Connecting data publication to the research workflow: a preliminary analysis

Sünje Dallmeir-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

IDCC17, Edinburgh, 22 February 2017
Understanding what principles and practices to reflect in our guidance material for organisations, to help them

- Provide effective research data services
- Understand options available
- Emerging standards and best practice
WDS-RDA Data Publishing Workflows Group

“A data publishing reference model comprised of generic components which need to be interoperable and seamlessly accessible”
WDS-RDA Data Publishing Workflows Group

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

http://dx.doi.org/10.5281/zenodo.34542
Definition of “Data Publishing Workflows”

“the activities and processes that lead to the publication of research data, associated metadata and accompanying documentation and software code on the Web.”
<table>
<thead>
<tr>
<th>Required elements for published product</th>
<th>Additional elements for increased context, quality, visibility</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary data entry with PID</strong></td>
<td><strong>Context</strong></td>
</tr>
<tr>
<td>Repository entry with persistent ID</td>
<td>Parallel descriptions / dependent objects</td>
</tr>
<tr>
<td>Metadata generation/review</td>
<td><strong>Quality</strong></td>
</tr>
<tr>
<td>Curation</td>
<td>QA/QC, Submission support</td>
</tr>
<tr>
<td>Distribution/discovery</td>
<td><strong>Visibility/Accessibility</strong></td>
</tr>
<tr>
<td></td>
<td>Rich documentation</td>
</tr>
<tr>
<td></td>
<td>Editing process</td>
</tr>
<tr>
<td></td>
<td>Enable indexing/machine readability</td>
</tr>
<tr>
<td></td>
<td>Link to results article</td>
</tr>
<tr>
<td></td>
<td>Enhanced curation by domain experts</td>
</tr>
<tr>
<td></td>
<td>Access support, value added services</td>
</tr>
<tr>
<td></td>
<td>Link to data article</td>
</tr>
<tr>
<td></td>
<td>Peer review</td>
</tr>
<tr>
<td></td>
<td>Push to impact/evaluation management platforms, metrics aggregators, author pages, 3rd parties</td>
</tr>
<tr>
<td></td>
<td>Submission support</td>
</tr>
</tbody>
</table>

Data access + steps to add context, quality and visibility

"Key Components of Data Publishing"


Recommendations for Organisations Establishing Workflows

<table>
<thead>
<tr>
<th>Required elements for published product</th>
<th>Additional elements for increased context, quality, visibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary data entry with PID</td>
<td><strong>Context</strong> Parallel descriptions / dependent objects</td>
</tr>
<tr>
<td></td>
<td><strong>Quality</strong> QA/QC, Submission support</td>
</tr>
<tr>
<td></td>
<td><strong>Visibility/Accessibility</strong></td>
</tr>
</tbody>
</table>

1. Start small using modular, open source and shareable components
2. Follow standards for interoperability
3. Facilitate data citation, e.g. through use of identifiers
4. Document roles, workflows and services
context - problem - approach - results - implications - next what about “upstream” data preparation?

Using RISE (Rans and Whyte, 2017)
what about upstream data preparation?

1) Is the intention to ‘publish’ data changing researcher practices?

2) What ‘active data management’ services support that?

3) How might our recommendations be adapted to fit?
context - problem - approach - results - implications - next call for examples

Requesting your input: Research workflows informed by the intent to publish data

Groups audience: RDA/WDS Publishing Data Workflows WG

Greetings, all,

The RDA-WDS Data Publishing Workflows Working Group is looking for examples of research workflows that demonstrate actions taken at earlier stages in the research lifecycle (i.e., prior to scholarly publication), which subsequently facilitate data sharing and publishing at a later stage.

We are collecting well-documented workflows that clearly describe or depict, using non-specialist terms and/or visuals: who does what, what the relevant steps are, what are their inputs and outputs, what tools are used to transform the data, etc. during steps relevant to data publishing. Pictures are wonderful.

We are analysing multiple workflows from diverse research areas, with the aim of describing a range of steps being taken to prepare data for publishing and archiving early in the project lifecycle. We welcome submission of workflows from all research areas. Workflows may capture the
Received 12 responses

- Mature data centre workflows x 5
- Proof-of-concept exemplars x 2
- Working prototypes x 2
- ‘Whole lifecycle’ models x 3
We reviewed how the previous recommendations applied to these examples
e.g. use of standards, identifiers
context - problem - **approach** - results - implications - next

Data Centre e.g. National Centre Atmospheric Research

Relatively **inter-dependent** workflows for preparation and archiving
Institutional e.g. Rspace to Edinburgh DataShare

Relatively independent workflows for preparation and archiving
1) Is the intention to ‘publish’ data changing researcher practices?

2) What ‘active data management’ services support that?

3) How might our recommendations be adapted to fit?

Using RISE (Rans and Whyte, 2017)
Whole Lifecycle ‘Seamless Solutions’ e.g. OSF

What happens to workflows dependencies (?)

Mathew Spitzer
An Open Science Framework for Solving Institutional Challenges: Supporting the Institutional Research Method
context - problem - **approach** - **results** - implications - next

**Whole Lifecycle ‘Seamless Solutions’ e.g. Elsevier**

---

**What happens to workflows dependencies (?)**

*Mathew Spitzer*

**An Open Science Framework for Solving Institutional Challenges: Supporting the Institutional Research Method**
What Principles Guide Choice of Service Components?

Help institutions to integrate components in a ‘loosely coupled’ manner to avoid vendor lock-in

Using RISE (Rans and Whyte, 2017)
"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"
definition: Active Data Management

“service used to create or transform digital objects for the purposes of research”
definition: Preservation

“service offering to ensure digital objects meet a defined level of FAIRness - findability, accessibility, interoperability, and reusability - for a designated community and period of time”
definition: Publication

“services offering to enhance digital objects FAIRness by reviewing their quality on specified criteria, or connecting them to additional metadata”
definition: Guidance

“services offering practical guidance on choosing or using the above services”
Principles (1 of 5): 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”
Principles (2 of 5): Preservation and Publication

“should publish policies stating what digital object types they accept, for what communities, and on what terms and conditions”
Principles (3 of 5): Active Data Management, Preservation and Publication

“should make openly available sufficient metadata to enable reuse of their outputs, including all terms and conditions for third-party access and reuse”
Principles (4 of 5): Active Data Management, Preservation and Publication

“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”
Principles (5 of 5): Guidance

“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”
Next Steps

• RDA Data Publishing Interest Group will discuss potential for new Working Group on FAIR workflows.
• For FAIR data to be published, to what extent does the production workflow need to be FAIR?
• Further work to understand how researchers' perception of ‘data publication’ affects practices.
Comments, Questions?

- Do institutions need guidelines like this?
- If so, what practical steps are needed?
Thank you...