IJDC | *General Article*

Connecting Data Publication to the Research Workflow: A Preliminary Analysis

Sünje Dallmeier-Tiessen
CERN

Varsha Khodiyar
Springer Nature

Fiona Murphy
University of Reading

Amy Nurnberger
Columbia University

Lisa Raymond
Woodes Hole Oceanographic Institution

Angus Whyte
Digital Curation Centre

Abstract

The data curation community has long encouraged researchers to document collected research data during active stages of the research workflow, to provide robust metadata earlier, and support research data publication and preservation. Data documentation with robust metadata is one of a number of steps in effective data publication. Data publication is the process of making digital research objects 'FAIR', i.e. findable, accessible, interoperable, and reusable; attributes increasingly expected by research communities, funders and society. Research data publishing workflows are the means to that end. Currently, however, much published research data remains inconsistently and inadequately documented by researchers. Documentation of data closer in time to data collection would help mitigate the high cost that repositories associate with the ingest process. More effective data publication and sharing should in principle result from early interactions between researchers and their selected data repository. This paper describes a short study undertaken by members of the Research Data Alliance (RDA) and World Data System (WDS) working group on Publishing Data Workflows. We present a collection of recent examples of data publication workflows that connect data repositories and publishing platforms with research activity 'upstream' of the ingest process. We re-articulate previous recommendations of the working group, to account for the varied upstream service components and platforms that support the flow of contextual and provenance information downstream. These workflows should be open and loosely coupled to support interoperability, including with preservation and publication environments. Our recommendations aim to stimulate further work on researchers' views of data publishing and the extent to which available services and infrastructure facilitate the publication of FAIR data. We also aim to stimulate further dialogue about, and definition of, the roles and responsibilities of research data services and platform providers for the 'FAIRness' of research data publication workflows themselves.

Received 20 October 2016 ~ *Revision received* 23 February 2017 ~ *Accepted* 23 February 2017

Correspondence should be addressed to Angus Whyte, Digital Curation Centre, Argyle House Floor F West, 3 Lady Lawson Street, Edinburgh, EH3 9DH. Email: a.whyte@ed.ac.uk

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Background and Context

The data curation community has long encouraged researchers to document their collected research data during each active stage of the research workflow, to provide robust metadata earlier, and support research data publication (e.g. Frey, De Roure and Carr, 2002; Wallis et al., 2008). A great deal of work has been undertaken in the digital preservation community to improve preservation planning through early and effective interaction between data producers and archives (e.g. Farquar and Hockx-Yu, 2008; Schmidt et al., 2010; Waddington et al., 2012). Similar motivations led to the development of the Producer-Archive Interface Methodology Abstract Standard (PAIMAS) whose purpose is to structure the submission agreements between a producer and an archive (Huc et al., 2003).

This article is similarly concerned with digital object management 'upstream' of repositories or archives, but in the more specific context of research data publication. The authors are members of a joint working group of the Research Data Alliance and World Data System, whose objectives are 'to provide an analysis of a representative range of existing and emerging workflows and standards for data publishing… and provide reference models and implementations for application in new workflows.' To that end, the Publishing Data Workflows Working Group defines research data publication as:

> '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' (Austin et al., 2016).

This definition is consistent with the steps needed to make digital research objects 'FAIR', i.e. findable, accessible, interoperable, and reusable; and these attributes are increasingly expected by research communities, funders and society (Wilkinson et al., 2016). There are overlapping concerns between 'research data publication' and Open Access, as both seek to minimise the legal barriers to reuse. However, research data publication does not, per se, imply publication under any particular OA licensing regime.

The activities and processes involved in research data publication also overlap with those for preservation. Key areas of overlap include the provision of persistent access and, upstream of that, in trying to ensure data producers provide data documentation with robust metadata (Austin et al., 2016).

Workflows to support the processing, analysis and archiving of research data have been the subject of eScience research and development for some time (e.g. Gil et al., 2007), and a number of scientific workflow management tools are available, for example MyExperiment (Goble and De Roure, 2007) and Keppler (Ludascher et al., 2006). Despite this, and the maturation of digital preservation as a discipline, recent

surveys indicate that many researchers do not deposit data in repositories at all (e.g. Kowalczyk, 2014; Van den Eynden et al., 2016).

Technologies to support research data workflows have nevertheless proliferated, as have the actors involved. Various metaphors have been used to frame these and the attendant challenges, for example as an 'ecosystem', and these have in turn informed new models for dealing with data (Parsons and Fox, 2013). Data publication is one such metaphor that has gained currency, and led to calls for 'a novel publishing paradigm', where 'publishing' is defined as making a product online available, discoverable, peer-reviewable, re-usable according to given rights, real-time accessible, citable, and interlinked with its research activity and associated products' (Assante et al., 2015).

