Data Publishing Workflows: Models

RDA-WDS Publishing Data Workflows Working Group

WG Collaborative, December 2015

Today's session

- Where we've been
- Where we're going
- Getting there together: Domains

Where we've been

- What is data publishing
- Models in data publishing
- Recommendations
- Challenges

http://bit.ly/1TvGe9v

Data Publishing

““Research data publishing is ***the release of research data, associated metadata, accompanying documentation, and software code*** *(in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way.*

Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are ***well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable*** *– all aspects of data publishing that are important for future reuse of data by third party end-users.””*

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

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

Subjects of Review

Guidelines for data publication, e.g.,

- ENVRI reference model
- PREPARDE

Data journals, e.g.,

- Scientific Data
- F1000

Repositories, e.g.,

- Domain
  - National Snow & Ice Data Center (NSIDC)
  - ICPSR (Social Sciences)
- General
  - Dryad
  - Arkivum + Figshare
- Institutional
  - Stanford Digital Repository
  - Data Repository for the University of Minnesota (DRUM)

https://zenodo.org/record/34542#.VmVJqMrWlqc

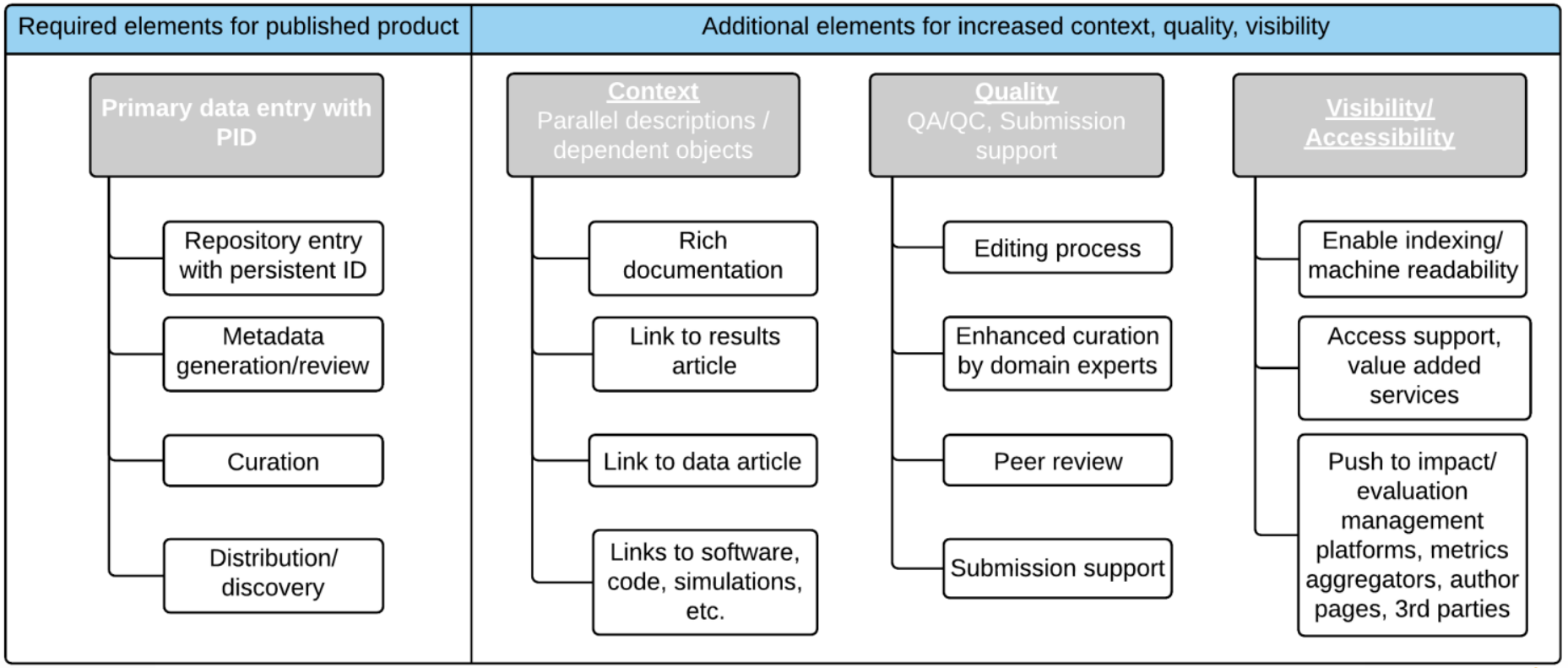
Elements of Analysis

- Discipline
- Function of workflow
- The assignment of persistent identifiers (PIDs) to datasets
- The PID type used -- e.g., DOI, ARK, etc.
