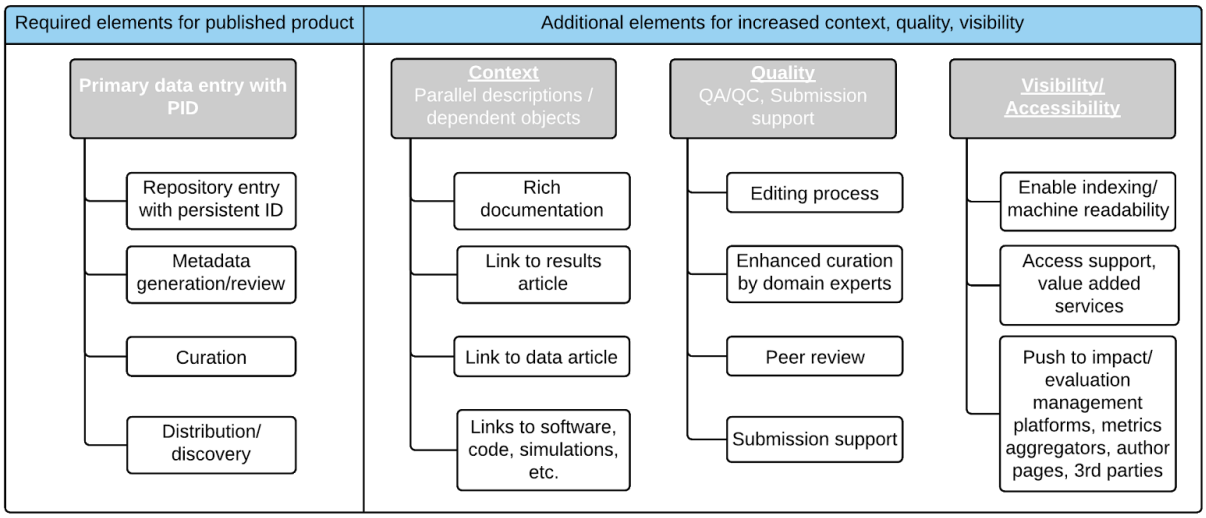
WDS-RDA Publishing Data Workflows Working Group Final Outputs

Co-Chairs: Amy Nurnberger, Varsha Khodiyar, Fiona Murphy, Sünje Dallmeier-Tiessen
Full report available here: http://dx.doi.org/10.5281/zenodo.34542

Executive Summary

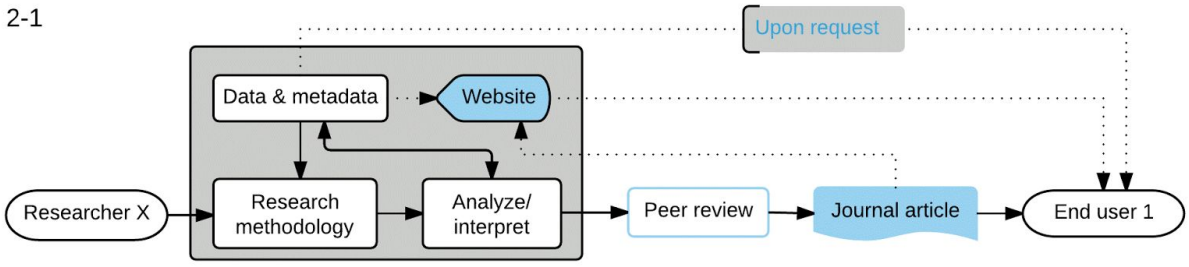
The key outputs of this group are the reference model featuring best practice recommendations for others to build upon. The results are based on a case study analysis which has been shared and discussed widely with the community. The feedback from various disciplines and stakeholder group has been implemented, potential extensions were identified so that the group's output also includes suggestions for next steps. A number of distinguished stakeholders from the community have already adopted the results from this WG.
The identified mandatory and recommended key components of data publishing are given below (Chart 1)

Chart 1: Key components for data publishing are listed in the following table:

