Int J Digit Libr (2017) 18:77–92
DOI 10.1007/s00799-016-0178-2

Key components of data publishing: using current best practices to develop a reference model for data publishing

Claire C. Austin[1,2] · **Theodora Bloom**[3] · **Sünje Dallmeier-Tiessen**[4] · **Varsha K. Khodiyar**[5] · **Fiona Murphy**[6] · **Amy Nurnberger**[7] · **Lisa Raymond**[8] · **Martina Stockhause**[9] · **Jonathan Tedds**[10] · **Mary Vardigan**[11] · **Angus Whyte**[12]

Received: 30 June 2015 / Revised: 13 May 2016 / Accepted: 25 May 2016 / Published online: 20 June 2016

Abstract The availability of workflows for data publishing could have an enormous impact on researchers, research practices and publishing paradigms, as well as on funding strategies and career and research evaluations. We present the generic components of such workflows to provide a reference model for these stakeholders. The RDA-WDS Data Publishing Workflows group set out to study the current data-publishing workflow landscape across disciplines and institutions. A diverse set of workflows were examined to identify common components and standard practices, including basic self-publishing services, institutional data repositories, long-term projects, curated data repositories, and joint data journal and repository arrangements. The results of this examination have been used to derive a data-publishing reference model comprising generic components. From an 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 wider community frameworks. It is clear that the data-publishing landscape is varied and dynamic and that there are important gaps and challenges. The different components of a data-publishing system need to work, to the greatest extent possible, in a seamless and integrated way to support the evolution of commonly understood and utilized standards and—eventually—to increased reproducibility. We therefore advocate the implementation of existing standards for repositories and all parts of the data-publishing process, and the development of new standards where necessary. Effective and trustworthy data publishing should be embedded in documented workflows. As more research communities seek to publish the data associated with their research, they can build on one or more of the components identified in this reference model.

Keywords Data publishing · Open data · Open Science · World Data System · Research Data Alliance

Author statement: All authors 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. Authors contributed to the writing of the article itself and significantly to the analysis. Contributors Timothy Clark, Eleni Castro, Elizabeth Newbold, Samuel Moore and Brian Hole shared their workflows with the group (for the analysis). The authors are listed in alphabetical order.

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

✉ Fiona Murphy
fionalm27@gmail.com

1 Research Data Canada, Toronto, Canada
2 Carleton University, Ottawa, Canada
3 BMJ, London, UK
4 CERN, Geneva, Switzerland
5 Nature Publishing Group, London, UK
6 University of Reading, Reading, UK
7 Columbia University, New York, USA
8 Woods Hole Oceanographic Institution, Woods Hole, USA
9 German Climate Computing Centre (DKRZ), Hamburg, Germany
10 University of Leicester, Leicester, UK
11 University of Michigan/ICPSR, Leicester, UK
12 Digital Curation Centre, Edinburgh, Scotland, UK

1 Data availability

Data from the analysis presented in this article are available in:

Bloom, T., Dallmeier-Tiessen, S., Murphy, F., Khodiyar, V.K., Austin, C.C., Whyte, A., Tedds, J., Nurnberger, A., Raymond, L., Stockhause, M., Vardigan, M. *Zenodo* doi:10.5281/zenodo.33899 (2015)

2 Introduction

Various data-publishing workflows have emerged in recent years to allow researchers to publish data through repositories and dedicated journals. While some disciplines, such as the social sciences, genomics, astronomy, geosciences, and multidisciplinary fields such as Polar science, have established cultures of sharing research data.[1] via repositories,[2] it has traditionally not been common practice in all fields for researchers to deposit data for discovery and reuse by others. Typically, data sharing has only taken place when a community has committed itself towards open sharing (e.g. Bermuda Principles and Fort Lauderdale meeting agreements for genomic data),[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.

A significant barrier to moving forward is the wide variation in best practices and standards between and within disciplines. Examples of good practice include standardized data archiving in the geosciences, astronomy, and genomics. Archiving for many other kinds of data is only just beginning to emerge or is non-existent [1]. A major disincentive for sharing data via repositories is the amount of time required to prepare data for publishing, time that may be 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 [2].

The advent of publisher and funding agency mandates to make accessible the data underlying publications 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 openness and transparency, and a drive towards regarding data as a first-class research output. Data publishing can provide significant incentives for researchers to share their data by providing measurable and citable output, thereby accelerating an emerging paradigm shift. Data release is not yet considered in a comprehensive manner in research evaluations and promotions, but enhancements and initiatives are under way within various funding and other research spaces to make such evaluations more comprehensive [3]. While there is still a prevailing sense that data carry less 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 [4,5].

The rationale for sharing data is based on assumptions of reuse— if data are shared, then users will come. However, the ability to share, reuse, and repurpose data depends upon the availability of appropriate knowledge infrastructures. Unfortunately, many attempts to build infrastructure have failed because they are too difficult to adopt. The solution may be to enable infrastructure to develop around the way scientists and scholars actually work, rather than expecting them to work in ways that the data center, organizational managers, publishers, or funders would wish them to [6]. Some surveys have found that researchers' use of repositories ranks a distant third—after responding to individual requests and posting data on local websites [7].

Traditionally, independent replication 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 reproduced using the data, code, and methods employed by the researcher [8–10]. Here, too, a profound culture change is needed if reproducibility is to be integrated into the research process [11–13]. Data availability

[1] When we use the term 'research data' we mean data 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 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 (Research Data Canada (2015), Glossary of terms and definitions, http://dictionary.casrai.org/Category:Research_Data_Domain).

[2] A repository (also referred to as a 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. A repository is a managed location (destination, directory or 'bucket') where digital data objects are registered, permanently stored, made accessible and retrievable, and curated (Research Data Alliance, Data Foundations and Terminology Working Group. http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page). Repositories preserve, manage, and provide access to many types of digital material 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 (Research Data Canada (2015), Glossary of terms and definitions, http://dictionary.casrai.org/Category:Research_Data_Domain). 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.

[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.html.

is key to reproducible research and essential to safeguarding trust in science.

As a result of the move towards increased data availability, a community conversation has begun about the standards, workflows, and quality assurance practices used by data repositories and data journals. Discussions and potential solutions are primarily concerned with how best to handle the vast amounts of data and associated metadata in all their various formats. Standards at various levels are being developed by stakeholder groups and endorsed through international bodies such as the Research Data Alliance (RDA), the World Data System of the International Council for Science (ICSU-WDS), and within disciplinary communities. For example, in astronomy there has been a long process of developing metadata standards through the International Virtual Observatory Alliance (IVOA),[5] while in the climate sciences the netCDF/CF convention was developed as a standard format including metadata for gridded data. Even in highly diverse fields such as the life sciences, the BioSharing[6] initiative is attempting to coordinate community use of standards. Increasingly, there is a new understanding that data publishing ensures long-term data preservation and hence produces reliable scholarship, demonstrates reproducible research, facilitates new findings, enables repurposing, and thereby realizes benefits and maximizes returns on research investments.