The Problem: Connecting Upstream Workflows to Data Publication

Currently much published research data remains inconsistently and inadequately documented by researchers (e.g. Tenopir et al., 2011). Researchers often miss the opportunity to capture accurate and sufficient metadata during the data generation phase (Jahnke, Asher and Keralis, 2012). That opportunity for better documented, and thus more reusable, research data is also an opportunity for repositories to make cost efficiencies. Documentation of data closer in time to data collection would help mitigate the high cost that repositories associate with the ingest process (Beagrie, Lavoie and Woollard, 2008-2010).

Against this background it is important to understand how the intention to 'publish' research data influences decisions earlier in the research workflow. We mean by this not only the executable components of such workflows, but the decision-making to enable data publication. This entails a broad view of workflows that considers the organisational process and policy context. Our definition is as follows;

> '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' (Austin et al., 2016).

The above definition is, however, part of a reference model intended to support interlinking of repositories and other platforms used 'downstream' in the research cycle, i.e. as finalised outputs are publicly shared. Further upstream towards data production, decisions affecting the final published products may involve a range of stakeholders in performing or facilitating data preparation (e.g. data producer, research administrator, research data support service, repository, publisher etc.) Ideally, this preparatory work would be performed in a continuous, considered, and consistent manner, close to the point of data collection, to ensure a full record of the provenance of the digital object throughout its journey from source to publication. The relevant preparatory steps include:

- Assignment of persistent identifiers (PIDs) to datasets, code, models etc;
- Creation of metadata to support data citation and discovery;
- Adoption of recognised metadata standards;

- Data documentation e.g. describing data using both domain-relevant and generalised terminology so that others may understand how and why the data, code, models, etc were produced;
- Linking research data documentation to author PIDs (e.g. ORCID) and, where relevant grant information;
- Linking research data documentation to other research products e.g. data management plan, data paper, journal article;
- Technical review, e.g. describing cleaning, de-identification, or quality assurance;
- Peer review of data, e.g. by researchers or by editorial reviewers.

While many of these tasks have been researched and practiced for many years in the data preservation and open access repository communities, there have been few empirical studies of them in the data publishing context. Although the term 'data publishing' promises research data producers and users some added value by linking across platforms and providers to give their curated digital objects more context, it is not yet clear how that promise influences the flow of metadata from its source.

Aims of the Study

The authors' review of a selection of research workflows aimed to identify connections between the goal of research data publication and the incorporation of preparatory steps into the research lifecycle. We consider a number of points when that exchange of metadata and identifiers is likely to happen. These include data management planning, data collection, creation, analysis, and use of data; data selection and access decisions; resolving ethical issues through de-identification, and publication (Addis, 2015).

Our review builds on the four recommendations of the RDA/WDS Publishing Data Workflows Working Group, which resulted from a review of data publishing workflows (Austin et al., 2016). Those recommendations were aimed at repositories and providers of other 'downstream' or end-of-research-lifecycle publication services, and were as follows:

1. Start small, building workflows from modular, open source and shareable components;
2. Follow standards that facilitate interoperability and permit extensions;
3. Facilitate data citation, e.g. through use of digital object PIDs, data/article/person/software linkages, researcher PIDs;
4. Document roles, workflows and services.

In this review we reflect on how these recommendations, initially targeted at the end of the research cycle, may be adapted to better reflect early preparation of data for publishing in the more dynamic context of research processes, with their diversity of tools, platforms, and disciplinary practices.

Method and Results

Collecting Examples of Upstream Workflow Solutions

To help reflect on the applicability of our recommendations to upstream workflows, as a first step we collected examples of these workflows from members of the RDA/WDS Publishing Data Workflows Working Group and participants in workshops held during RDA Plenary Meetings. These were supplemented through a call for participation, disseminated via the RDA/WDS Working Group listserv[1] and website. This resulted in a collection of 12 examples listed below. To gauge how extensively these examples cover the research data management process we applied a classification adopted in Addis (2015). We then identified characteristics of workflows that we believed demonstrated aspects of the recommendations in Austin et al. (2016).

Table 1. Integrated research data publishing workflows.

Workflow name	Workflow type addressed	Source document /contributors
CERN Analysis Preservation	Collection and processing, Selection, Publication	Dallmeier-Tiessen et al (2014)
Electronic lab notebook to data repository (RSpace to DataShare)	Collection and processing, Selection, Publication	Ward, MacNeil and Whyte; response in Dallmeier-Tiessen et al. (2016)
Elsevier RDM solutions workflow	Collection and processing, Selection, Publication	Haak, De Waard, Zudilova-Seinstra, Shell, Jones, Cousijn and Koers; response in Dallmeier-Tiessen et al. (2016)
EOL Quality Control of Dropsonde Data	Collection and processing	Callaghan and CEDA (2013)
Galaxy-ISA-Gigascience-Nanopublication	Collection and processing, Selection, Publication	González-Beltrán et al. (2015)
Imperial College: RDM by researchers to meet institutional policy	Planning, Collection and processing, Selection, De-identification, Publication	Addis (2015)
IPCC Data Distribution Center (IPCC-DDC)	Planning, Collection and processing, Selection	Stockhause et al. (2012)
NCAR EOL Data Management Group Workflow	Planning, Collection and processing, Selection, Publication	Callaghan and CEDA (2013)
NCAR/EOL Atmospheric Sounding Processing Procedures	Collection and processing	Callaghan and CEDA (2013)

