Research Data as a (not just) storage challenge

Stefan Kramer Associate Director for Research Data Services American University Library skramer@american.edu

Library of Congress Designing Storage Architectures Meeting, Sep. 9, 2015



Source: http://d7.library.gatech.edu/research-data/home

What do we mean by "research data"?

"Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings...."

(from OMB Circular A-110, Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations http://www.whitehouse.gov/omb/circulars_a110#36)



The problems of research data accessibility

- Availability of research data over time – from bad to terrible: in one study of 516 life science publications, only 37% of the data from two-year-old publications could be located, and only 7% for 20-year-old publications
- Problem for validation & replication of scientific claims, or repurposing/reuse of already collected data



26 PEER REVIEW CONGRESS

From: Timothy H. Vines et al. *How Does the Availability of Research Data Change With Time Since Publication?* Plenary session abstracts of the Seventh International Congress on Peer Review and Biomedical Publication, Sep. 8-10, 2013, Chicago, IL. (http://www.peerreviewcongress.org/abstracts_2013.html#31)

Storage of "the raw data" (and knowing how much?) is *part* of the challenge...

JESLIB 2012; 1(2): 63-78 doi:10.7191/jeslib.2012.1008

Figure 7: Responses to the question: "Given the NSF expectation to share data with other researchers, how much data would you intend to share?"



From article "Prepared to Plan? A Snapshot of Researcher Readiness to Address Data Management Planning Requirements" – surveyed NSF PIs at Cornell U.

... but there's also the metadata

From poster presentation Research study is conceived and The Lifecycle of Social Science Research Data: Improved Idea planned, Discovery through better Metadata and Search Tools methodologies at the 2011 ASIS&T Research Data Access and Preservation selected, funding sources explored Summit - Source: http://hdl.handle.net/1813/22464 By search tools utilizing Metadata metadata from data stores. Existing data new research data can sources are sought Search & become available for finding and explored – also and exploring by researchers happens for basic Discovery research needs **Research data management** Research instruments Final datasets are are designed; data deposited for long-term Design & are collected through preservation - e.g., into Archiving surveys, interviews, institutional or domain Collection etc. - and from repository existing data sources Final datasets are made Collected data are publicly accessible – e.g. merged, cleaned, Analysis & via researcher's and/or analyzed, sub-Publication department's and/or setted, coded, Processing journal publisher's web harmonized, linked, site etc.



The crucial role of metadata in research data discoverability and usability

- Many data file formats are binary and proprietary – unlike documents, do not lend themselves to fulltext indexing
- Even where full-text indexing possible, often little/any meaningful content to index!
- This makes good metadata, as a subset of data documentation, vital
 - Descriptions of contents, formats, geographic and time coverage and granularity, relations to other files, etc.