But what exactly is data publishing? Parsons and Fox [14] question whether publishing is the correct term when dealing with digital information. They suggest that the notion of data publishing can be limiting and simplistic and recommend that we explore alternative paradigms such as the models for software release and refinement, rather than one-time publication [14]. Certainly, version control[7] 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, e.g. monitoring data and longitudinal studies [15]. The data journal *Earth System Science Data* is addressing this challenge with its approach to 'living data'.[8] The RDA Dynamic Citation Working group has also developed a comprehensive specification for citing everything from a subset of a dataset to data generated dynamically, 'on-the-fly' [16]. International scientific facilities typically plan periodic scientifically processed data releases through the lifetime of a mission (e.g. XMM-Newton X-ray Telescope source catalogue, [17], in addition to making underlying datasets available through archives according to embargo policies.

In 2011, Lawrence et al. [18] defined the act of 'publishing 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. [19] 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 content-based nature (e.g. are the data 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 of which are essential to data reuse. An important consideration for our study is that support for 'normal' curation falls short of best practice standards. For example, having conducted a survey of 32 international online data platforms [20], the Standards & Interoperability Committee of Research Data Canada (RDC)[9] 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, and developed guidelines for the deposit and preservation aspects of publishing research data.

With the present study, a first step is taken towards a reference model comprising generic components for data publishing—which should help in establishing standards across disciplines.

We describe selected data-publishing solutions, the roles of repositories and data journals, and characterize workflows currently in use. Our analysis involved the identification and description of a diverse set of workflows, including basic self-publishing services, long-term projects, curated data repositories, and joint data journal and repository arrangements. Key common components and standard practices were then identified as part of a reference model for data publishing. These could help with standardizing data-publishing activities in the future (while leaving enough room for disciplinary or institutional practices). It is worth noting that there is continued discussion about many of the key definitions. The working group presents core data-publishing

[5] http://www.ivoa.net.

[6] http://biosharing.org.

[7] Version control (also known as 'revision control' or 'versioning') is control over a time period of changes to 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. 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.

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

[9] Research Data Canada (RDC) is an organizational member of Research Data Alliance (RDA) and from the beginning has worked very closely with RDA. See: "Guidelines for the deposit and preservation of research data in Canada, http://www.rdc-drc.ca/wp-content/uploads/Guidelines-for-Deposit-of-Research-Data-in-Canada-2015.pdf and, "Research Data Repository Requirements and Features Review", http://hdl.handle.net/10864/10892.

terms (definitions) based on the analysis. We compare, contrast, and evaluate the key components, and identify and assess their utility and value-enhancing capabilities. We discuss the challenges inherent in citing and disseminating data, and then give context to already existing initiatives in this space. We outline continuing gaps and challenges—themselves opportunities for further research—and finally include a practical, modular set of recommendations as part of our conclusions.

3 Methods and materials

The RDA-WDS Publishing Data Workflows Working Group (WG) was formed to provide an analysis of a reasonably 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. The present work was specifically focused on articulating a draft reference model comprising generic components for data-publishing workflows that others can build upon. We also recognize the need for the reference model to promote workflows that researchers find usable and attractive.

To achieve this, the working group followed the OASIS definition of a reference model as:

"…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".[10]

A particularly relevant example is the OAIS Reference Model for an Open Archival Information System.[11] This model has shaped the Trusted Digital Repository (TDR) standard which frames repository best practice for ingesting, managing, and accessing archived digital objects. These have recently been exemplified by the DSA-WDS Catalogue of Requirements[12] and are particularly relevant for their emphasis on making workflows explicit.

Our specific concerns in the working group build on such standards, to guide implementation of quality assurance and peer review of research data objects, their citation, and linking with other digital objects in the research and scholarly communication environment.

A case study approach was in keeping with this aim. Case studies explore phenomena in their context and generalize to theory rather than to populations [21]. Similarly, drafting a conceptual model does not require us to make generalizable claims to the repository population as a whole, but it does commit us to testing its relevance to repositories, and other stakeholders, through community review and amendment.

As the membership of the RDA-WDS Publishing Data Workflows, WG was reasonably diverse in terms of disciplinary and stakeholder participation, we drew upon that group's knowledge and contacts, and issued calls to participate under the auspices of the RDA and WDS, in collaboration with the Force11 Implementation Group[13] to identify best practices and case studies in data-publishing workflows. Presentations and workshops at RDA plenary meetings were used to validate the approach and progress. With this iterative approach, we identified an initial set of repositories, projects, and publishing platforms which were thought to be reasonably representative of institutional affiliation and domain-specific or cross-disciplinary focus. These workflows served as a case study for the analysis to identify likely examples of 'data publishing' from repositories, projects, and publishing platforms, whether institutional, domain-specific, or cross-disciplinary.

Publicly available information was used to describe the workflows on a common set of terms. In addition, repository representatives were invited to present and discuss their workflows via videoconference and face-to-face meetings. Emphasis was given to workflows facilitating data citation and the provision of 'metrics' for data was added as a consideration. Information was organized into a comparison matrix and circulated to the group for review, whereupon a number of annotations and corrections were made. Empty fields were populated, where possible, and terms were cross-checked and harmonized across the overall matrix. Twenty-six examples were used for comparison of characteristics and workflows. However, one workflow (Arkivum) was judged not to qualify for the definition of 'data publishing' as it emerged in the

[10] Source: OASIS, http://www.oasis-open.org/committees/soa-rm/faq.php.

[11] "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.

[12] Draft available at: http://rd-alliance.org/group/repository-audit-and-certification-dsa-wds-partnership-wg/outcomes/dsa-wds-partership.