1 Amy Nurnberger 8 Dec. 2015 'Requesting your input: Research workflows informed by the intent to publish data' post to RDA listserv, available at: https://rd-alliance.org/group/rdawds-publishing-data-workflows-wg/post/requesting-your-input-research-workflows-informed

Workflow name	Workflow type addressed	Source document /contributors
Ontologies for research data tools workflow	Collection and processing	Aguiar Castro, Ribeira, Roca da Silva, and Carvalho Amorim; response in Dallmeier-Tiessen et al. (2016)
Science 2.0 Repositories	Collection and processing, Selection, Publication	Assante et al. (2015)
Use of DOIs for computational chemistry data	Planning, Collection and processing, Selection, Publication	Addis (2015)

The examples vary in maturity, from conceptual models (e.g. Assante et al., 2015) through proof-of-concept exemplars (e.g. González-Beltrán et al., 2015) and prototypes (e.g. Ward et al.) to fully implemented processes (e.g. Callaghan and CEDA, 2013). Three of the examples are further described in a dataset for this article (Dallmeier-Tiessen et al., 2016). Relevant workflow tools and models were also highlighted. These included the Berkeley Initiative for Transparency in the Social Sciences, the Open Science Framework, and Taverna. These are briefly described in the Appendix.

Applicability of the Data Publishing Recommendations to Upstream Workflows

The collected examples provide different contexts for data publication preparation by researchers and/or service providers. We acknowledge the sample is very small and cannot be considered representative. Nevertheless, they offer a basis for characterising some challenges for our recommendations that we describe below.

Starting small, building modular, open source and shareable components

Research workflow examples, such as those detailed by IPCC-DDC (WDCC) and CERN, provide additional components that are small, modular, open source and shareable, and which clearly complement the more static 'downstream' data publication workflows presented by Austin et al. (2016). They illustrate more complex research workflows and the 'work in progress' nature of some of the content elements. They operate by establishing a counterpart that allows early referencing and versioning, and that often facilitates collaborative communication elements. It should be noted that access is frequently restricted where content is 'work in progress', and subsequently published openly.

Nanopublication is one such approach, aiming to enhance reproducibility by employing data modelling frameworks and executable workflows. González-Beltrán et al. (2015) for example reproduce results from a selected life science paper using a range of nanopublication methodologies. Their paper provides useful insights into the merits of these, and argues that better systems are needed to support reproducibility. The authors assert that wider testing of the principles of nanopublication could strengthen the scholarly communications lifecycle: from research, through to peer review and publication.

Some of the workflows illustrate heavily computational areas of research. The integration of the Galaxy platform with the data journal *Gigascience* and with open

RDM platforms such as myExperiment[2] exemplifies the steps required to ensure future reproducibility of computational research. These include implementation of standardized, automated components into an integrated and executable workflow, along with instructions on how to use data and related materials (see, for example, Gil et al., 2007).

The use of small, modular, shareable components may help ensure platforms offer sufficient flexibility to support variety, both in terms of the workflows supported, and the content these produce for publication. This diverse collection of workflow solutions clearly exemplifies how necessary it is to address the diverse content and needs of a research community (metadata, restrictions, publication products). The prototype nature of the collected examples underlines the necessity to work step by step together with community members in order to connect data publication with the research workflow.

Following standards that facilitate interoperability and permit extensions

All of the workflow support examples identified provide some form of standardized interface between workflow components through the use of metadata standards for data discovery and citation (e.g. DataCite, Dublin Core) and standards for packaging, exchanging and exposing content (e.g METS, SWORD, Linked Open Data). Solutions that enable straightforward data and metadata generation early in the research cycle, and in accordance with community defined and accepted standards, help expose these intellectual products and enhance their reuse.

In the examples reviewed, generic data citation standards (DataCite) were commonly used. Specific disciplines or communities mentioned included life and biomedical sciences, climate sciences and high energy physics. With the exception of the ISA group of standards for life and biomedical sciences, referred to in González-Beltrán et al. (2015), we had insufficient evidence of domain-specific content standards being used upstream to draw any conclusions.

Metadata captured upstream in the research process also needs to be exposed in standard formats if the research data is to be published, reused by others and the benefits fully realised. Regarding data packaging, exchange and exposure, our preliminary analysis suggests more widespread use of proprietary APIs than of interoperability standards for these purposes. Among our examples initial steps in this direction were being taken using METS, SWORD, JSON and JSON-LD.