- Peer review of data (e.g., by researcher and by editorial review)
- Curatorial review of metadata (e.g., by institutional or subject repository)
- Technical review and checks (e.g., for data integrity at repository/data centre on ingest)
- Discoverability: Was there indexing of the data, and if so, where?
- Formats covered
- Persons/Roles involved, e.g., editor, publisher, data repository manager, etc.
- Links to additional data products (data paper; review; other journal articles) or "stand-alone" product
- Links to grants, usage of author PIDs
- Whether data citation was facilitated
- Whether the data life cycle was referred to
- Standards compliance

Elements of Analysis

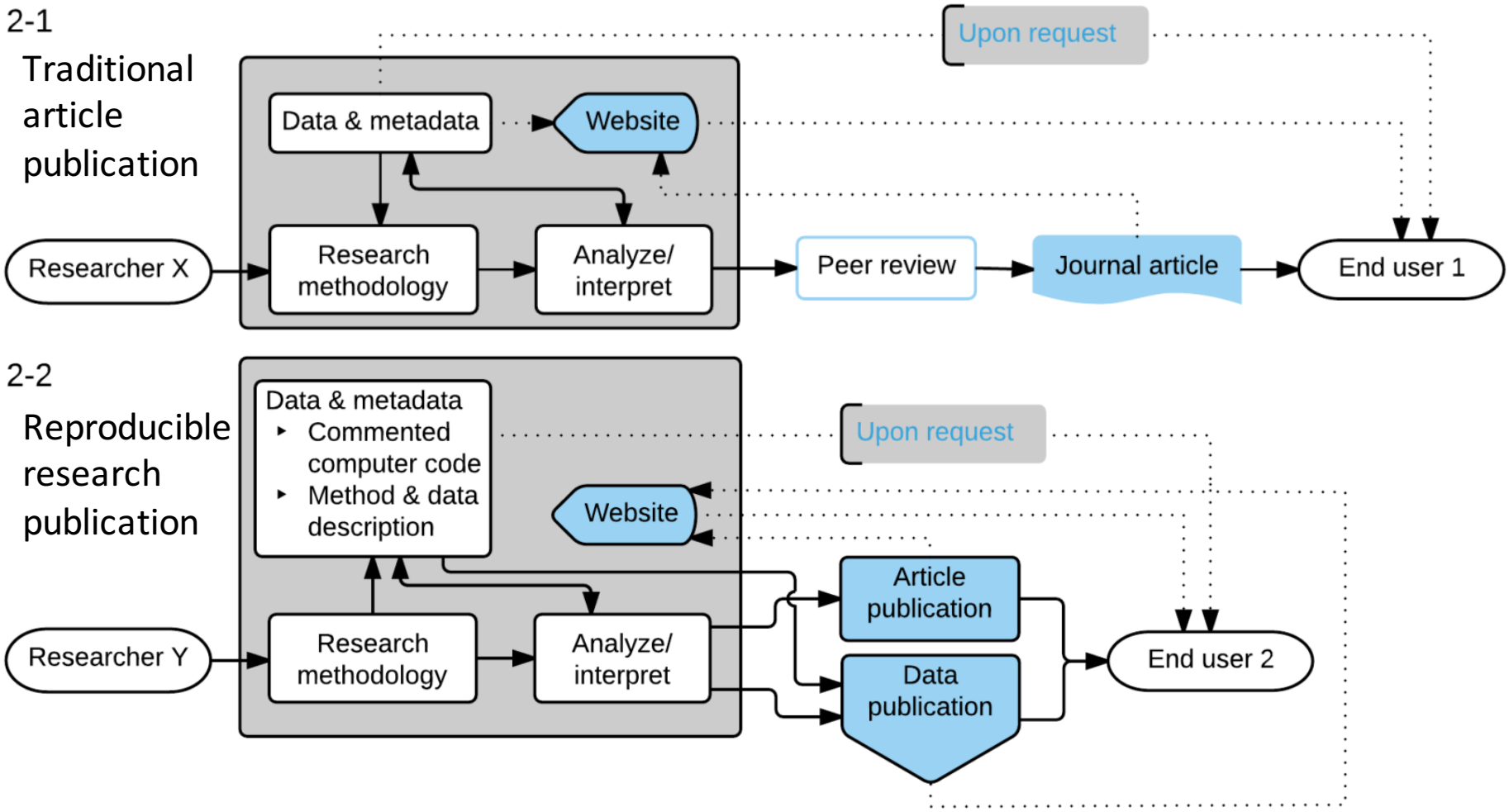
- **Discipline**
- Function of workflow
- The assignment of persistent identifiers (PIDs) to datasets
- The PID type used -- e.g., DOI, ARK, etc.
- Peer review of data (e.g., by researcher and by editorial review)
- Curatorial review of metadata (e.g., by institutional or subject repository)
- Technical review and checks (e.g., for data integrity at repository/data centre on ingest)
- Discoverability: Was there indexing of the data, and if so, where?
- **Formats covered**
- Persons/Roles involved, e.g., editor, publisher, data repository manager, etc.
- Links to additional data products (data paper; review; other journal articles) or "stand-alone" product
- Links to grants, usage of author PIDs
- Whether data citation was facilitated
- Whether the data life cycle was referred to
- Standards compliance