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

Table 1 Repositories, projects, and publishing platforms selected for analysis of workflows and other characteristics

Workflow provider name	Workflow provider type	Workflow provider specialist research area, if any	Deposit initiator
ENVRI reference model	Guidelines	Environmental sciences	Project-led
PREPARDE	Guidelines	Earth sciences	Researcher led (for *Geoscience Data Journal*)
Ocean Data Publication Cookbook	Guidelines	Marine sciences	Researcher-led
Scientific Data, Nature Publishing Group	Journal		Researcher (author) led
F1000Research	Journal	Life sciences	Researcher led; editorial team does a check
Ubiquity Press OHDJ	Journal	Life, health and social sciences	Researcher led
GigaScience	Journal	Life and biomedical sciences	Researcher (author) led
Data in Brief	Journal		Author led
Earth System Science Data Journal, Copernicus Publications	Journal	Earth sciences	Researcher led for data article. Researcher led for data submission to repository
Science and Technology Facilities Council Data Centre	Repository	Physics and space sciences	Researcher led as part of project deliverables
National Snow and Ice Data Center	Repository	Polar Sciences	Project or researcher led
INSPIRE Digital library	Repository	High energy Physics	Researcher led
UK Data Archive (ODIN)	Repository	Social sciences	Researcher led
PURR Institutional Repository	Repository		Researcher-/librarian led
ICPSR	Repository	Social and behavioural sciences	Researcher, acquisitions officer, and funder led
Edinburgh Datashare	Repository		Researcher led, librarian assists
PANGAEA	Repository	Earth sciences	Researcher led
WDC Climate	Repository	Earth sciences	Researcher or project led
CMIP/IPCC-DDC	Repository	Climate sciences	Project-led[a]
Dryad Digital Repository	Repository	Life sciences	Researcher led
Stanford Digital Repository	Repository		Researcher led
Academic Commons Columbia	Repository		Researcher and repository staff
Data Repository for the University of Minnesota (DRUM)	Repository		Researchers from institution
ARKIVUM and Figshare	Repository		Researcher led
OJS/ Dataverse data repository, all disciplines	Repository		Researcher led; part of journal article publication process

[a] Data Citation concept for CMIP6/AR6 is available as draft at: http://www.earthsystemcog.org/projects/wip/resources/

course of the research, so the final table consists of 25 entities (Table 1).

Workflows were characterized in terms of the discipline, function, data formats, and roles involved. We also described the extent to which each exhibited the following ten characteristics associated with data publishing:

- The assignment of persistent identifiers (PIDs) to datasets, and 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?
- Links to additional data products (data paper; review; other journal articles) or "standalone" product.
- Links to grant information, where relevant, and usage of author PIDs.
- Facilitation of data citation.
- Reference to a data life cycle model.
- Standards compliance.

The detailed information and categorization can be found in the analysis dataset comprising the comparison matrix [22].

4 Analysis and results: towards a reference model in data publishing

4.1 Definitions for data-publishing workflows and outputs

The review of the comparison matrix of data-publishing workflows produced by the RDA-WDS Publishing Data Workflows WG [22] revealed a need for standardization of terminology. We therefore propose definitions for six key terms: research data publishing, research data-publishing workflows, data journal, data article, data review, and data repository entry.

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."

This definition applies also to the publication of confidential and sensitive data with the appropriate safeguards and accessible metadata. A concrete example of such a workflow may be a published journal article that includes discoverability and citation of a dataset by identifying access criteria for reuse.[14] Harvard University is currently developing a tool that will eventually be integrated with Dataverse to share and use confidential and sensitive data in a responsible manner.[15]

Research data-publishing workflows

> Research data-publishing workflows are activities and processes 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 their 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).

Data article

> A data article is a 'data-publishing' product, also known as a 'data descriptor' that may appear in a data journal or any other journal. When publishers refer to 'data publishing', they usually mean a data article rather than the underlying dataset. Data articles focus on making data discoverable, interpretable, and reusable, rather than testing hypotheses or presenting new interpretations (by contrast with traditional journal articles). Whether linked to a dataset in a separate repository, or submitted in tandem with the data, the aim of the data article is to provide a formal route to data sharing. The parent journal may choose whether or how standards of curation, formatting, availability, persistence, or peer review of the dataset are described. By definition, the data article provides a vehicle to describe these qualities, as well as some incentive to do so. The length of such articles can vary from micro papers (focused on one table or plot) to very detailed presentation of complex datasets.

Data journal

> A data journal is a journal (invariably open access) that publishes data articles. The data journal usually provides templates for data description and offers researchers guidance on where to deposit and how to describe and present their data. Depending on the journal, such templates can be generic or discipline focused. Some journals or their publishers maintain their own repositories. As well as supporting bi-directional linking between a data article and its corresponding dataset(s), and facilitating persistent identification practices, data journals provide workflows for quality assurance(i.e. data peer review) and should also provide editorial guidelines on data quality assessment.

Data review

> Data review comprises a broad range of quality assessment workflows, which may extend from a technical review of metadata accuracy to a double-blind peer review of the adequacy of data files and documentation and accuracy of calculations and analyses. Multiple variations of review processes exist and are dependant upon factors such as publisher requirements, researcher expectations, or data sensitivity. Some workflows may be similar to traditional journal workflows, in which specific roles and responsibilities are assigned to editors and reviewers to assess and ensure the quality of a

[14] Indirect linkage or restricted access—see e.g. Open Health Data Journal, http://openhealthdata.metajnl.com.

[15] http://privacytools.seas.harvard.edu/datatags.

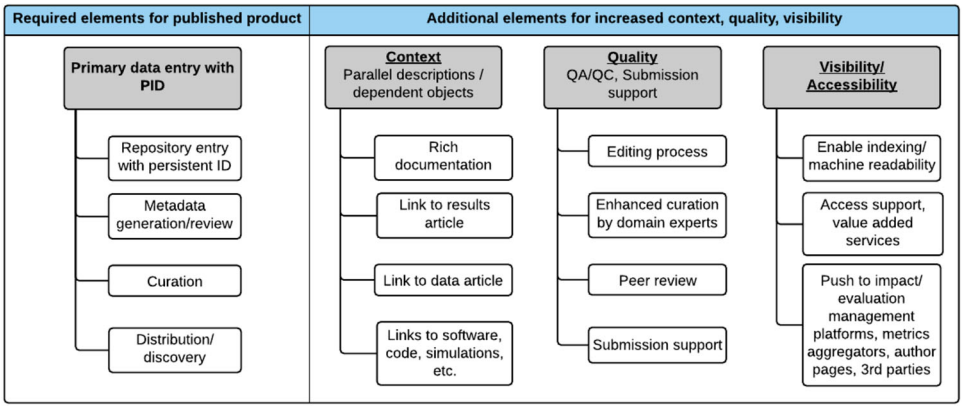

Fig. 1 Data-publishing key components. Elements that are required to constitute data publication are shown in the *left panel*, and optional services and functions in the *right panel*

data publication. The data review process may therefore encompass a peer review that is conducted by invited domain experts external to the data journal or the repository, a technical data review conducted by repository curation experts to ensure data are suitable for preservation, and/or a content review by repository subject domain experts.

Data repository entry

A data repository entry is the basic component of data publishing consisting of a persistent, unique identifier pointing to a landing page that contains a data description and details regarding data availability and the means to access the actual data [22]

4.2 Key components of data publishing

Analysis of workflows by the RDA-WDS data publishing WG identified the components that contribute to a generic reference model for data publishing. We distinguish basic and add-on services. The basic set of services consists of entries in a trusted data repository, including a persistent identifier, standardized metadata, and basic curation (Fig. 1).