Several examples use electronic laboratory notebooks (ELNs), in keeping with long-standing aims of using such tools for ‘curation at source’ i.e. creating and eliciting metadata as data is produced (Frey, 2008). In one, ‘RSpace ELN to DataShare Repository’, open standards are deployed to enable researchers to deposit directly from the RSpace environment to an institutional data repository, Edinburgh DataShare.

The ELN content is exported as XML documents, and packaged as a zip archive with METS descriptive header, including the DataCite minimum metadata required, and deposited using the SWORD protocol. This workflow results from a partnership between University of Edinburgh and Research Space, a provider of electronic lab notebook (ELN) software.

In the second example, ‘Ontologies for research data tools’, the workflow is supported by Dendro (da Silva et al., 2014) an ontology-based collaborative platform for research data. Dendro offers researchers a file management environment with a tool for creating metadata descriptors as Linked Open Data (LOD), optionally picking recommended terms from published vocabularies, including elements from well-recognized standards like Dublin Core.

2 myExperiment: http://wiki.myexperiment.org/index.php/Galaxy

Curators can work with Dendro to design domain-specific metadata models, and enrich the terms available to researchers they work with. The Dendro workflow optionally includes Labtablet, a mobile application designed to allow researchers to capture metadata on fieldwork. Locally relevant terms are packaged with the data for deposit in a public repository, while the terms themselves are published on the web as candidate ontologies for the researchers' domain, allowing for their evolution through broader community reuse.

Facilitating data citation, e.g. through use of identifiers and linkages

The pervasive use of identifiers can help instantiate data citation as an active practice, and the Joint Declaration of Data Citation principles appear to be becoming accepted by general consensus.[3] The examples show that some stakeholders are already getting involved in services to assess publishing and reuse patterns. Exposing information about content and their identifiers in a machine readable way facilitates such exercises.

When dealing with complex workflows and dynamic content, it is important for the purposes of reproducibility to be able to identify data, software, and documentation correctly and uniquely. Hence, it is not surprising that most of the described approaches clearly commit to the use of PIDs and include versioning capabilities. Independent of any software environment, PIDs can be used to connect content such as data, software and publications. Ideally, solutions would be able to track changes to a digital object through internal, restricted and public modules.

Persistent identifiers, such as DOIs, are required for any digital object that may be cited. However, we note that identifiers can serve data publication purposes by aiding reproducibility in other ways. Objects may have internal identifiers if they are only temporarily required for tracking reasons, or would not be cited because they are at too granular a level.

Also, it should be noted that identifiers may be applied to physical objects and the persons involved in the processes. The advent of ORCID as a unique identifier for contributors allows easier attribution to individuals. It could be expected that researchers use several independent systems throughout their research process, and hence such IDs could be used to connect contents automatically across these systems, where permitted.

Documenting roles, workflows and services

User documentation for research data workflow support services should ideally promote transparency and generate service uptake, which in turn can assist in documenting benefits to each user community. Compared with data repository and other downstream publishing platforms, it is evident in our examples that the responsibilities for documentation may be spread across multiple providers of upstream services.

The examples commonly documented roles and responsibilities of providers and users. This is accepted practice in research data management, and is reflected for example in the information researchers are expected to provide in data management plans (see, for example, DMPonline[4]). Nevertheless, the examples identified were mostly works in progress. As such it would be wrong to assume that comprehensive documentation is available.

3 Joint Declaration of Data Citation Principles: https://www.force11.org/group/joint-declaration-data-citation-principles-final

4 DMPonline: https://dmponline.dcc.ac.uk/

Providing clear checklist-style documentation can pay off by acting as a first step towards partial automation of the deposition workflow. The 'Rspace ELN to DataShare' workflow, which is based around a deposition checklist, illustrates this. As a result, researchers can capture data in a structured way during the research process, and then retain and deposit this structure without duplication of the initial effort, and with potential benefit for reproducibility.

This example and others (e.g. Science 2.0 Repository model) highlight the need for coordination of documentation where service providers take different roles as data flows downstream from production. A 'trustworthy' repository, certified according to the emerging standards such as the Data Seal of Approval[5], may outsource or delegate certain data management or curation functions. Where researchers use tools that they and their institution trust, this may facilitate delegation of curation tasks to research groups, and reduce repository ingest costs. To the extent that trust is transitive, delegation of service provision functions could work to the mutual advantage of providers and the user groups to which they delegate functions. This implies coordination between the various actors involved in providing these functions, possibly to varying degrees at different stages of data management. Service providers could facilitate that coordination by publicly documenting their respective roles in the research data management workflows they aim to support.

Curators have a role in connecting research workflows to publishing platforms

The organisational aspects of research workflows come into focus when workflows connect the practices of a number of different tool or service providers. The examples submitted to the Working Group often imply some measure of intermediation by curators to enable workflows to be joined up effectively. This could range from simply making researchers aware of tools, through enabling elements of automation, through to supporting uptake.

