Data Curation: Technical Challenges Facing Repositories

Brianna Marshall

Jan. 9, 2014

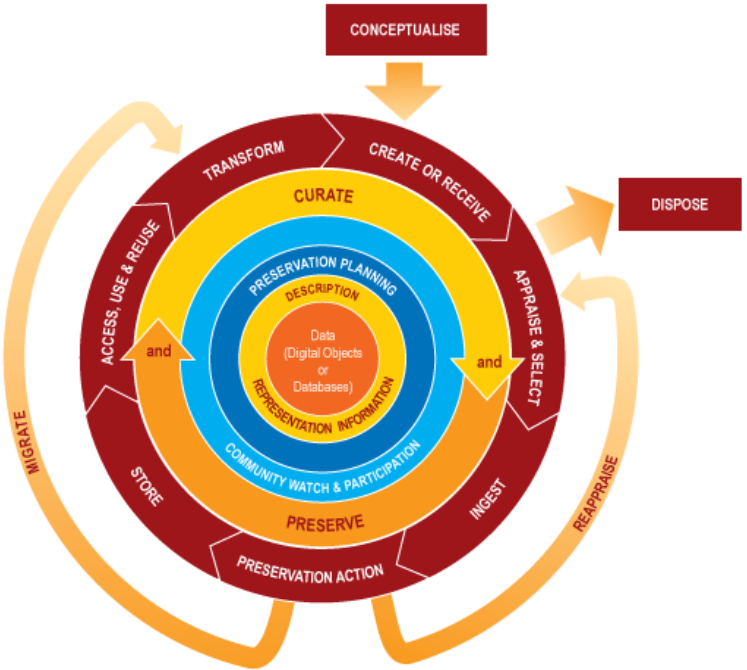
Defining digital curation

“Maintaining and adding value to a trusted body of digital information for current and future use; specifically... the active management and appraisal of data over the life-cycle of scholarly and scientific materials.”

-Digital Curation Centre

DCC Curation Lifecycle Model

1. Create or receive
2. Appraise and select
3. Ingest
4. Preservation action
5. Store
6. Access, use, and reuse
7. Transform