Optional add-ons could include components such as contextualization through additional embedding into data papers or links to traditional papers. Some authors and solutions make a distinction between metadata publication and data publication. 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 for the correct use, understanding, and interpretation of the data.

Important add-ons are quality assurance/quality control (QA/QC)[16] and peer review services. Different variations of such services exist, ranging from author-led, editor-driven, librarian-supported solutions, to (open) peer review. Such components are crucial enablers of future data reuse and reproducible research. Our analysis found that many services offer or are considering offering such services. The third group of add-ons aims to improve visibility, as shown on the right panel of Fig. 1. This set of services is not currently well established and this hampers data reuse. Other emerging services include connection of data-publishing workflow with indexing services, research information services (CRIS), or metrics aggregators.

To ensure the possibility of data reuse, data publishing should contain at least the basic elements of curation, QA/QC, and referencing, plus additional elements appropriate for the use case (Fig. 1). Depending on the use case, however, it might be appropriate to select a specific set of elements from the key components (following some best practices). In the light of future reuse, we would argue that the basic elements of curation, QA/QC, and referencing should always be included.

4.3 Detailed workflows and dependencies

We present a traditional article publication workflow (Fig. 2-1), a reproducible research workflow (Fig. 2-2), and a data publication workflow (Fig. 2-3).

The workflow comparison found 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. Datasets may be published in a repository with or without an associated data article. 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 scientific and technical [18,23] quality assurance activities regarding dataset content, description, format, and metadata quality before data are archived for the long term. The typical data repository creates an entry for a specific dataset or a collection

[16] 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 (Research Data Canada, 2015, Glossary of terms and definitions, http://dictionary.casrai.org/Category:Research_Data_Domain).

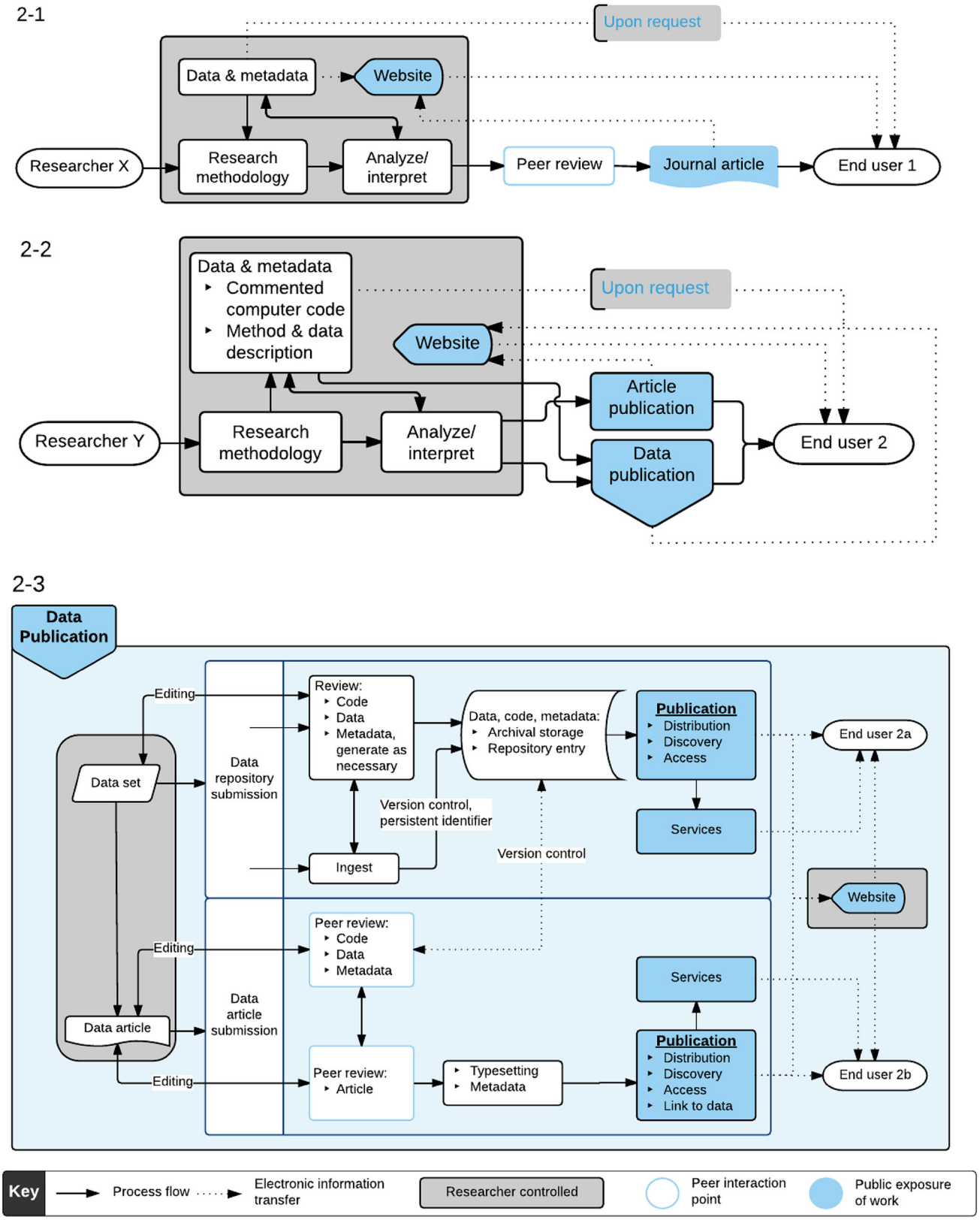

Fig. 2 Research data publication workflows. We present a traditional article publication workflow (**2-1**), a reproducible research workflow (**2-2**), and—as a more dynamic version of Fig. 1—a data publication workflow (**2-3**)

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 [24].[17] Some repositories facilitate third-party access for discoverability or metrics services.

As shown in Fig. 2, researchers can and do follow a number of different pathways to communicate about their data. Traditionally, research results are published in journals, and readers (end user 1) interested in the data would need to contact authors to access underlying data or attempt to access it from a researcher-supported website (Fig. 2-1). Emerging processes supporting greater reproducibility in research include some form of data publication (Fig. 2-2). This includes the special case of standalone[18] data publications with no direct connection to a paper. These are common in multiple domain areas (e.g. the large climate data intercomparison study CMIP).[19] Figure 2-3 illustrates the two predominant emerging data publication workflows emerging from our analysis: (a) submission of a dataset to a repository; and, (b) submission of a data article to a data journal. Both workflows require that datasets are submitted to a data repository.