In some cases, the intermediation between different service components is provided by an institutional research data service (e.g. University of Edinburgh, Imperial College). Beyond encouraging or requiring individual providers to offer online documentation, as already described, research data services can provide overall service catalogues, and advise on the use cases each service component is intended to meet.

There was also innovation in the methods that curators use to engage with researchers and understand the workflows they are integrating. An example of this was 'Ontologies for research data tools'. Here the authors describe their approach to defining context-specific domain ontologies, in which they invite researchers to an interview about their data activities, requirements and their expectations regarding data sharing. This interview is based on the Data Curation Profile Toolkit (Witt, 2009). The authors describe complementing this through content analysis of researchers' publications, and then discussing with them the fragments of information that others will required to interpret the dataset (da Silva et al., 2014).

Implications for Upstream Data Publishing

The reference model and recommendations of the RDA/WDS Publishing Data Workflows working group (Austin et al., 2016) offer a 'joined up' approach involving repositories, publishers and other services (such as persistent identifiers). It would be

5 Data Seal of Approval: http://www.datasealofapproval.org

premature to update that model to reflect the upstream workflows found in the small collection of examples identified here, although that will be desirable when more solutions are available and in use.

Community engagement to support uptake of workflow tools and services that connect with data publishing is critical. This is a task for the repository community, e-infrastructure providers, funders, thought leaders within disciplines, institutions, research managers, and other key stakeholders. More work is needed to understand how (and to what extent) research groups are connecting upstream tools to downstream repositories, and any added value they expect to get from greater provenance or context for their published outputs. Further examples and data on actual usage are of uttermost importance to understand whether and how workflow support tools work in the context of research data publishing.

Meanwhile, our review gives a partial snapshot of a landscape that appears to be changing in long-anticipated directions. In the interest of stimulating further debate on good practice, we consider below how our previous recommendations may be adapted to better reflect shared characteristics of the upstream examples collected.

One of the most noticeable common features our examples manifest is a desire by service providers to offer integrated 'whole lifecycle' solutions. In the months since we solicited examples for our review there have already been substantial developments of that nature in the upstream data management landscape. These developments include changes in the market for commercial services, and for commons-based infrastructure.

A key development in the commercial arena involves one of the reviewed examples; Elsevier's RDM solutions workflow. The acquisition of the Hivebench lab notebook by Elsevier is intended to offer researchers the ability to link notes in that environment with research outputs managed in other Elsevier platforms, including Pure and Mendeley. This development may potentially make data management easier for researchers using the integrated toolset.[6] It has not been universally welcomed, adding to concerns about the sustainability and governance of research workflows that Bilder, Lin and Neylon (2015) have articulated in a set of *Principles for Open Scholarly Infrastructure*.

There have been further key developments in the public arena, for example in the solutions available through EU research infrastructures. These include the EUDAT 'B2 Service Suite',[7] which offers an integrated set of data management services. They also include the OpenAIRE infrastructure for open access. Through the Zenodo service this enables researchers to link data, code and articles, and further support is intended for workflow integration based on notification services.[8]

The desire to provide whole-lifecycle support is echoed in a recent report on RDM workflows in UK Higher Education Institutions (Addis, 2015). This offers scenarios for linking preservation and publishing platforms with archival storage, based on approaches across disciplines and institutions of varying size and research-intensity. Whilst acknowledging the impossibility of a 'one size fits all' solution, the report conclusions include the following:

- Researchers may be more likely to engage with data publishing if presented with clear and seamless support that integrates data publishing with their entire workflow.

6 See: https://www.elsevier.com/connect/putting-data-management-in-the-hands-of-researchers-with-hivebench-acquisition

7 EUDAT B2 Service Suite: https://www.eudat.eu/b2-service-suite

8 OpenAIRE: Open Science as-a-Service: https://www.digitalinfrastructures.eu/content/openaire-open-science-service

- Automation should aim, where possible, to drive the speed, accuracy and cost-efficiency of RDM workflows, and support institutions to share service provision through single points of contact or interfaces. Automated support for curation is essential to deal with the exponential growth in data.

A further point worth reiterating is that a more diverse set of stakeholders are involved when integrating upstream and downstream workflows. A broader set of decisions will be made by researchers and institutions about tools, platforms and providers involved. An array of influences from disciplinary cultures to institutional policies and personalities will shape decisions made at any stage in the workflow, with consequences for downstream choices. For example, researchers may be influenced in their choice of metadata standards by collaborators' working practices, which may in turn constrain their choice of downstream repositories. Or they may be guided by their institution towards licensing choices that affect where they may publish data subsequently.

Our collected workflow examples offer brief descriptions of their contexts. More detailed case studies should indicate whether the proposed solutions allow users enough cultural and political 'room to manoeuvre', to make free and informed choices that account for preference and circumstance. We would expect an integrated data publishing infrastructure to draw on existing good practice and aim for 'loose coupling', a widely applied computing and business process management concept. This refers to the desirability of limiting inter-dependence, i.e. the need for components to encapsulate knowledge of each other's internal operation (see, for example, Kamoun, 2007).