Key Components of Data Publishing



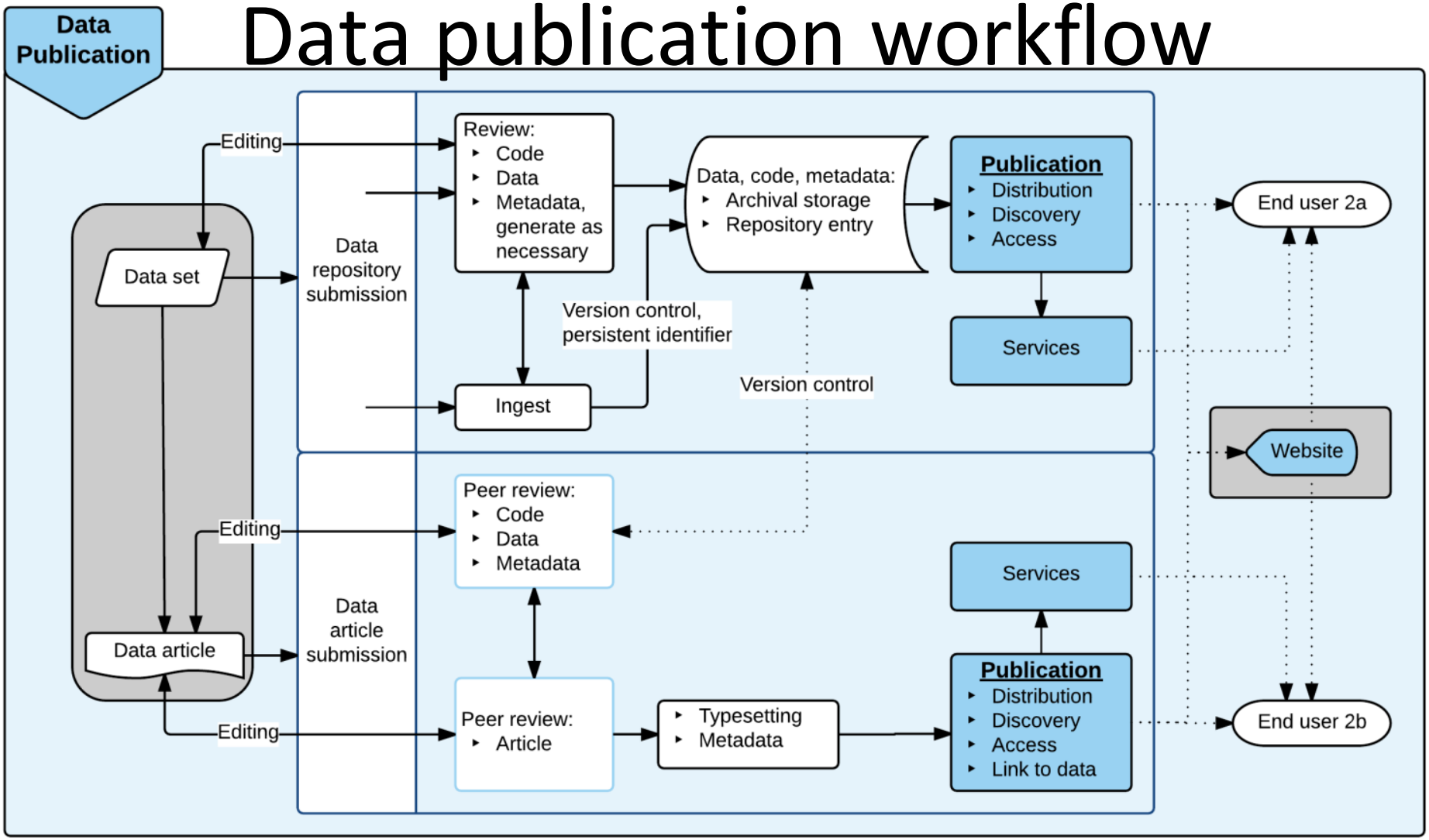
https://zenodo.org/record/34542#.VmVJqMrWlqc