🛄 sp426:	2-Data.sav - SPSS	Data Editor								_ 8 ×
File Edit	View Data Tran	sform Analyze Gr	aphs Utilities Add-	ons Window He	lp					
1:Q1		1								
	Q1	Q2	Q3	Q4	Q5A	Q5B	Q5C	Q6D	Q6A	Q6B
1	1.00	1.00	1.00	1.00	10.00	7.00	96.00	96.00	52.00	
2	5.00	3.00	5.00	5.00	10.00	96.00	96.00	96.00	7.00	
3	1.00	3.00	1.00	4.00	10.00	96.00	96.00	96.00	7.00	
4	4.00	4.00	3.00	4.00	9.00	96.00	96.00	96.00	7.00	
5	4.00	4.00	4.00	4.00	7.00	96.00	96.00	96.00	28.00	
6	2.00	4.00	6.00	3.00	10.00	96.00	96.00	96.00	60.00	
7	2.00	2.00	1.00	3.00	10.00	96.00	96.00	96.00	60.00	
8	4.00	5.00	4.00	4.00	10.00	7.00	96.00	96.00	60.00	_
9	5.00	4.00	6.00	4.00	10.00	47.00	96.00	96.00	60.00	_
10	2.00	1.00	1.00	2.00	7.00	96.00	96.00	96.00	60.00	_
11	1.00	4.00	5.00	2.00	10.00	7.00	96.00	96.00	60.00	_
12	4.00	5.00	3.00	4.00	10.00	7.00	96.00	96.00	28.00	_
13	4.00	3.00	6.00	2.00	10.00	7.00	96.00	96.00	18.00	
14	2.00	2.00	2.00	3.00	10.00	7.00	96.00	96.00	60.00	_
15	4.00	3.00	3.00	3.00	10.00	96.00	96.00	96.00	60.00	
16	4.00	4.00	5.00	4.00	7.00	10.00	96.00	96.00	28.00	_
17	5.00	5.00	5.00	4.00	7.00	96.00	96.00	96.00	10.00	_
18	2.00	2.00	6.00	2.00	7.00	10.00	96.00	96.00	60.00	
19	1.00	4.00	5.00	2.00	10.00	7.00	96.00	96.00	60.00	_
20	2.00	3.00	6.00	2.00	10.00	52.00	96.00	96.00	60.00	
21	1.00	4.00	4.00	1.00	10.00	47.00	96.00	96.00	60.00	
22	4.00	4.00	6.00	3.00	3.00	10.00	96.00	96.00	47.00	
23	1.00	5.00	5.00	4.00	10.00	7.00	96.00	96.00	47.00	_
24	4.00	4.00	4.00	4.00	10.00	7.00	96.00	96.00	28.00	
25	2.00	2.00	3.00	2.00	10.00	3.00	7.00	96.00	60.00	_
26	3.00	6.00	3.00	3.00	10.00	7.00	96.00	96.00	28.00	
27	4.00	5.00	5.00	4.00	10.00	7.00	96.00	96.00	60.00	
28	3.00	3.00	4.00	2.00	7.00	96.00	96.00	96.00	60.00	_
29	5.00	4.00	5.00	5.00	7.00	96.00	96.00	96.00	60.00	
30	2.00	1.00	6.00	2.00	10.00	96.00	96.00	96.00	60.00	
31	1.00	1.00	1.00	1.00	10.00	7.00	50.00	96.00	3.00	
32	2.00	2.00	1.00	2.00	10.00	52.00	96.00	96.00	60.00	
	4.00	, 5.00	1.00	4.00	10.00	7.00	96.00	96.00	60.00	×11*
🛃 Start	🛛 🥥 Electronic Data Center : 🔯 ICPSR images		🛗 sp426	💼 sp4262-Data.sav - SP		To Output1 - SPSS Viewer		🕘 🄗 2:15 PM		

Source of screen snapshot: http://einstein.library.emory.edu/icpsr_to_SPSS_ASCII.shtml

RDA Repository Platforms for Research Data IG

Repository Platforms for Research Data



Group details

Status: Recognised & Endorsed Chair(s): David Wilcox, Stefan Kramer, Ralph Müller-Pfefferkorn Secretariat Liaison: Herman Stehouwer TAB Liaison: Jamie Shiers and Peter Wittenburg Case Statement: Download

Total Members: 69 Total Posts: 34

Institutions, developers, and other members of the research data community struggle to choose, utilize, deploy or develop the best possible repository platform to meet particular research data needs. The Repository Platforms for Research Data Interest Group will gather and analyze research data use cases in the context of repository platform requirements. The primary deliverable will be a matrix relating use cases with functional requirements for repository platforms. The primary target audience for the aforementioned matrix consists of developers and service providers of repository software. The functional requirements will influence the development of repository software and related services to better serve the use cases of the research data community.

https://rd-alliance.org/groups/repository-platforms-research-data.html

What *might* be on a research data repository wishlist (aside from storing the data)?

Automated embargo - START of access

Automated embargo - END of access

assigning and searching metadata elements of geographic and time extent and granularity

assignment of DOIs

expression of relationship of files in a study (data, input, output, etc.)

automatic creation of checksums, digital fingerprints, or other fidelity verification

identify and allow upload of a file containing metadata (e.g., DDI XML) that is specially parsed

must require file type ID upon ingest (e.g., that .DOC = MS Word)

delegation of submitter role (e.g., grad. asst. for faculty member)

presentiation of submission licenses that are data-specific, stating statistical disclosure control, privacy, confidentiality issues

versioning of submission licenses

tracking of submission license version consented to by submittor

field for ORCIDs

web-based submission process

field for funder identification (FundRef)

display suggested citation for data

export citation to bibliographic software (EndNote, RefWorks, Citavi, etc.)

repeating field for links to related publications (reports/articles based on the data's analysis)

visualization: interactive charts and tables

export/download datasets in multiple formats (Excel, CSV, Stata, SPSS, ...)

visualization: static charts and tables

express periodicity of data collection, if regular



different access levels for different digital objects in same study, e.g. documentation=public, dataset=requiring login or authorization clear presentation of access levels to user, e.g., through "traffic light" colors

Note: this is the presenter's own opinion, not that of the aforementioned RDA Interest Group