Broad data curation challenges

- Data are heterogeneous
  - Formats
  - Size
  - Complexity
  - Fixity

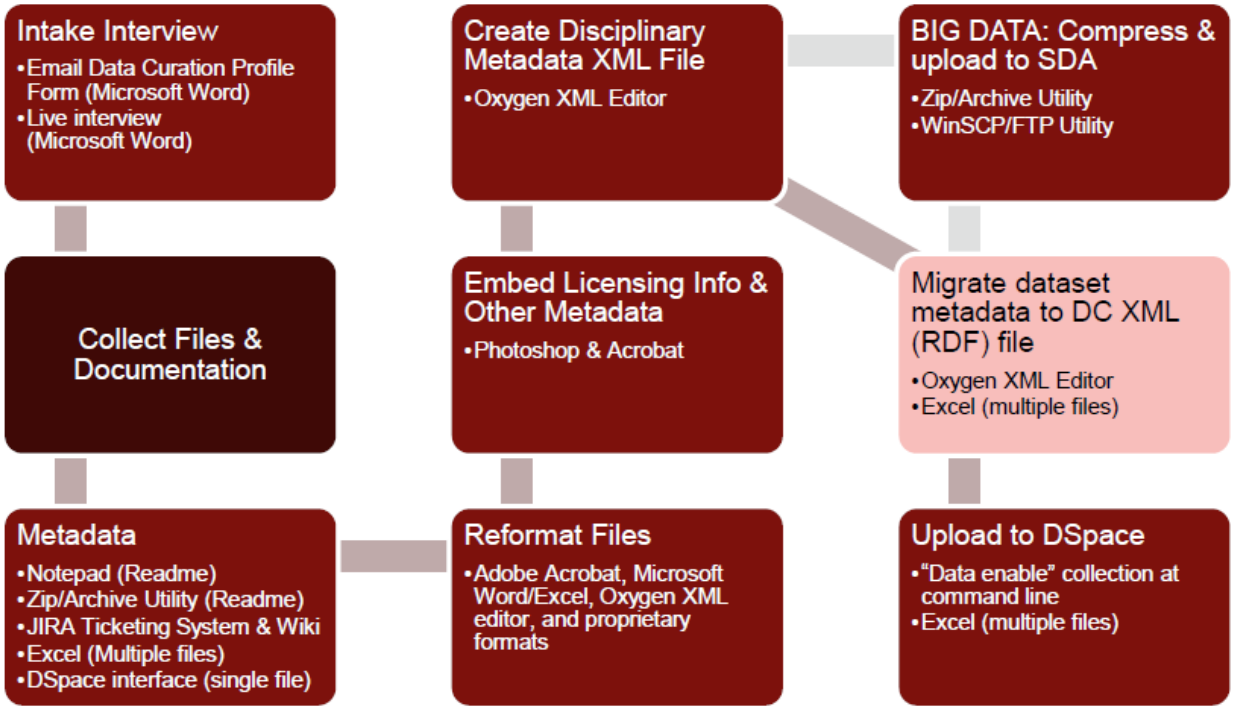
Broad data curation challenges

- For most IR platforms, data is an afterthought
  - Created initially for electronic theses and dissertations (ETDs), other publications
  - Not intended for active data collection

The real problem

So much of the active data curation that takes place is *outside* the repository.

Example workflow (IU)

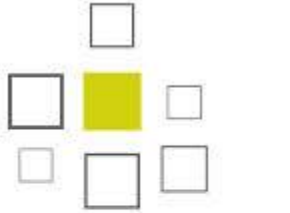


Konkiel, Snajdr, Marshall; LITA 2013

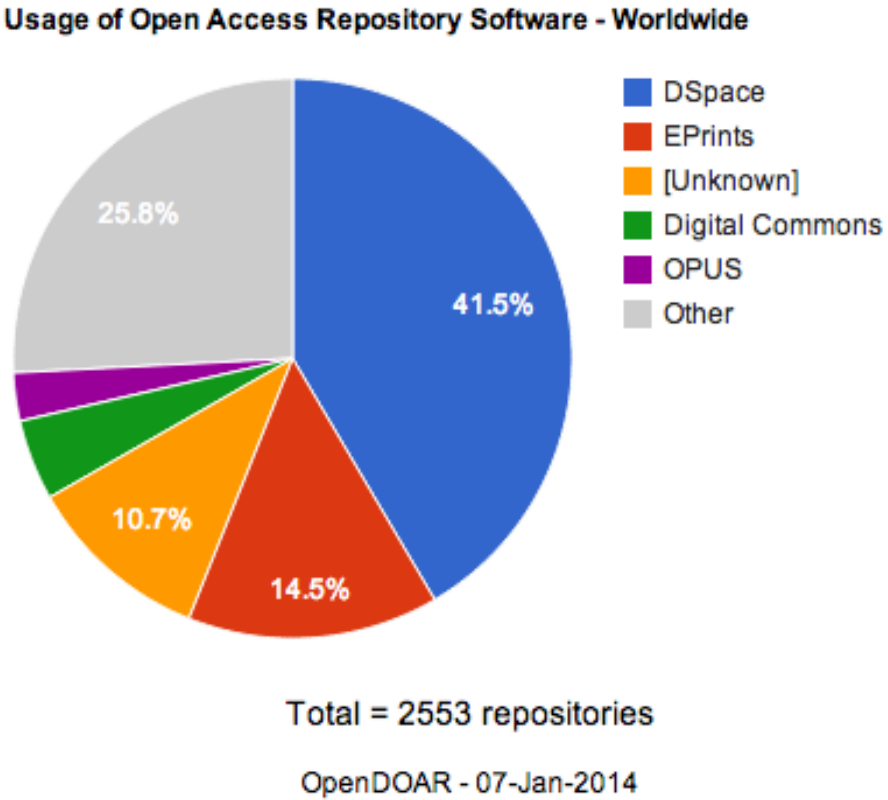
Worth pondering...

- Should we build additional technical functionality into a repository?
- Or just create better workflows between repositories and data curation tools?
  - Electronic lab notebooks
  - Data Curation Profiles Toolkit
  - DMPTool

Existing repository platforms



D SPACE



http://www.opendoar.org/

What are repositories doing well?