The data publication process shown in Fig. 2-3 may be initiated at any time during research once the data are sufficiently complete and documented, and may follow a variety of paths. A repository will typically provide specific templates for metadata and additional documentation (e.g. methodology or code-specific metadata). The submission may then be reviewed from a variety of perspectives depending on the policies and practices of the repository. These review processes may include formatting issues, content, metadata, or other technical details. Some repositories may also require version control of the dataset. There is a great deal of variability between repositories in the type of data accepted, available resources, the extent of services offered, and workflows. Figure 2-3 illustrates the elements common to the workflows of the data repositories selected for the present study (Fig. 2-3) are consistent with those shown in Fig. 1.

A researcher may also choose to initiate the data publication process by submitting a data article for publication in a data journal. This workflow is also illustrated in Fig. 2-3, and while it is in part dependent on data repositories (data journals typically identify approved repositories),[20] the data article publication process has the opportunity to more consistently provide some of the advantages of data publication as represented in the 'Additional elements' of Fig. 1. Data journals are similar to the traditional research journal (Fig. 2-1), in that their core processes consist of peer review and dissemination of the datasets. Naturally, reviewers must have pre-publication access to the dataset in a data repository, and there needs to be version control solutions for datasets and data papers. Whether publishing data via a data article or a data repository, both workflows have the potential to be incorporated into the current system of academic assessment and reward in an evolutionary process rather than a disruptive departure from previous systems.

Data publication workflows supporting reproducible research give end users access to managed and curated data, code, and supporting metadata that have been reviewed and uploaded to a trusted repository (Fig. 2, end-user 2a). If an associated data article is published, end users will also have further contextual information (Fig. 2, end-user 2b). The traditional journal article may be published as usual and may be linked to the published data and/or data article as well. There are some hard-wired automated workflows for data publishing (e.g. with the Open Journal Systems-Dataverse integration [25], or there can be alternate automated or manual workflows in place to support the researcher (e.g. Dryad).

4.4 Data deposit

We found that a majority of data deposit mechanisms underlying data-publishing workflows are initiated by researchers, but their involvement beyond the initial step of deposition varied across repositories and journals. 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), screens submissions and assesses the levels of metadata and support required. Data journals, however, typically adopt a 'hands-off' approach: the journal is the 'publication' outlet, but the data are housed elsewhere. Hence, the journal publishing team often relies on external parties—repository managers and the research community in general[21] —to manage data deposit and to assess whether basic standards are met for data deposition or if quality standards are met (see details below).

[17] http://www.force11.org/datacitationimplementation.

[18] Defined in e.g. [18].

[19] Program for Climate Model Diagnosis and Intercomparison. (n.d.). Coupled Model Intercomparison Project (CMIP). Retrieved November 11, 2015, from http://www-pcmdi.llnl.gov/projects/cmip/.

[20] Approved by the data journal.

[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 and the latter journal's website: http://www.earth-system-science-data.net/peer_review/interactive_review_process.html.

4.5 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 given the greater diversity of use cases and practice around data from diverse disciplines. Some discipline-specific repositories have multiple-stage processes including several QA/QC processes and workflows based on OAIS. Many IRs have adopted a broader approach to ingest necessitated by their missions, which involves archiving research products generated across their campuses, especially those found in the long tail of research data, including 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 improve.

When data journals rely on external data repositories to handle the actual data curation, there needs to be a strong collaboration between the journal and repository staff beyond trust that the repository will pursue data management and ingestion according to acceptable standard procedures. Data journals and data repositories are encouraged to make public and transparent any such agreements (e.g. service-level agreements). Ultimately, however, this level of one-to-one interaction is not scalable and automated procedures and repository standards will be needed.

4.6 Quality assurance (QA) and quality control (QC)

We found that QA/QC typically occurs at three points during the data-publishing workflow: (1) during data collection and data processing, prior to submission of the data to a repository; (2) during submission and archiving of the data; and (3) during a review or the editorial procedure. We distinguish between traditionally understood peer review and the internal reviews that repositories and journals also generally conduct (Fig. 2), which may touch on content, format, description, documentation, metadata, or other technical details.

QA/QC procedures vary widely and may involve authors/reviewers for QA of the content and documentation, and data managers/curators, librarians, and editors for technical QA. Quality criteria can include checks on data, metadata, and documentation against repository, discipline.[22] and project standards.

Most repositories and all of the data journals that we reviewed had some QA/QC workflows, but the level and type of services varied. Established data repositories (e.g. ICPSR or Dataverse [22]) tended to have dedicated data curation personnel to help in standardizing and reviewing data upon submission and ingestion, especially in the area of metadata. Some domain repositories (e.g. ICPSR) go farther and conduct in-depth quality control checks on the data, revising the data if necessary in consultation with the original investigator. Other repositories responsible for the long-term archiving of project data (e.g. the IPCC-DDC[23]) document their QA results. Some data repositories rely on researchers for the QA/QC workflows to validate the scientific aspects of data, metadata, and documentation. Technical support, data validation, or QA/QC was also done by some repositories, but the level of engagement varied with the service and the individual institutions: some checked file integrity, while others offered more complex preservation actions, such as on-the-fly data format conversions. Some multi-purpose repositories provided support to researchers for QA/QC workflows, but this was not a standard practice. Overall, QA/QC in data publishing is a 'hot-button' topic and is debated heavily and continuously within the community. Mayernik et al. describe a range of practice in technical and academic peer review for publishing data [26].

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, and also between publisher and data repository in workflow co-ordination, versioning, and consistency.

Given the wide range of QA/QC services currently offered, future recommendations should consider the following:

- 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 reuse. This also simplifies the peer review process for associated data journals and lowers barriers to uptake by researchers.
- General research data repositories which must accommodate a wide variety of data may have some limitations in QA/QC workflows and these should be made explicit.
- Information about quality level definitions and quality assessment procedures and results should be explicit and readily available to users (and also possibly to third parties, such as aggregators or metric services).

There appears to be a trend towards data being shared earlier in the research workflow, at a stage where the data are still

[22] An example for a discipline standard is the format and metadata standard NetCDF/CF used in Earth system sciences: http://cfconventions.org/.

[23] Intergovernmental Panel on Climate Change Data Distribution Centre (IPCC-DDC): http://ipcc-data.org.

dynamic (see for example Meehl et al. [27]). There is a need, therefore, for QA/QC procedures that can handle dynamic data.

4.7 Data administration and long-term archiving

Data administration and curation activities may include dealing with a variety of file types and formats, creation of access-level restrictions, the establishment and implementation of embargo procedures, and assignment of identifiers. We found an assortment of practices in each of these areas. These vary from providing file format guidelines alone 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. Several discipline-specific repositories already have a long track record of preserving data and have detailed workflows for archival preservation. Other repositories are fairly new to this discussion and continue to explore potential solutions.