Publication workflows



https://zenodo.org/record/34542#.VmVJqMrWlqc

Data publication workflow



Recommendations

- **Start small and build open source/shareable 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.
- **Follow standards** whenever available to facilitate interoperability and to permit extensions based on the work of others using the same standards.
- Implement and adhere to **standards for data citation**, including the use of persistent identifiers (PIDs). Linkages between data and publications can be automatically harvested if DOIs for data are used routinely in papers. The use of researcher PIDs such as ORCID can also establish connections between data and papers or other research entities such as software. The use of PIDs can also enable linked open data functionality.
- **Document** roles, workflows and services.

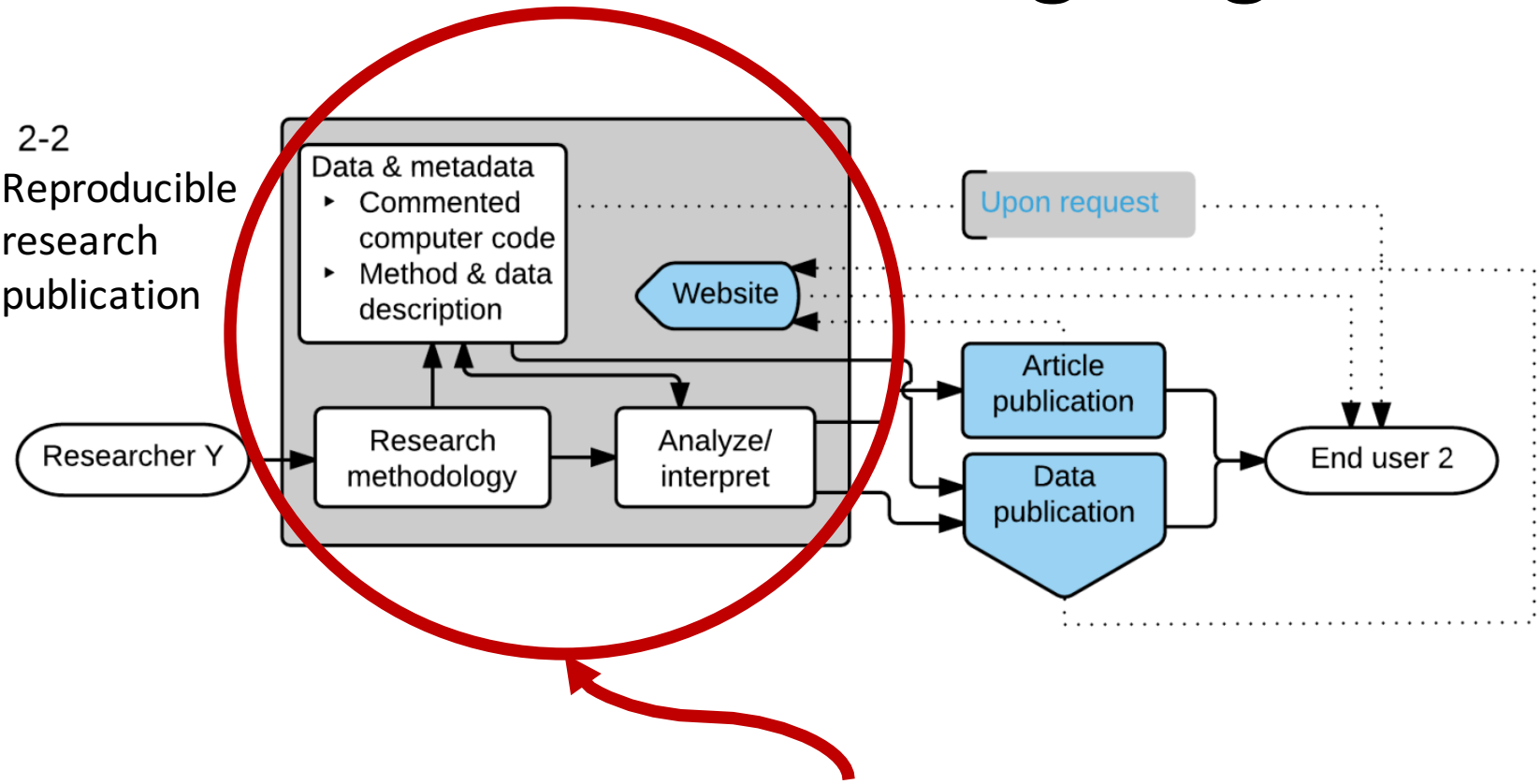
Challenges

- Bi-directional linking.
- Software management.
- Version control / dynamic data
- Sharing restricted-use data.
- Role clarity.
- Business models.
- Data citation support.
- Metrics.
- Incentives.

Where we're going

Where we're going

2-2
Reproducible research publication



Here

Where we're going

- How does the intent to make research data public inform the research workflow?
- Can we extend data publication to cover the research workflow better/at all?
  - Who does that? Where? How?
  - What are the challenges?

Intent to publish research data informing the research workflow

Traditional research workflows

- Searching literature (know any good references?)
- Looking for diverse domain examples

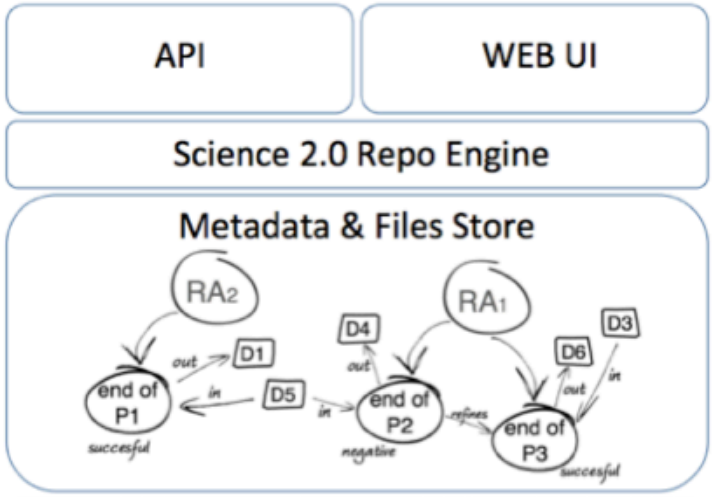
Research workflows integrating data publication

- Canvassing community: diverse domains desired
- http://bit.ly/1N48NHf

SEAMLESS ASTRONOMY

http://projects.iq.harvard.edu/seamlessastronomy/home

Supported by Microsoft Research



Open Science Framework

Taverna