Loose coupling is desirable in the business processes that software workflows support and are embedded in, as well as to the software components of those workflows, especially where the business processes themselves bisect organisational boundaries (Hagel and Seely-Brown, 2005). Loose coupling is generally favoured as a strategy for ensuring software flexibility and interoperability, and is enhanced by application of open standards. In the business process management domain, loose coupling is associated with business objectives of flexibility and interoperability, and with strategies that align with our recommendations for data publishing:

- Modular design of workflow components,
- Standard vocabularies and protocols to describe components,
- Standardized ways of specifying capabilities and performance requirements,
- Significant investment in building trust-based relationships among participants (Hagel and Seely-Brown, 2005).

In the research context, we see a role for curators in promoting loose-coupling strategies for research data services, to help mitigate the risk of over-dependence between upstream and downstream components in services they provide. Curators and others who support researchers to manage their own data publishing workflows could also support a loose coupling approach to those workflows, by seeking platform providers that use open standards in their APIs, and offering researchers support to make informed choices of platform and provider at the main decision points in their research workflow. This might include, for example, the choice of a storage provider or metadata editing platform early in the data management lifecycle, or a repository for data publication at later stages.

The work reported by González-Beltrán et al. (2015) exemplifies aspects of a 'loosely coupled' workflow, as it employs standardised methods to explicitly declare the

elements of experimental design, variables, and findings. Generally, however we lack information on whether integrated data publishing solutions, such as those offered in the workflow examples we gathered, offer researchers enough flexibility in downstream service components they need to manage and publish the research objects they produce.

Conclusions and Recommendations

This first step towards a landscape review shows that data publication practices and products are emerging to better serve upstream research workflows. They extend the 'traditional' data publishing model (Austin et al., 2016) to preserve internal 'work in progress', i.e. make dynamic content ready for preservation and publication earlier in the research process. The examples also indicate the desirability of:

- Curation support for researchers to and choose service components that will maintain interoperability across their data publishing workflows,
- Solutions that enable workflows to link computational and content preservation components,
- Solutions that are easily extendable: facilitated by APIs and new data models,
- More work to embed such tools and workflows into the 'business as usual' experience of the critical mass of researchers.

Working on the assumption that upstream data management platforms should normally be loosely coupled with those for downstream data preservation and publication, it is worth considering how the recommendations may apply differently to the various platform providers involved in a connected workflow.

Differentiating Between Upstream and Downstream Roles

Recommendations that account for different roles in an integrated workflow can begin with that of the integrator. In practice this might be any organisation, but for our purposes we assume this will be operated by a research institution, funder or research infrastructure provider and take the form of a 'research data service'. This may have technical and organisational aspects, e.g. human curation, and can be defined as follows:

> Research data service: A means of delivering value to the producers and users of digital objects by facilitating outcomes they want to achieve without the ownership of specific costs or risks.[9]

Service components to support research data workflows would include the following:

- **Active data management service:** A service offering to create or transform digital objects for the purposes of research.
- **Research data preservation service:** A service offering to ensure digital objects meet a defined level of FAIRness – findability, accessibility,

9 Derived from the definition of a service employed by the ITIL (IT Infrastructure Library) standard for service management, see: http://itsmtransition.com/2014/01/what-is-itil-service/

interoperability, and reusability – for a designated community and period of time.

- **Research data publication service:** A service offering to enhance digital objects FAIRness by reviewing their quality on specified criteria, or connecting them to additional metadata.
- **RDM guidance service:** A service offering research data service users practical guidance, including on choosing or using the above services.

Recommendations for Research Data Services on Integrating Workflows

Using the above working definitions we offer the following recommendations for research data services that seek to integrate components in open, loosely coupled workflows. We welcome further comment on these.

1. Active data management services should use open standards to express and expose the objects and metadata they offer to downstream services, including their access and reuse terms.
2. Preservation and publication services should publish policies stating what digital object types they accept, for what communities, and on what terms and conditions.
3. Active data management, preservation and publication services should make openly available sufficient metadata to enable reuse of their outputs, including all terms and conditions for third-party access and reuse.
4. Active data management, preservation and publication services should make sufficient detail of their workflows available to support the provenance of digital objects the workflows produce, and the reproducibility of research they support.
5. Guidance services should support users of other services to make an informed choice of downstream service capabilities, informed by consideration of relevant compliance, risk, and data value factors and based on independent guidance.
6. Guidelines 1-5 should be implemented using content that is findable, accessible, interoperable and reusable (FAIR).

Final Remarks

This preliminary analysis offers indications of how the service ecosystem is evolving to join up research data management workflows, spanning the research lifecycle from data production to publishing. RDA/WDS Working group sessions highlighted a considerable interest in such solutions. Further work is needed to collect examples and provide a more comprehensive picture of integrated RDM workflows. That work should clarify to what extent data publication motivates the collection of metadata and identifiers early in the research lifecycle, by what actors and service components, and at which stages.