Most repositories in our sample have indicated a commitment to persistence and the use of standards. The adoption of best practices and standards would increase the likelihood that published data will be maintained over time and lead to interoperable and sustainable data publishing. Repository certification systems 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. The range of certification schemes[24] includes those being implemented by organizations such as the Data Seal of Approval (DSA)[25] and the World Data System (ICSU-WDS).[26] Improved adoption of such standards would have a big impact on interoperable and sustainable data publishing.

4.8 Dissemination, access, and citation

Data packages in most repositories we analyzed 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 recognition that data must be citable if it is to be optimally useful.[27] 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 culture shift needed 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. The launch of the Data Citation Index[28] by Thomson Reuters and projects such as the *Data Discovery Index*[29] are working on addressing the important challenge of data discovery and could serve as an accelerator to a paradigm shift for establishing data publishing within research 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 fairly comprehensive combination of data, metadata, and platforms to work with. Eighteen months later, the project was able to report both a cultural shift in researcher behaviour and a significant increase in the potential reproducibility of relevant research. As discussed in Bandrowski et al. [28], 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. This provided a compelling rationale, coherent journal policies (which necessitated compliance for would-be authors to publish), and a specific workflow for the registration process (complete with skilled, human support if required). Further work is needed to determine exactly how this use case can be leveraged across the wider gamut of subjects, communities, and other players.

FAIR principles[30] and other policy documents [10] explicitly mention that data should be accessible. Data-publishing solutions ensure that this is the case, but some workflows allow only specific users to access sensitive data. An exam-

[24] 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.

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

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

[27] 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 the data accessible and reusable, a key to changing the culture around scholarly credit for research data.

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

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

[30] http://www.force11.org/group/fairgroup/fairprinciples.

ple is survey data containing information that could lead to the identification of respondents. In such cases, a prospective data user could access the detailed survey metadata to determine if meets his/her research needs, but a data use agreement would need to be signed before access to the dataset would be granted. The metadata, data article, or descriptor could be published openly, perhaps with a Creative Commons license, but the underlying dataset would be unavailable except via registration or other authorization processes. In such a case, the data paper would allow contributing researchers to gain due credit, and it would facilitate data discovery and reuse.[31]

Citation policies and practice also vary by community and culture. Increasingly, journals and 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*[32] 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.[33]

4.9 Other potential value-added services and metrics

Many repository or journal providers look beyond workflows that gather information about the research data and also want to make this information visible to other information providers in the field. This can add 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 (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 researcher. In addition, services such as ORCID.[34] 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 our investigation, 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 shared materials can enrich the content on the original platforms and encourage users to engage in further data sharing or curation activities. Such information is certainly of interest to infrastructure and research funders.[35]

4.10 Diversity in workflows

While workflows may appear to be fairly straightforward and somewhat similar to traditional static publication procedures, the underlying processes are, in fact, quite complex and diverse. The diversity was most striking in the area of curation. 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 QA had more complex workflows with additional steps, possibly consecutive. They might facilitate more collaborative work at the beginning of the process, or include standardized preservation steps.

There was metadata heterogeneity across discipline-specific repositories. Highly specialized repositories frequently focused on specific metadata schemas and pursued curation accordingly. Some disciplines have established metadata standards, similar to the social sciences' use of the Data Documentation Initiative standard.[36] In contrast, more general repositories tended to converge on domain-agnostic metadata schemas with fields common across disciplines, e.g. the mandatory DataCite fields.[37]

Data journals are similar in overall workflows, but differ in terms of levels of support, review, and curation. As with repositories, the more specialized the journal (e.g. a discipline in the earth sciences with pre-established data-sharing practices), the more prescriptive are the author guidelines and the more specialized the review and QA processes. With the rise of open or post-publication peer review, some data journals are also inviting the wider community to participate in the publication process.

The broader research community and some discipline-based communities are currently developing criteria and practices for standardized release of research data. The services supporting these efforts, whether repositories or journals, also generally show signs of being works in progress

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

[32] Data Citation Synthesis Group, 2014. Accessed 17 November 2015: http://www.force11.org/group/joint-declaration-data-citation-principles-final.

[33] 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.

[34] http://orcid.org/.

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

[36] Accessed 17 November 2015: http://www.ddialliance.org.

[37] Accessed 17 November 2015: http://schema.datacite.org.

or proof-of-concept exercises rather than finished products. This is reflected in our analysis dataset [22]. Depending partly on their state of progress during our review period (1 February–30 June 2015), and also on the specificity of the subject area, some workflow entries were rather vague.

5 Discussion

Although the results of our analysis show wide diversity in data-publishing workflows, the key components were fairly similar across providers. The common components were grouped and charted in a reference model for data publishing. Given the rapid developments in this field and in light of the disciplinary differences, diversity of workflows might be expected to grow even further. Through the RDA Working Group we will seek further community review and endorsement of the generic reference model components and carry out further analyses of such disciplinary variations. However, the results of our study suggest that new solutions (e.g. for underrepresented disciplines) could build on the identified key components that best match their use case. Some evident gaps and challenges (described below) hinder global interoperability and adoption of a common model.

5.1 Gaps and challenges

Whilst our analysis extended across all the data-publishing entities we studied (repositories, journals, and projects), many of the most obvious gaps and challenges were observed amongst the repository category.

While there are still many disciplines for which no specific domain repositories exist, we are seeing a greater number of repositories of different types (re3data.org indexes over 1200 repositories). In addition to the disciplinary repositories, there are many new repositories designed to house broader collections, e.g. Zenodo, Figshare, Dryad, Dataverse, and the institutional repositories at colleges and universities. "Staging" repositories are also being established that extend traditional workflows into the collaborative working space—e.g. Open Science Framework[38] which has a publishing workflow with Dataverse. Another example is the Sustainable Environment Actionable Data (SEAD)[39] project, which provides project spaces in which scientists manage, find, and share data, and which also connects researchers to repositories that will provide long-term access and preservation of data.

Despite much recent data-publishing activity, our analysis of the case studies found that challenges remain, in particular when considering more complex workflows. These include:

- *Bi-directional linking* How do we link data and publications persistently in an automated way? Several organizations, including RDA and WDS,[40] are now working on this problem. A related issue is the persistence of links themselves.[41]
- *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 is needed to establish a wider framework, including community updating and initiatives involving linking to associated data .
- *Version control* In general, we found that repositories handle version control in different ways, which is potentially confusing. While some version control solutions might be tailored to discipline-specific challenges, there is a need to standardize. This issue also applies to provenance information.
- *Sharing restricted-use data* Repositories and journals are generally 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 expertise in this area.
- *Role clarity* Data publishing relies on collaboration. For better user guidance and greater confidence in the services, an improved understanding of roles, responsibilities, and collaboration is needed. Documentation of '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 and Houghton[42] produced a synthesis of data centre studies combining quantitative and qualitative approaches to quantify value 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.[43] However, much more work is needed to understand viable financial models for publishing data[44] and to distinguish trustworthy collaborations.
- *Data citation support* Although there appears to be widespread awareness, there is only partial implementation of the practices and procedures recommended by

[38] https://osf.io/.