Massimiliano Assante, Leonardo Candela, Donatella Castelli, Paolo Manghi and Pasquale Pagano, *Science 2.0 Repositories: Time for a Change in Scholarly Communication*, DOI: 10.1045/january2015-assante http://nemis.isti.cnr.it/groups/infrascience

RDA RESEARCH DATA ALLIANCE

ICSU WORLD DATA SYSTEM

Getting there together

RDA Publishing Workflows: Research Workflows

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 life-cycle (i.e. prior to scholarly publication), which subsequently facilitate data sharing and publishing at a later stage. We welcome submission of workflows from all research areas. Workflows may capture the steps taken early in the research life cycle by any relevant stakeholder (e.g. data producer, research administrator, research data support service, repository, publisher etc), to facilitate and encourage good research data management practices, data sharing, publishing and archiving.

Workflow authors will be acknowledged as contributors in any resulting scholarly works.

Due 6Jan2016

http://bit.ly/1N48NHf

RESEARCH DATA ALLIANCE

ICSU WORLD DATA SYSTEM

What we're asking:

Describe the research workflow & how it integrates practices that enable data publication:

1) Roles - who is involved in the stage
2) Inputs - outputs from previous stages
3) Actions - steps / activities, both optional and required
4) Outputs - products that become inputs to next stages
5) Tools - both current and desired, as relevant

Describe the results of the workflow:

1) Achieved
2) Yet to be achieved & what is needed

http://bit.ly/1N48NHf

RDA RESEARCH DATA ALLIANCE

ICSU WORLD DATA SYSTEM

Extending data publication to cover the research workflow

- Current practices
- Current tools
- Nascent opportunities
  - Who does that? Where? How?
  - What are the challenges?

Invited contributions

Submit your research workflow:

http://bit.ly/1N48NHf

Working paper:

http://bit.ly/1Km4hlZ

Thank you!

RESEARCH DATA ALLIANCE

ICSU WORLD DATA SYSTEM