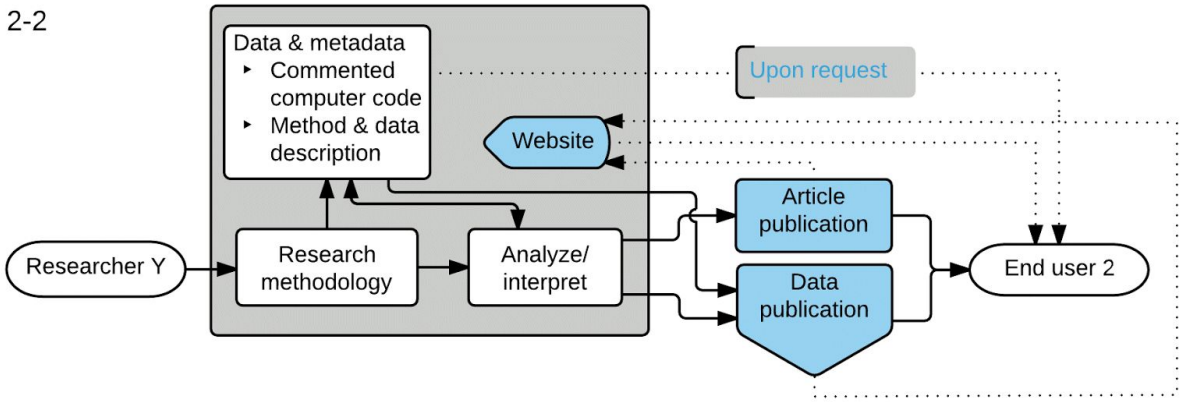


The proposed reference model for (static) data publishing is summarized in the following chart. It is based on case study analysis presented in Austin et al. (2015).
Figure 2-1: Traditional article publication workflow
Figure 2-2: Reproducible research workflow
Figure 2-3: Data publication workflow

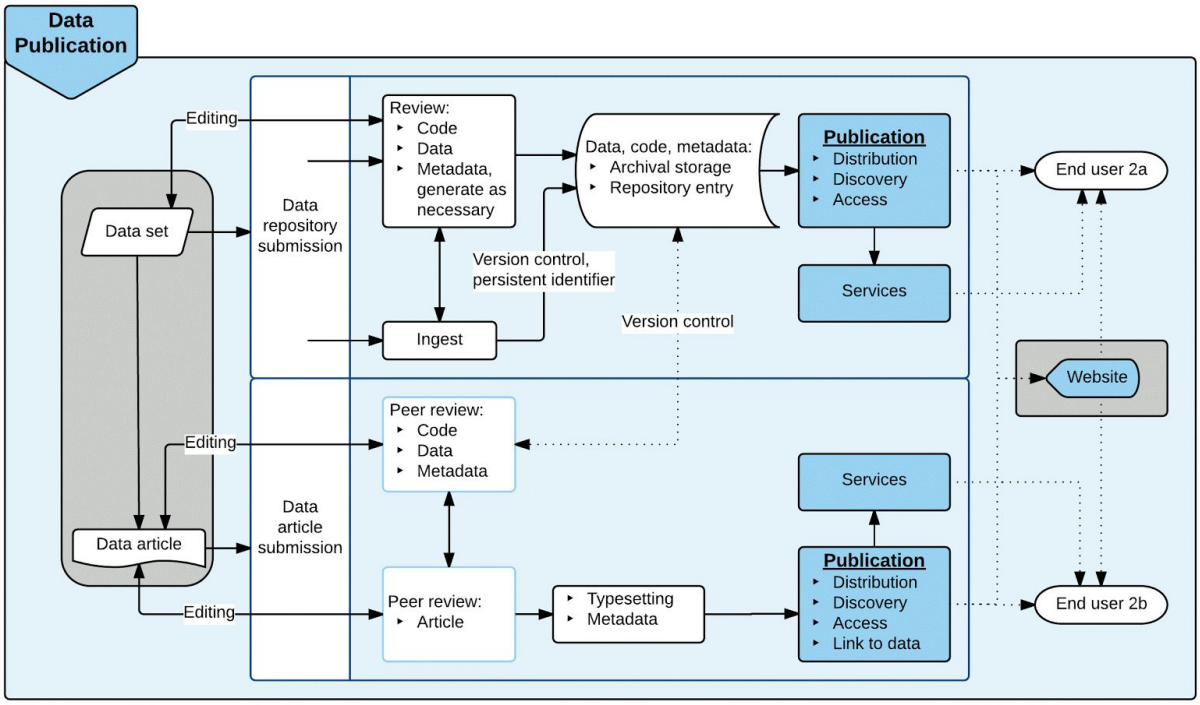
2-1

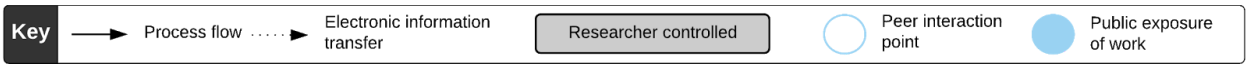

Upon request
Data & metadata
Website
Researcher X
Research methodology
Analyze/ interpret
Peer review
Journal article
End user 1

