documented workflows. This helps to establish collaborations with potential partners and gives guidance to the researchers. The latter enables and encourages the deposit of reusable research data that will be persistent while preserving provenance.

REFERENCES

Austin CC; Brown S; Fong N; Humphrey C; Leahey L; Webster P (2015). "Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements." Presented at the IASSIST Annual Conference, Minneapolis MN, June 2-5; IASSIST Quarterly Preprint. International Association for Social Science, Information Services, and Technology. https://drive.google.com/file/d/0B_SRWahCB9rpRF96RkhsUnh1a00/view

Beagrie N; Houghton JW (2014). "The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres." JISC http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

Brase J; Lautenschlager M; Sens I (2015). "The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite." D-Lib Magazine. Volume 21, Number 1/2. doi: 10.1045/january2015-brase

Boulton G. et al (2012). "Science as an Open Enterprise", Royal Society Report, Available at https://royalsociety.org/policy/projects/science-public-enterprise/Report/

Callaghan S, Murphy F, Tedds J, Allan R, Kunze J, Lawrence R, Mayernik MS, Whyte A (2013). "Processes and Procedures for Data Publication: A Case Study in the Geosciences." The International Journal of Digital Curation, 8(1), doi:10.2218/ijdc.v8i1.253

Castro E., Garnett A. (2014): Building a Bridge Between Journal Articles and Research Data: The PKP-Dataverse Integration Project. International Journal of Digital Curation, 9(1), pp. 176-184. doi:10.2218/ijdc.v9i1.311

CCSDS (2012). "Recommendation for Space Data System Practices: Reference Model for an Opean Archival Information System (OAIS), CCSDS 650.0-M-2." http://public.ccsds.org/publications/archive/650x0m2.pdf DataCite (2015). "DataCite Metadata Schema for the Publication and Citation of Research Data." http://dx.doi.org/10.5438/0010

DCC (n.d.) ‘Research Data Management’ entry in ‘Digital Curation Glossary’ Retrieved from: http://www.dcc.ac.uk/digital-curation/glossary#R

Ember C, Hanisch R (2013). Sustaining Domain Repositories for Digital Data: A White Paper. Inter-university Consortium for Political and Social Research (ICPSR). http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

Force11 (2014). Joint Declaration of Data Citation Principles. Data Citation Synthesis Group, Martone M. (ed.). San Diego CA. https://www.force11.org/group/joint-declaration-data-citation-principles-final

Force11 (2015). Future Of Research Communications and e-Scholarship https://www.force11.org/group/data-citation-implementation-group

George BJ, Sobus JR, Phelps LP, Rashleigh B, Simmons JE, Hines RN(2015). Raising the Bar for Reproducible Science at the U.S. Environmental Protection Agency Office of Research and Development. Toxicological Sciences, 145(1), 16–22. http://toxsci.oxfordjournals.org/content/145/1/16.full.pdf+html

Harvard (2015). Privacy tools project - Data tags. Harvard School of Engineering and Applied Sciences. http://privacytools.seas.harvard.edu/datatags

Higgins, S. (2008). The DCC curation lifecycle model. International Journal of Digital Curation, 3(1).pp. 134-140 doi:10.2218/ijdc.v3i1.48

ISO 16363 (2012). Space data and information transfer systems — Audit and certification of trustworthy digital repositories. http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510.

Kratz J, Strasser C (2014). “Data Publication Consensus and Controversies [v3; ref status: indexed, http://f1000r.es/4ja].” F1000Research. http://f1000research.com/articles/3-94/v3

Landry BC, Mathis BA., Meara NM, Rush JE, Young CE (1970). Definition of some basic terms in computer and information science. Journal of the American Society for Information Science, 24(5), 328–342.

