Connecting data publication to the research workflow: a preliminary analysis

Authors: Sünje Dallmeier-Tiessen; Varsha Khodiyar; Fiona Murphy; Amy Nurnberger; Lisa Raymond; Angus Whyte

Acknowledgements: 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

Introduction

The data curation community has long encouraged researchers to document their collected research data early in the research workflow in a manner that will support publication and preservation, as well as future understanding, interoperability, and reuse. Introducing smoothly continuous, considered, and consistent data documentation practices to the research workflow, as close to the data collection point as possible, should reduce the data sharing burden that researchers associate with the data deposition process. Documentation of data closer in time to data collection would also help mitigate the high cost that repositories associate with the ingest process, as seen in the work by Beagrie *et al.* (2008 and 2010). In this extension to the Data Publishing Workflows working group's initial report (Austin et al., 2015), we provide here a preliminary review of a selection of research workflows, with the intent of identifying connections between the goal of research data publication and the incorporation of such 'upstream' barrier-reducing measures into the research workflow. These measures, intended to facilitate data publication, might include data preparation practices that offer the possibility of interaction with a repository before that data is ready for publication, use of workflow platforms that support such practices, and participation in "a novel publishing paradigm where "publishing" is intended 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). A sampling of the research workflows that may be influenced in such a way were described by Addis (2015) and include data management planning; collection, creation, analysis, and use of data; data selection and access decisions; resolving ethical issues through de-identification, and publication.

Frequently, moving data curation activities closer to the research activity implies integrating elements such as code, software, models, documentation, products from the research process. It is important to understand how the intention to publish data might be made integral to the research workflow. For example, considering how best to cross-link the scholarly literature with different software releases and different versions of data versions, as well as considering integration with labbooks, ipython notebooks and so on, could be important contributions to making data publishing a usual part of the researcher's workflow.

Given the range of disciplinary practices in the research workflow, it is to be assumed that such an analysis will show stronger differences than in this group's previous analysis (Austin et al, 2015). The latter focused on high-level outcomes that are broadly applicable to multiple disciplines and are less subject to the dynamic changes that are frequently found in research processes. The work of Williams and Pryor (2009) underlines the complexity and diversity of research workflows and respective information flows across disciplines.

As a first step in the direction of the upstream analysis, we consider how the examples collected instantiate the report's recommendations on data publishing..

Examples of upstream workflows

Twelve workflows were collected, representing a range of disciplines and broader projects. Here we present a first step towards a state of the art review of current data publishing solutions or projects that tie in with an active research workflow. They are put into the context of the recommendations made by Austin et al. (2015).

Workflow table

Workflow name	**Workflow types (planning, collection and processing, selection and access, deidentification, publication)**	**Source document/ Contributors**
CERN Analysis Preservation	Collection and processing, Selection, Publication	Dallmeier-Tiessen
Electronic lab notebook to data repository (RSpace to DataShare)	Collection and processing, Selection, Publication	Ward, MacNeil, and Whyte

Elsevier RDM solutions workflow	Collection and processing, Selection, Publication	Haak, De Waard, Zudilova-Seinstra, Shell, Jones, Cousijn, and Koers
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, Deidentification, Publication	Addis (2015)
IPCC Data Distribution Center (IPCC-DDC)	Planning, Collection and processing, Selection	Stockhause
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)
Ontologies for research data tools workflow	Collection and processing	Aguiar Castro, Ribeira, Roca da Silva, and Carvalho Amorim
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)

In addition to these research workflows, several workflow tools and guidelines that support and enable data publication producing research workflows were recommended to us during our collection process. These included the Berkeley Initiative for Transparency in the Social Sciences, the Open Science Framework, and Taverna, which are detailed in more depth in the Appendix.

Implications for data publishing recommendations

Start 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 previous static data publication workflow analysis (Austin et al. 2015). Other workflows accommodate the more complex research workflows and the "work in progress" nature of some of the content elements, by establishing a counterpart that allows early referencing, versioning and often facilitates collaborative communication elements. It should be noted that access is often restricted to content that is "work in progress". The diverse content that might be accumulated in such an approach, could be published openly, following the example of nanopublications (for data and software).