2-2


Data & metadata
Commented computer code
Method & data description
Upon request
Website
Article publication
Researcher Y
Research methodology
Analyze/ interpret
Data publication
End user 2

2-3


Data Publication
Editing
Review:
Code
Data
Metadata, generate as necessary
Data, code, metadata:
Archival storage
Repository entry
Publication
Distribution
Discovery
Access
End user 2a
Data set
Data repository submission
Version control, persistent identifier
Services
Version control
Ingest
Website
Peer review:
Code
Data
Metadata
Editing
Services
Data article
Data article submission
Publication
Distribution
Discovery
Access
Link to data
Peer review:
Article
Typesetting
Metadata
Editing
End user 2b



Following the presented reference model and the best practices, we envisage a workflow that both supports and results in all scholarly objects being connected, linked, citable, and persistent. This includes cross-linking of documentation, code, data, and journal articles in an integrated manner. Furthermore, in the ideal workflow, all of these objects would be well documented to enable others (researchers, citizen scientists, funders etc) to reuse the data to make novel discoveries. In this ideal workflow, information about the quality assurance procedures applied to the data, would also be available. 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 funding and academic reward systems will need to value data publications, data-driven secondary analysis and the reuse of existing data, as first class scholarly outputs. More attention to these outputs (i.e. more perceived value) by funders will be key to changing the current paradigm, allowing researchers to navigate smoothly between differently outputs and enabling reproducible research.
The results from this working group predominantly focus on “static” data publishing, i.e. data publishing as one of the final steps of the research process. To accommodate an emerging interest to connect the research workflow and data publishing, a second workflow analysis has also been initiated.

Full report available here: http://dx.doi.org/10.5281/zenodo.34542

Recommendations:

- Implement core components of the reference model according to the needs of the stakeholder concerned, by engaging with relevant stakeholders to understand requirements and current practices.
- 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. Make sure components are interoperable and that they build on best practices (and shared code).
- Follow existing standards where possible, to facilitate interoperability with the work of others.
- It is especially important to implement and adhere to standards for data citation, including the use of persistent identifiers (PIDs).
    - Linkage 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.
    - The use of PIDs can also enable linked open data functionality. It would help if which PID to use could be standardised!
- Document the role of each stakeholder in workflows and services for data publication.
    - 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.

- Lack of comprehensive documentation also impacts potential users of the services. Documentation could promote user engagement, as well as assist other stakeholders to identify trusted partners. The latter is in particular relevant when the service/platform/project is looking into an extension of the service (e.g. to cover upstream research workflows).

Additional information:

Maintenance statement:

The working group has been working intensively with various partners around the world to establish a reference model for data publishing. It has been adopted by several stakeholders and gained considerable interest in the community. Data Publishing offers incentives for researchers to share their data. By highlighting core components from the data publishing reference model, others can easily reuse the concepts and contribute to establishing a new data publishing paradigm. This could facilitate better uptake of Open Science practices across research communities.

Discussions are ongoing on as to whether RDA and WDS should continue with the Interest Group for data publishing. This IG would work to support uptake in the different communities in the mid and long term.

As data publishing is adopted more widely and solutions are emerging to cover "upstream" workflows (i.e. the actual research practices) with respect to data publishing, the IG would also perform a critical function by adapting and refining the core components highlighted in the reference model. This will require coordinated community feedback once further data publishing solutions have been put in place. The forum of the interest group could serve as a starting point for this refinement.

The final decision on the potential IG will be made at the Plenary 7 in Tokyo.

Adoption statement:

Many data organisations engaged in data publication have already adopted and are employing elements of the reference model. The public release of this model intends to make it available for relevance testing, community review, and amendment.

Adopting organizations include: *GigaScience*, Research Space, The University of Edinburgh DataShare, Elsevier Research Data Management Solutions, *Scientific Data*, the Jisc project "Giving Researchers Credit for their Data" and the Digital Curation Center (DCC). DCC states, "The DCC (Digital Curation Centre) is drawing on the reference model for a new title in its series of How-to guides, which offer practical guidance to research organisations on delivering support for Research Data Management. The new guide 'How to Document Workflows for Data Preservation and Publishing' is to be published in 2016, and will set out the benefits of workflow documentation, describe simple tools for doing so, and illustrate examples that build on the Working Group's report. "

Inclusion of term definitions in recognized resources such as the CASRAI[1] glossary and the RDA Data Foundations and Terminology tool[2] is a step forward in developing a common language wherein the reference model can be further developed.

---

[1] http://dictionary.casrai.org/Category:Research_Data_Domain

[2] http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page