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

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

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

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

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

[44] RDA/WDS Publishing Data Costs IG addresses this topic: http://rd-alliance.org/groups/rdawds-publishing-data-ig.html.

the Data Citation Implementation Group. There is a wide range of PIDs emerging, including ORCID, DOI, FundRef, RRID, IGSN, ARK, and many more. Clarity and ease of use need to be brought to this landscape.[45]

- *Metrics* Creators of data and their institutions and funders need to know how, and how often, their data are being reused.
- *Incentives* Data publishing offers potential incentives to researchers, e.g. a citable data product, persistent data documentation, and information about the impact of the research. Also, many repositories offer support for data submission. Benefits of data publishing need to be better communicated to researchers. In addition, stakeholders should disseminate the fact that formal data archiving results in greater numbers of papers and thus more science, as Piwowar and Vision, and Pienta et al. [4,5] have shown. There should also be increased clarity with respect to institutional and funder recognition of the impact of research data.

The challenges of more complex data—in particular, big data and dynamic data—need also to be addressed. Whereas processes from the past 10 years focus on irrevocable, fully documented data for unrestricted (research) use, data publishing needs to be 'future proof' (Brase et al. [29]). There is a requirement from research communities[46] to cite data before it has reached an overall irrevocable state and before it has been archived. This particularly holds true for communities with high volume data (e.g. high-energy physics; climate sciences), and for data citation entities including multiple individual datasets for which the time needed to reach an overall stable data collection is long. Even though our case study analysis found that data citation workflows are implemented or considered by many stakeholder groups involved in data publishing, dynamic data citation challenges have not been widely addressed. Version control and keeping a good provenance record[47] of datasets are also critical for citation of such data collections and are indispensable parts of the data-publishing workflow.

With respect to gaps and challenges, we recognize that the case studies we analyzed are limited in scope. This relates to an overall challenge we encountered during the project: it is difficult to find clear and consistent human-readable workflow representations for repositories. The trust standards (e.g. Data Seal of Approval,[48] Nestor, ISO 16363 and World Data System) require that repositories document their processes, so this may change in the future, but we would add our recommendation that repositories publish their workflows in a standard way for greater transparency. This would bolster confidence in repositories and also increase user engagement.

The diversity we found is not surprising, nor is it necessarily undesirable. Case studies and ethnographies of data practices have found that workflows for dealing with data 'upstream' of repositories are highly diverse. Data sharing practices vary considerably at the sub-disciplinary level in many cases (e.g. Cragin et al. [30]), so there is likely to be continued need to support diverse approaches and informed choice rather than unified or monolithic models (Pryor [31]). Our analysis shows that a variety of workflows has evolved, and more are emerging, so researchers may be able to choose their best fit on the basis of guidance that distinguishes relevant features, such as QA/QC and different service or support levels.

5.2 Best practice recommendations and conclusions

Based on selected case studies, key components in data publishing have been identified, leading to a reference model in data publishing. The analysis, and in particular the conversations with the key stakeholders involved in data-publishing workflows, highlighted best practices which might be helpful as recommendations for organizations establishing new workflows and to those seeking to transform or standardize existing procedures:

- Start small and build components one by one in a modular way 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.
- 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 researcher PIDs such as ORCID can also establish connections between data and papers or other research entities such as software.

[45] http://project-thor.eu/.

[46] 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 dataset.

[47] 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 are crucial for citations of such data collections. Technical solutions are being developed, e.g. by the European Persistent Identifier Consortium (EPIC).

[48] http://datasealofapproval.org.

The use of PIDs can also enable linked open data functionality.[49]

- Document roles, workflows and services. A key difficulty we had in conducting the analysis of the workflows was 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.

In summary, following the idea of the presented reference model and the best practices, 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 and to enable reproducible research. This includes linkages between documentation, code, data, and journal articles in an integrated environment. Furthermore, in the ideal workflow, all of these objects need to be well documented to enable other researchers (or citizen, scientists, etc.) to reuse the data for new discoveries. We would like to see information standardized and exposed via APIs and other mechanisms so that metrics on data usage can be captured. We note, however, that biases in funding and academic reward systems need value data-driven secondary analysis and reuse of existing data, as well as data publishing as a first class object. More attention (i.e. more perceived value) from funders will be key to changing this paradigm.

One big challenge is that there is a need to collaborate more intensively among the stakeholder groups. For example, repositories and higher education institutions (holding a critical mass of research data) and the large journal publishers (hosting the critical mass of discoverable, published research) have not yet fully engaged with each other. Although new journal formats are being developed that link data to papers and enrich the reading experience, progress is still being impeded by cultural, technical, and business model issues.

We have demonstrated that the different components of a data-publishing system need to work, where possible, in a seamless fashion and in an integrated environment. We therefore advocate the implementation of standards, and the development of new standards where necessary, for repositories and all parts of the data-publishing process. Data publishing should be embedded in documented workflows, to help establish collaborations with potential partners and to guide researchers, enabling and encouraging the deposit of reusable research data that will be persistent while preserving provenance.

[49] At the time of writing, CrossRef had recently announced the concept and approximate launch date for a 'DOI Event Tracker', which could also have considerable implications for the perceived value of data publishing as well as for the issues around the associated metrics (Reference: http://crosstech.crossref.org/2015/03/crossrefs-doi-event-tracker-pilot.html by Geoffrey Bilder, accessed 26 October 2015).