We have re-articulated the previous recommendations of the RDA/WDS Data Publishing Workflows Working Group to account for the varied upstream service components and platforms that support the flow of contextual and provenance information downstream. These workflows should be open and loosely coupled to support interoperability, including with preservation and publication environments. We recognise the limitations of the evidence and analysis we have gathered to date, but aim to stimulate further work on researchers' views of data publishing and the extent to which available services and infrastructure facilitate the publication of FAIR data. We also aim to stimulate further dialogue and definition, e.g. through the RDA/WDS Interest Group, of the roles and responsibilities of research data services and platform providers for the 'FAIRness' of research data publication workflows themselves. That research and community dialogue will inform further development of the Reference Model for Data Publishing.

Acknowledgements

The authors gratefully acknowledge workflow contributions from Sarah Callaghan, Pauline Ward, Rory MacNeil, Paolo Manghi, Jonathan Tedds, Mary Vardigan, Claire Austin, Martina Stockhause, João Aguiar Castro, Cristina Ribeira, João Rocha da Silva, Ricardo Carvalho Amorim, Wouter Haak, Anita de Waard, Elena Zudilova-Seinstra, Joe Shell, Mike Jones, Helena Cousijn, and Matthew Addis. We also thank Neil Jefferies, Ruth Duerr, and three anonymous reviewers for their comments on earlier versions of this paper.

References

Addis, M. (2015). RDM workflows and integrations for HEIs using hosted services. doi:10.6084/m9.figshare.1476832.v3

Assante, M., Candela, L., Castelli, D., Manghi, P., & Pagano, P. (2015). Science 2.0 repositories: Time for a change in scholarly communication. *D-Lib Magazine*, *21*(1/2). doi:10.1045/january2015-assante

Austin, C.C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V.K., Murphy, F., Nurnberger, A., Raymond, L., Stockhause, M., Tedds, J., Vardigan, M., & Whyte, A. (2016). Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries.* doi:10.1007/s00799-016-0178-2

Beagrie, N., Lavoie, B., & Woollard, M. (2008, 2010). Keeping research data safe: Cost-benefit studies, tools, and methodologies focussing on long-lived data. Retrieved from http://www.beagrie.com/krds.php

Bilder, G., Lin, J., & Neylon, C. (2015). Principles for open scholarly infrastructures v1. doi:10.6084/m9.figshare.1314859.v1

Callaghan, S., & CEDA Team. (2013). Workflows for data publication, from repository to data journal. Retrieved from http://proj.badc.rl.ac.uk/preparde/attachment/wiki/DeliverablesList/D2_1_D2_2_PREPARDE_Workflows_combined_draft1.pdf

Dallmeier-Tiessen, S., Lavasa, A., Herterich, P., Rueda, L., Kotarski, R., & Newbold, E. (2014). A comparative analysis of disciplinary data management workflows. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 281–284). IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2740817

Dallmeier-Tiessen, S., Khodiyar, V., Murphy, F., Nurnberger, A., Raymond, L., & Whyte, A. (2016). RDA publishing workflows – Research workflows (questionnaire responses). doi:10.5281/zenodo.167046

Darch, P.T., Sands, A.E. (2015). Beyond big or little science: Understanding data lifecycles in astronomy and the deep subseafloor biosphere. In *iConference 2015 Proceedings*. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/73655/185_ready.pdf

Data Seal of Approval. (2013). Guidelines version 2. Retrieved from https://assessment.datasealofapproval.org/guidelines_52/html/

Farquhar, A., & Hockx-Yu, H. (2008). Planets: Integrated services for digital preservation. *International Journal of Digital Curation*, *2*(2), 88–99.

Frey, J., De Roure, D., & Carr, L. (2002). Publication at source: Scientific communication from a publication Web to a data Grid. Paper presented at the Euroweb 2002 Conference, Oxford, UK. Retrieved from http://ewic.bcs.org/content/ConWebDoc/4084

Frey, J.G. (2008). Curation of laboratory experimental data as part of the overall data lifecycle. *International Journal of Digital Curation*, 3(1), 44–62. doi:10.2218/ijdc.v3i1.41

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., ... & Myers, J. (2007). Examining the challenges of scientific workflows. *Ieee computer*, *40*(12), 26-34. Retrieved from http://eprints.soton.ac.uk/271187/

Goble, C.A., & De Roure, D.C. (2007). myExperiment: Social networking for workflow-using e-scientists. In *Proceedings of the 2nd workshop on Workflows in support of large-scale science* (pp. 1–2). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1273361

González-Beltrán, A., Li, P., Zhao, J., Avila-Garcia, M.S., Roos, M., Thompson, M., et al. (2015). From peer-reviewed to peer-reproduced in scholarly publishing: The complementary roles of data models and workflows in bioinformatics. *PLoS ONE 10*(7): e0127612. doi:10.1371/journal.pone.0127612

Hagel, J., & Brown, J.S. (2005). *The only sustainable edge: Why business strategy depends on productive friction and dynamic specialization*. Harvard Business Press.