- Storage & access
- Indexing items for search engine discoverability
- Authentication / authorization
- Time-based access control (embargoes)

Researchers need more, though

- Persistent identifiers
  - For datasets
  - For data creators (disambiguation)
- Metadata
- Altmetrics
- Improved data ingest

Persistent identifiers

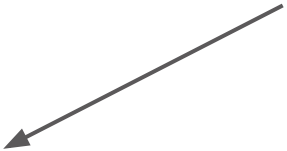
- DOIs
  - Growing push to cite datasets (DataCite)
  - Data sharing proven to increase citations (Piowawar & Vision, 2013)
  - DSpace 4.0 now offers DOIs in addition to handles

Author identifiers (disambiguation)

Metadata

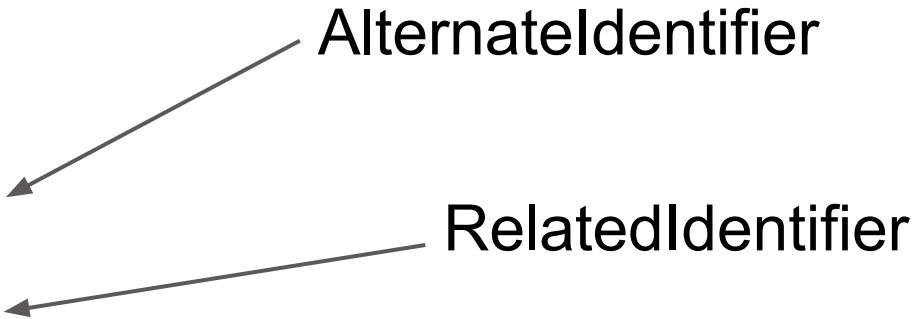
- Currently, Dublin Core is default
- Why not the DataCite metadata schema?

ID	Property
6	Subject (with scheme attribute)
7	Contributor (with type attribute)
8	Date (with type attribute)
9	Language
10	ResourceType
11	AlternateIdentifier (with type attribute)
12	RelatedIdentifier (with type and relation type attributes)
13	Size
14	Format
15	Version
16	Rights
17	Description (with type attribute)

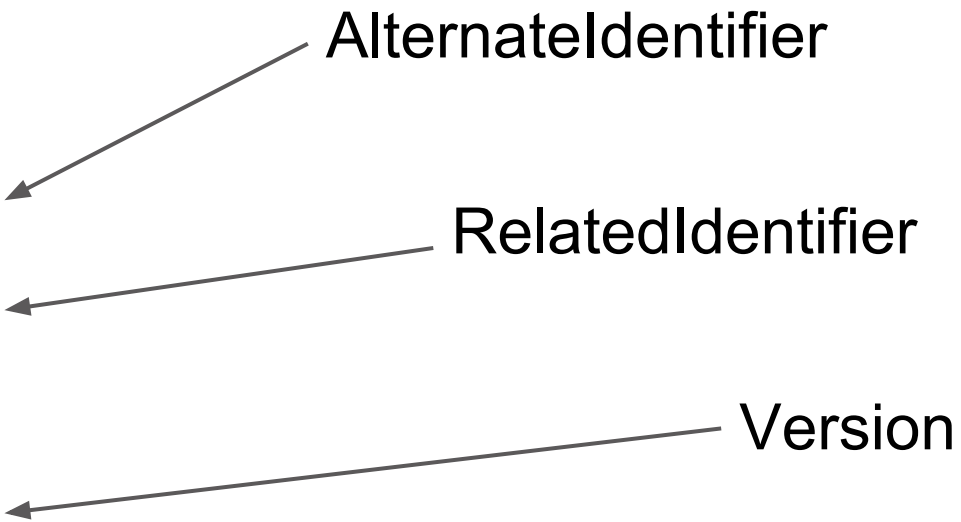


AlternateIdentifier

ID	Property
6	Subject (with scheme attribute)
7	Contributor (with type attribute)
8	Date (with type attribute)
9	Language
10	ResourceType
11	AlternateIdentifier (with type attribute)
12	RelatedIdentifier (with type and relation type attributes)
13	Size
14	Format
15	Version
16	Rights
17	Description (with type attribute)



ID	Property
6	Subject (with scheme attribute)
7	Contributor (with type attribute)
8	Date (with type attribute)
9	Language
10	ResourceType
11	AlternateIdentifier (with type attribute)
12	RelatedIdentifier (with type and relation type attributes)
13	Size
14	Format
15	Version
16	Rights
17	Description (with type attribute)



ImpactStory.