Nanopublications offer, in part, an established method for enhancing reproducibility by way of data modelling frameworks and executable workflows. González-Beltrán et al (2015) conducted an experiment to reproduce the results from a selected life science paper using a range of nanopublication methodologies. Their resulting paper provides useful insights into both the relative merits of the systems themselves and the reasons why better systems are needed to support reproducibility. The authors also point out that, if the principles of nanopublication can be evaluated and accepted by a critical mass of the research community, they could strengthen the scholarly communications model throughout its lifecycle: at points within the research, throughout the review process, and in the publication model.

Some of the represented workflows pay more attention to the computational components, which is reasonable for those areas of research that are heavily computational. The implementation of standardized, automated components (with instructions on how to use data and related materials), is considered an important step for future reproducibility of research (see for example Gil *et al*, 2007). One example of an executable workflow, is the integration of the Galaxy platform with the data journal *Gigascience* and with open RDM platforms such as myExperiment [1].

A closer connection between digital research infrastructure, traditional repositories and research communities is proposed in the concept of Science 2.0 Repositories. Assante et al. (2015) suggest that research infrastructure services should intercept and publish research products, whilst providing researchers with social networking tools for discovery, notification, sharing, discussion, and assessment of research outputs. Even though it is possible that the concept could be implemented in a modular way, the overall concept provides a single locus that would serve one or multiple research communities with a complex information system.

In summary, with the diverse content that results from an upstream research workflow, there is a need to address the individual needs (metadata, restrictions, publication products) step by step together with the research community.

[1] http://wiki.myexperiment.org/index.php/Galaxy

Follow standards that facilitate interoperability and permit extensions

This preliminary analysis underlines the need to understand and distinguish between the different available standards: e.g. disciplinary and generic metadata standards and standards around exchanging and exposing data. As dependencies between modules and objects might be more prevalent upstream in the research workflow (e.g. data that can only be analyzed with a specific software), it is vital to ensure components can exchange information smoothly and with minimal information loss. This reinforces the ongoing work on FAIR Data Principles being coordinated by Force11. At time of writing, the principles are still open for community consultation, but in essence they strongly encourage an approach to make research data "Findable, Accessible, Interoperable, and Re-usable" for both humans and machines.[2]

The advanced solutions identified by this analysis predominantly serve specific disciplines or communities, including life and biomedical sciences, climate sciences and High Energy Physics. All of them provide standardized interfaces between the components (closed and open counterparts) and data curation and standardization support. One example is IPCC-DCC which, for example, uses detailed project naming conventions for directory structures, data header information and file names. It appears that these solutions are being developed relatively closely with the research communities employing them.

Improvements to workflows for data discovery and exchange are still required; for example, current solutions utilize JSON and JSON-LD. Many solutions in this preliminary analysis use APIs to exchange and expose information about their content. While APIs at least enable the exchange of metadata across workflows, more open and sustainable approaches are based on open access protocols, and vocabularies openly published as Linked Open Data. Metadata captured upstream in the research process needs to be clearly exposed if it is to be reused by others and the benefits fully realised.

There is a growing number of electronic laboratory notebooks (ELNs) intended to help incorporate metadata curation into the data production workflow. The term 'curation at source' (Frey, 2008) has been used for such attempts to make metadata creation more effective, efficient, and less error-prone. Two of the examples submitted to the Working Group illustrate this: "RSpace ELN to DataShare Repository" and "Ontologies for research data tools". In the first, open standards are deployed to enable deposit from a proprietary ELN to an institutional data repository. In the second Linked Open Data is used to enrich research workflows with relevant descriptors, which may be published as domain ontologies further upstream.

In the first example, "RSpace ELN to DataShare Repository", the workflow enables researchers to deposit directly from the RSpace electronic lab notebook (ELN) to DataShare institutional data repository. The ELN content is exported as XML documents, and packaged as a zip archive with METS descriptive header, including the DataCite minimum metadata required for DataShare. The packaged content including citation metadata is deposited to DataShare using

[2] https://www.force11.org/group/fairgroup/fairprinciples - accessed 6 February 2016.