Huc, C., Boucon, D., Sawyer, D.M., & Garrett, J.G. (2003). The Producer-Archive Interface Methodology Abstract Standard (PAIMAS). Retrieved from http://arc.aiaa.org/doi/pdf/10.2514/6.2004-649-446

Jahnke, L., Asher, A., & Keralis, S.D.C. (2012). The problem of data, with an introduction by Charles Henry. Council on Library and Information Resources. Publication 154. ISBN 978-1-932326-42-0

Kamoun, F. (2007). The convergence of business process management and service oriented architecture. *Ubiquity.* Retrieved from http://dl.acm.org/citation.cfm?id=1276167&coll=portal&dl=ACM

Kowalczyk, S.T. (2014). Where does all the data go: Quantifying the final disposition of research data. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–10. doi:10.1002/meet.2014.14505101044

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., … Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, *18*(10), 1039–1065.

Parsons, M.A., & Fox, P.A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*(0), WDS32–WDS46.

Schmidt, R., King, R., Jackson, A., Wilson, C., Steeg, F., & Melms, P. (2010). A framework for distributed preservation workflows. *International Journal of Digital Curation*, *5*(1), 205–217. doi:10.2218/ijdc.v5i1.154

Silva, J.R.d, Ribeiro, C., Lopes, J.C. (2014) The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. In Proceedings of the 11th International Conference on Digital Preservation iPRES2014, Melbourne, Australia, 6-10 October 2014, pp. 189-193. Retrieved from https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf

Stockhause, M., Höck, H., Toussaint, F., Lautenschlager, M. (2012). Quality assessment concept of the World Data Center for Climate and its application to the CMIP5 data. *Geoscientific Model Development* 5(4):1023-1032. doi:10.5194/gmd-5-1023-2012

Tenopir, C., Suzie, A., Douglass, K., Aydinoglu, A., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE, 6*(6). Retrieved from http://dx.plos.org/10.1371/journal.pone.0021101

Van den Eynden, V., Knight, G., Vlad, A., Radler, B., Tenopir, C., Leon, D., Manista, F., Whitworth, J., & Corti, L. (2016). Survey of Wellcome researchers and their attitudes to open research. doi:10.6084/m9.figshare.4055448.v1

Waddington, S., Green, R.A., & Awre, C.L. (2012). CLIF: Moving repositories upstream in the content lifecycle. *Journal of Digital Information*, *13*(1). Retrieved from https://hydra.hull.ac.uk/resources/hull:5580

Wallis, J.C., Borgman, C.L., Mayernik, M.S., & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation*, *3*(1), 114–126. doi:10.2218/ijdc.v3i1.46

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., … et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/

Williams, R. & Pryor, G. (2009). Patterns of information use and exchange: Case studies of researchers in the life sciences. London: Research Information Network and British Library. Retrieved from http://www.dcc.ac.uk/projects/life-science-case-studies

Witt, M. et al. (2009). Constructing data curation profiles. *International Journal of Digital Curation, 4*(3), pp.93–103. doi:10.2218/ijdc.v3i1.46

Appendix: Workflow Tools

The following examples of data publishing tools were identified in responses to the RDA/WDS Working Group on Data Publishing Workflows call for upstream workflow examples. The list is not meant to be exhaustive or definitive.

- **Open Science Framework:** The Center for Open Science[10] has developed the Open Science Framework (OSF)[11], which is part network of research materials, part version control system, and part collaboration software. The purpose of the software is to support the scientist's workflow and help increase the alignment between scientific values and scientific practices.
- **Berkeley Initiative for Transparency in the Social Sciences:** The Berkeley Initiative for Transparency in the Social Sciences[12] is an international network of researchers and institutions committed to improving the standards of openness and integrity in economics, political science, psychology, and related disciplines. Central to BITSS efforts is the identification of useful tools and strategies for increasing transparency and reproducibility in research, including the use of study registries, pre-analysis plans, version control, data sharing platforms, disclosure standards, and replications. A best practices manual[13] offers suggestions for managing workflow in a transparent and systematic way.
- **Taverna:** Taverna[14] is a workflow tool that supports implementations of workflows intended to result in the publication of research data in all domains, predominantly in the biological and life science domain[15]. The open source tool is able to connect to various data resources and enables computational (re)implementation of (research) workflows.

10 Center for Open Science: http://centerforopenscience.org/
11 Open Science Framework: https://osf.io/?_ga=1.164844473.72750444.1430154145
12 Berkeley Initiative for Transparency in the Social Sciences: http://www.bitss.org/
13 BITSS Best Practice Manual: https://github.com/garretchristensen/BestPracticesManual/blob/master/Manual.pdf
14 Taverna: http://www.taverna.org.uk
15 See: http://nar.oxfordjournals.org/content/34/suppl_2/W729.shor