Tracks:

- Scholarly impact
- Bookmarking
- Blog mentions
- Popular impact
- News outlet mentions

(Konkiel, 2013)

Follow Download as

Introduction
Materials and Methods
Results
Discussion
Conclusions
Supplemental Information
Additional Information and Declarations

Peer Review history
Questions 6
Links

Subject areas

Bioinformatics
Science Policy

Data reuse and the open data citation advantage

Heather A. Piwowar [1,2], Todd J. Vision [1,2,3]

PubMed ID: 24109559

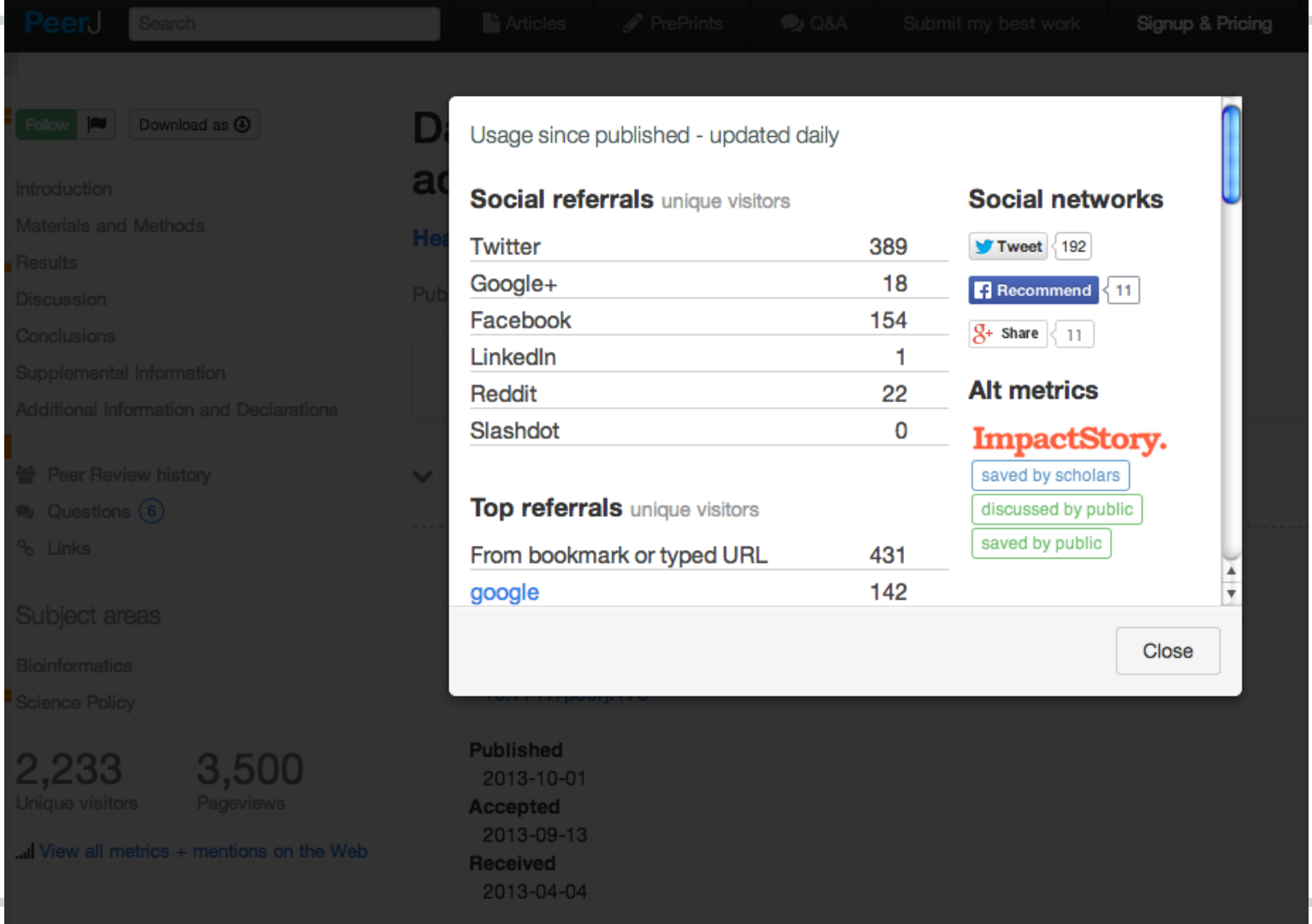
Note that a PrePrint of this article also exists.