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 employs Dendro (da Silva et al, 2014) an ontology-based collaborative platform for research data management. 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. 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.

Facilitate data citation, e.g. through use of digital object PIDs, data/article/person/software linkages, researcher PIDs

With more complex workflows and dynamic content, it is even more important for humans and machines to be able to identify the data, software, and documentation correctly and uniquely for the purposes of reproducibility. Hence, it is not surprising that most solutions clearly commit to the use of PIDs and their versioning capabilities. Independent of any software environment, PIDs can be used to connect content such as data, software and publications.

It should be noted that the use of PIDs applies to not only the digital objects (data, software, any text document, etc), but also the physical objects and the persons involved in the processes. The advent of ORCID as a unique identifier for contributors allows an easy attribution of content to the individual person. 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, as permitted.

Many of the workflows that were studied incorporate PIDs: RSpace & Datashare, Elsevier, Imperial College London, Computational Chemistry, IPCC-DDC, CERN Analysis Preservation, Galaxy and Science 2.0 Repos. This ensures that content can be tracked throughout any module or workflow. Ideally, solutions would be able to track changes to a digital object through internal, restricted and public modules.

The pervasive use of such identifiers can assist in instantiating the active practice of data citation. It appears that most solutions today try to facilitate data citation and the Joint Declaration of Data Citation principles have been finding general consensus.[3] The analysis shows that some stakeholders are already getting involved in services to assess publishing and

[3] https://www.force11.org/group/joint-declaration-data-citation-principles-final - accessed 6 February 2016

reuse patterns. Exposing information about content and their identifiers in a machine readable way facilitates such exercises.

Document roles, workflows and services

Some of the examples identified in this analysis are still works in progress at the time of writing. Hence it is to be assumed that documentation is not yet comprehensive. However, one can note that documentation of roles and responsibilities in such solutions is significant. This would also help researchers to include such relevant information into data management plans (according to DMPonline[4]). Given the more complex nature of upstream systems, often involving a collaborative approach amongst several partners, there is a need for documented service level agreements and respective guidance for partners. User support is particularly relevant in order to generate the uptake of the service in the user community. If the added benefit is not highlighted explicitly it might be difficult to harness interest for a new tool, for instance.

In the “Rspace ELN to DataShare” example an institutional data repository (DataShare) provided a checklist approach to deposit. This subsequently facilitated a partnership with ELN software providers Research Space. The resulting partial automation of the deposition workflow shows how clear documentation can offer direct benefits to repository depositors and users. 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. Retaining the original structure of the research in a packaged form that may be associated with a publication benefits the reproducibility of the research.

Repositories partnering with third-parties can extend the trusted repository model to their partners by delegating certain data management or curation functions to them. Standards bodies, for example the Data Seal of Approval, recognize that a data service may be partially outsourced (Data Seal of Approval, 2013). Repositories can partner with providers of research tools and upstream services, as well as downstream integration with journal publishers, or with harvesting and aggregating services. Trust is transitive, and where researchers use tools that they and their institution trust this can facilitate a level of delegation of curation tasks to the research group, and may reduce repository ingest costs. There is also potential for economic and research benefits further ‘downstream’, to the extent that well-packaged research data facilitates easier integration with publication platforms and ease of reuse.

The curators’ role in connecting research workflows to publishing platforms

The examples submitted to the Working Group often identified 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 the uptake of services. This points to another area of 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

[4] https://dmponline.dcc.ac.uk/ - accessed 6 Februrary 2016

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). That process is complemented by performing content analysis in researcher's publications, and discussing with the researchers the fragments of information that should be provided along with the dataset to help others interpret it.

Summary

This first step towards a state of the art review shows that practices and products are emerging to serve upstream research workflows better in data publishing.