Lawrence B, Jones C, Matthews B, Pepler S, Callaghan S (2011). Citation and Peer Review of Data: Moving Toward Formal Data Publication.” *The International Journal of Digital Curation.* http://www.ijdc.net/index.php/ijdc/article/view/181/265. doi:10.2218/ijdc.v6i2.205

Mayernik M.S., Callaghan S., Leigh R., Tedds J.A. and Worley S., 2015: Peer Review of Datasets: When, Why, and How. Bulletin of the American Meteorological Society, 96(2), 191–201. http://dx.doi.org/10.1175/BAMS-D-13-00083.1

Murphy, Fiona et al.. (2015). WDS-RDA Publishing Data Workflows Working Group Analysis sheet. Zenodo. 10.5281/zenodo.19107

Parsons M, and Fox P (2013). Is data publication the right metaphor? Data Science Journal, Volume 12. https://www.jstage.jst.go.jp/article/dsj/12/0/12_WDS-042/_article

Peng RD (2011). Reproducible research in computational science. Science, 334(6060), 1226-1227.

Pienta AM, Alter GC, & Lyle JA (2010). *The enduring value of social science research: The use and reuse of primary research data*. Retrieved from http://hdl.handle.net/2027.42/78307

Piwowar H, Vision T. (2013). “Data reuse and the open data citation advantage.” PeerJ Computer Science. https://peerj.com/articles/175/

RDA (2014a). TeD-T: Term definition tool. Research Data Alliance, Data Foundations and Terminology Working Group. http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

RDA (2014b). Data Foundation and Terminology DFT 3: Snapshot of DFT Core Terms. Research Data Alliance, Data Foundations and Terminology Working Group. https://www.rd-alliance.org/system/files/DFT3%20-%20Snapshot%20of%20core%20terms.pdf

RDA (2014c). Data Foundation and Terminology DFT 4: Use cases. https://rd-alliance.org/system/files/DFT4%20-%20Use%20Cases.pdf

RDA (2015). Scalable dynamic citation. Research Data Alliance data citation working group. https://rd-alliance.org/sites/default/files/attachment/Scalable%20Dynamic%20Data%20Citation_v1.pdf

RDC (2014). Glossary of terms and definitions. Research Data Canada. http://www.rdc-drc.ca/glossary/

RDC-SINC (2015a). Research Data Repository Requirements and Features Review. Research Data Canada, Standards and Interoperability Committee http://hdl.handle.net/10864/10892

RDC-SINC (2015b). Guidelines for the deposit and preservation of research data in Canada. Research Data Canada. http://www.rdc-drc.ca/wp-content/uploads/Guidelines-for-Deposit-of-Research-Data-in-Canada-2015.pdf

Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1:e1 https://dx.doi.org/10.7717/peerj-cs.1

Stodden V, Bailey DH, Borwein J, LeVeque RJ, Rider W, and Stein W (2013). Setting the Default to Reproducible. Workshop in Reproducibility in Computational and Experimental Mathematics (December 10-14, 2012), Institute for Computational and Experimental Research in Mathematics. https://icerm.brown.edu/tw12-5-rcem/icerm_report.pdf

Thayer KA, Wolfe MS, Rooney AA, Boyles AL, Bucher JR, and Birnbaum LS (2014). Intersection of systematic review methodology with the NIH reproducibility initiative. Environmental Health Perspectives. 122, A176–A177. http://ehp.niehs.nih.gov/wp-content/uploads/122/7/ehp.1408671.pdf

Tucker, D (2014) “Arkivum and Figshare Announce Partnership”. http://arkivum.com/news/arkivum-figshare-announce-partnership/ (accessed 29 June 2015)

Vines TH, Albert AYK, Andrew RL, DeBarre F, Bock DG, Franklin MT, Kimberly J. Gilbert KJ, Moore JS, Renaut S,Rennison DJ (2014). “The Availability of Research Data Declines Rapidly with Article Age.” *Current Biology,* 24(1), 94-97.

Whyte, A., Tedds, J. (2011). ‘Making the Case for Research Data Management’. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm

Zin et al. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. Journal of the American Society for information science and technology, 58(4):479–493.