Author and article information

[1] National Evolutionary Synthesis Center, Durham, NC, USA
[2] Department of Biology, Duke University, Durham, NC, USA
[3] Department of Biology, University of North Carolina - Chapel Hill, Chapel Hill, NC, USA

DOI
10.7717/peerj.175

Published
2013-10-01
Accepted
2013-09-13
Received
2013-04-04


PeerJ
Search
Articles
PrePrints
Q&A
Submit my best work
Signup & Pricing
Follow
Download as
Introduction
Materials and Methods
Results
Discussion
Conclusions
Supplemental Information
Additional Information and Declarations
Peer Review history
Questions 6
Links
Subject areas
Bioinformatics
Science Policy
2,233
Unique visitors
3,500
Pageviews
View all metrics + mentions on the Web
Usage since published - updated daily
Social referrals unique visitors
Twitter 389
Google+ 18
Facebook 154
LinkedIn 1
Reddit 22
Slashdot 0
Top referrals unique visitors
From bookmark or typed URL 431
google 142
Social networks
Tweet 192
Recommend 11
Share 11
Alt metrics
ImpactStory.
saved by scholars
discussed by public
saved by public
Close
Published
2013-10-01
Accepted
2013-09-13
Received
2013-04-04

Data ingest

- Current workflows are clunky and inconvenient
  - Command line or manual ingest
  - Batch ingest usually undertaken by institution
- Could benefit from integrations with researchers' local hard drive

So how did we get here?

- Dealing with the data deluge is new, overwhelming - no easy answers
- Institutions become “locked in” to the quirks of their particular repository system
- Takes developer time, either to adapt others’ code or create from scratch

New data curation models

- Hydra
  - ScholarSphere (Penn State)
  - Curate/Shared IR collaboration
    - "A Next Generation IR solution with emphasis on flexibility and responsiveness to emerging needs. Of particular priority will be the ability to accept and manage highly complex objects and **research data**." (https://wiki.duraspace.org/display/hydra/Shared+IR+project)
    - SHARE (SHared Access Research Ecosystem)
- Separate repositories for digital content
  - Purdue University
  - IUPUI
- Partnerships with external data repositories (Dryad)

Final thoughts

- Repositories are a crucial aspect of data curation, but they can’t do it all
- Current IRs enable storage and access… less so other aspects of data curation
- Overall, repository functionality must continue to adapt and anticipate researchers’ needs
- There is no one perfect solution (yet)

Resources

Chapman, J. W., Reynolds, D., and Shreeves, S. (2009). “Repository Metadata: Approaches and Challenges.” *Cataloging & Classification Quarterly*, 47:309–325.

De Castro, P., and Warner, Simeon (2013). “ORCID Implementation in Open Access Repositories and Institutional Research Information Management Systems.” Presentation at Open Repositories 2013. Charlottetown, PEI, Canada.

Konkiel, S., and Halliday, J. (2013). “IUScholarWorks, Statistics, and Altmetrics.” Presentation at the Digital Library Brown Bag series. https://scholarworks.iu.edu/dspace/handle/2022/16980

Konkiel, S. (2012). "Robust 'Altmetrics' as a Framework for Measuring Item Usage and Researcher Impact in Institutional Repositories." Poster presentation at 2012 LITA National Forum. Columbus, OH, USA. 5-7 October 2012.

“Metadata and Repository Services for Research Data Curation.” DuraSpace Community Webinar, October 2013.

Piwowar, H., and Vision, T. (2013). “Data reuse and the open data citation advantage.” *PeerJ*1:e175 http://dx.doi.org/10.7717/peerj.175

Thank you!

Brianna Marshall

bhmarsha@indiana.edu

@notsosternlib