- An extension of the "traditional" data publishing model (Austin et al. 2015) to preserve internal "work in progress", i.e. dynamic content early in the research process
- An extension of collaborative features that enable easy collaboration with colleagues when conducting research
- Active interventions by curators which lead to better connected workflows and more richly engaged researchers
- Solutions that enable computational workflows (including preserved content);
- Solutions that are easily extendable: facilitated by APIs and new data models
- More work needs to be done to embed such tools and workflows into the "business as usual" experience of critical mass of researchers

Further investigations are needed to determine how data publishing can accommodate the results of the research workflow. This preliminary analysis underlines that a few solutions are under way and the discussions within the working group (sessions) also highlighted a considerable interest in such solutions.

Over time the emerging developments might require an updating of the reference model proposed by this working group (Austin et al. 2015) to refine the upstream components in more detail. But this will require more in depth work once more solutions are "on the market" and in use.

More work is also needed to understand whether and how communities (i.e. the individual researchers) really use such tools. So far there is almost no data available on the actual usage which is of uttermost importance to understand whether and how workflows work. Only with considerable uptake by researchers can such upstream workflows work in the mid and long term. In the light of these developments the recommendations presented in Austin et al. (2015) may themselves need a versioned update in the near future.

Community engagement to support uptake of the services is critical. This is a task for e-infrastructure providers, funders, thought leaders within disciplines, research managers and other key stakeholders.

Some of the findings are also supported by a recent report provided by Matthew Addis who considers both the function and effects of various RDM workflows. Addis et al (2015) contains a number of case studies taken from UK-based Higher Education Institutions. There is a range of discipline, size - and research-intensity level - of institution, size of dataset, and so forth. Whilst acknowledging the impossibility of devising a ‘one size fits all’ solution, the report does discern a number of useful conclusions:

- When presented with clear and seamless workflows, researchers are more likely to engage with the whole of the data publishing cycle.
- Automation, wherever possible, will drive speed, accuracy, and the ability of groups of institutions to provide a high level of services, as well as keeping costs down.
- A single point of contact or interface, even where different workflows/funders/subject areas are concerned, will also support engagement.
- Providing trusted metrics - for funders as well as the institutions - contributes greatly to the value of the exercise, particularly if these can be specifically linked with tangible career enhancement.

References

Addis, M. (2015). RDM workflows and integrations for HEIs using hosted services. http://dx.doi.org/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). http://doi.org/10.1045/january2015-assante

Austin, Claire C et al.. (2015). Key components of data publishing: Using current best practices to develop a reference model for data publishing. Zenodo. http://dx.doi.org/10.5281/zenodo.34542

Beagrie, Neil, Lavoie Brian and Woollard, Matthew (2008, 2010) with contributions by the Universities of Cambridge, Oxford, and Southampton, the Archaeology Data Service, OCLC Research, UK Data Archive, & University of London Computer Centre
Keeping Research Data Safe 2008, 2010 http://www.beagrie.com/krds.php

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

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 https://www.ideals.illinois.edu/bitstream/handle/2142/73655/185_ready.pdf?sequence=2

Data Seal of Approval (2013). Guidelines version 2, July 19, 2013 https://assessment.datasealofapproval.org/guidelines_52/html/

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

González-Beltrán A, Li P, Zhao J, Avila-Garcia MS, 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. http://dx.doi.org/10.1371/journal.pone.0127612

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. http://eprints.soton.ac.uk/271187/

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
https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf

Williams, R. & Pryor, G. (2009, November). Patterns of information use and exchange: Case studies of researchers in the life sciences. London: Research Information Network and British Library. Available at: 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.v4i3.117

Appendix: Workflow Tools

This series of examples aimed to to enhance understanding of how researchers locate, evaluate, organise, manage, transform and communicate information resources as part of the research process.

The Center for Open Science has developed the Open Science Framework (OSF), 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. http://centerforopenscience.org/

The Berkeley Initiative for Transparency in the Social Sciences 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 offers suggestions for managing workflow in a transparent and systematic way. http://www.bitss.org/

Taverna 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 (http://nar.oxfordjournals.org/content/34/suppl_2/W729.short). The open source tool is able to connect to various data resources and enables computational (re)implementation of (research) workflows. http://www.taverna.org.uk/