References

1. Schmidt, B., Gemeinholzer, B., Treloar, A.: Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey (2015). http://docs.google.com/document/d/1jRM5ZlJ9o4KWIP1GaW3vOzVkXjIIBYONFcd985qTeXE/ed
2. Vines, T.H., Albert, A.Y.K., Andrew, R.L., DeBarre, F., Bock, D.G., Franklin, M.T., Gilbert, K.J., Moore, J.S., Renaut, S., Rennison, D.J.: The availability of research data declines rapidly with article age. Curr. Biol. **24**(1), 94–97 (2014)
3. Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., Rafols, I.: Bibliometrics: The Leiden Manifesto for research metrics. Nature **520**, 429–431 (2015). http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351. Accessed 10 November 2015
4. Piwowar, H., Vision, T.: Data reuse and the open data citation advantage. PeerJ Comput. Sci. (2013). http://peerj.com/articles/175/. Accessed 10 November 2015
5. Pienta, A.M., Alter, G.C., Lyle, J.A.: The enduring value of social science research: the use and reuse of primary research data (2010). http://hdl.handle.net/2027.42/78307. Accessed 10 November 2015
6. Borgman, C.L.: Big data, little data, no data: scholarship in the networked world. MIT Press, Cambridge (2015)
7. Wallis, J.C., Rolando, E., Borgman, C.L.: If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One **8**(7), e67332 (2013). doi:10.1371/journal.pone.0067332
8. Peng, R.D.: Reproducible research in computational science. Science **334**(6060), 1226–1227 (2011)
9. Thayer, K.A., Wolfe, M.S., Rooney, A.A., Boyles, A.L., Bucher, J.R., Birnbaum, L.S.: Intersection of systematic review methodology with the NIH reproducibility initiative. Environ. Health Perspect. **122**, A176–A177 (2014). http://ehp.niehs.nih.gov/wp-content/uploads/122/7/ehp.1408671.pdf. Accessed 10 November 2015
10. George, B.J., Sobus, J.R., Phelps, L.P., Rashleigh, B., Simmons, J.E., Hines, R.N.: Raising the bar for reproducible science at the US Environmental Protection Agency Office of Research and Development. Toxicol. Sci. **145**(1), 16–22 (2015). http://toxsci.oxfordjournals.org/content/145/1/16.full.pdf+html
11. Boulton, G., et al.: Science as an open enterprise. R. Soc. Lond. (2012). https://royalsociety.org/policy/projects/science-public-enterprise/Report/. Accessed 10 November 2015
12. Stodden, V., Bailey, D.H., Borwein, J., LeVeque, R.J., Rider, W., Stein, W.: Setting the default to reproducible. Reproducibility in computational and experimental mathematics. Institute for Computational and Experimental Research in Mathematics (2013). http://icerm.brown.edu/tw12-5-rcem/icerm_report.pdf. Workshop report accessed 10 November 2015
13. Whyte, A., Tedds, J.: Making the case for research data management. DCC briefing papers. Digital Curation Centre, Edinburgh (2011). http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm. Accessed 10 November 2015
14. Parsons, M., Fox, P.: Is data publication the right metaphor? Data Sci. J. **12** (2013). doi:10.2481/dsj.WDS-042. Accessed 10 November 2015
15. Rauber, A., Pröll, S.: Scalable dynamic data citation approaches, reference architectures and applications RDA WG Data Citation

position paper. Draft version (2015). http://rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html. Accessed 13 November 2015
16. Rauber, A., Asmi, A., van Uytvanck, D., Pröll, S.: Data citation of evolving data: recommendations of the Working Group on Data Citation (WGDC) Draft—request for comments (2015). Revision of 24th September 2015. http://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150924.pdf. Accessed 6 November 2015
17. Watson, et al.: The XMM-Newton serendipitous survey. V. The Second XMM-Newton serendipitous source catalogue. Astron. Astrophys. **493**(1), 339–373 (2009). doi:10.1051/0004-6361:200810534
18. Lawrence, B., Jones, C., Matthews, B., Pepler, S., Callaghan, S.: Citation and peer review of data: moving toward formal data publication. Int. J. Digital Curation (2011). doi:10.2218/ijdc.v6i2.20r
19. Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M.S., Whyte , A.: Processes and procedures for data publication: a case study in the geosciences. Int. J. Digital Curation **8**(1) (2013). doi:10.2218/ijdc.v8i1.253
20. Austin, C.C., Brown, S., Fong, N., Humphrey, C., Leahey, L., Webster, P.: Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements. Presented at the IASSIST Annual Conference. IASSIST Quarterly Preprint. International Association for Social Science, Information Services, and Technology. Minneapolis (2015). http://drive.google.com/file/d/0B_SRWahCB9rpRF96RkhsUnh1a00/view. Accessed 13 November 2015
21. Yin, R.: Case study research: design and methods, 5th edn. Sage Publications, Thousand Oaks (2003)
22. Murphy, F., Bloom, T., Dallmeier-Tiessen, S., Austin, C.C., Whyte, A., Tedds, J., Nurnberger, A., Raymond, L., Stockhause, M., Vardigan, M.: WDS-RDA-F11 Publishing Data Workflows WG Synthesis FINAL CORRECTED. Zenodo. **2015** (2015). doi:10.5281/zenodo.33899. Accessed 17 November 2015
23. Stockhause, M., Höck, H., Toussaint, F., Lautenschlager, M.: Quality assessment concept of the World Data Center for Climate and its application to the CMIP5 data. Geosci. Model Dev. **5**(4), 1023–1032 (2012). doi:10.5194/gmd-5-1023-2012
24. Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R.R., Duerr, R., Haak, L.L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J.E., Lin, J., Nielsen, L.H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M., Clark, T.: Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Comput. Sci. **1**(e1) (2015). doi:10.7717/peerj-cs.1
25. Castro, E., Garnett, A.: Building a bridge between journal articles and research data: The PKP-Dataverse Integration Project. Int. J. Digital Curation **9**(1), 176–184 (2014). doi:10.2218/ijdc.v9i1.311
26. Mayernik, M.S., Callaghan, S., Leigh, R., Tedds, J.A., Worley, S.: Peer review of datasets: when, why, and how. Bull. Am. Meteorol. Soc. **96**(2), 191–201 (2015). doi:10.1175/BAMS-D-13-00083.1
27. Meehl, G.A., Moss, R., Taylor, K.E., Eyring, V., Stouffer, R.J., Bony, S., Stevens, B.: Climate Model Intercomparisons: preparing for the next phase. Eos Trans. AGU **95**(9), 77 (2014). doi:10.1002/2014EO090001
28. Bandrowski, A., Brush, M., Grethe, J.S., Haendel, M.A., Kennedy, D.N., Hill, S., Hof, P.R., Martone, M.E., Pols, M., Tan, S., Washington, N., Zudilova-Seinstra, E., Vasilevsky, N.: The Resource Identification Initiative: a cultural shift in publishing [version 1; referees: 2 approved] *F1000Research***4**, 134 (2015). doi:10.12688/f1000research.6555.1
29. Brase, J., Lautenschlager, M., Sens, I.: The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite. D-Lib Mag. **21**(1/2) (2015). doi:10.1045/january2015-brase
30. Cragin, M.H., Palmer, C.L., Carlson, J.R., Witt, M.: Data sharing, small science and institutional repositories. Philos. Trans. R. Soc. A **368**(1926), 4023–4038 (2010)
31. Pryor, G.: Multi-scale data sharing in the life sciences: Some lessons for policy makers. Int. J. Digital Curation **4**(3), 71–82 (2009). doi:10.2218/ijdc.v4i3.115