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

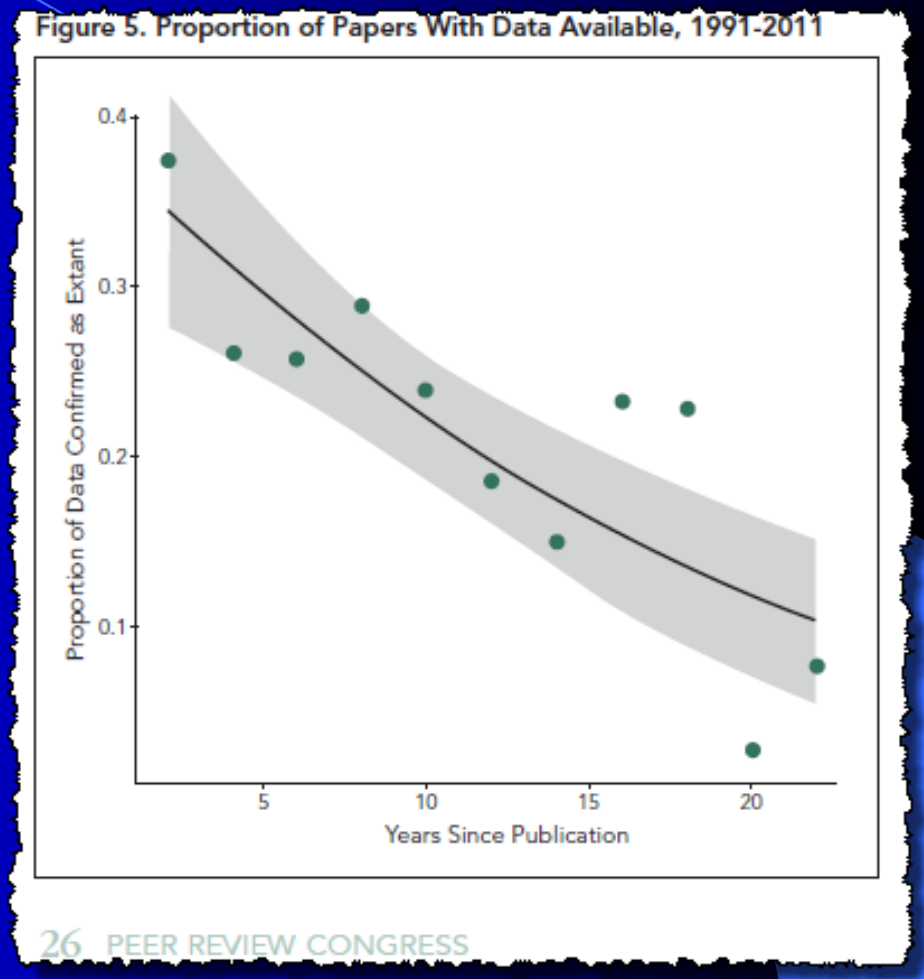
What do we mean by “research data”?

“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

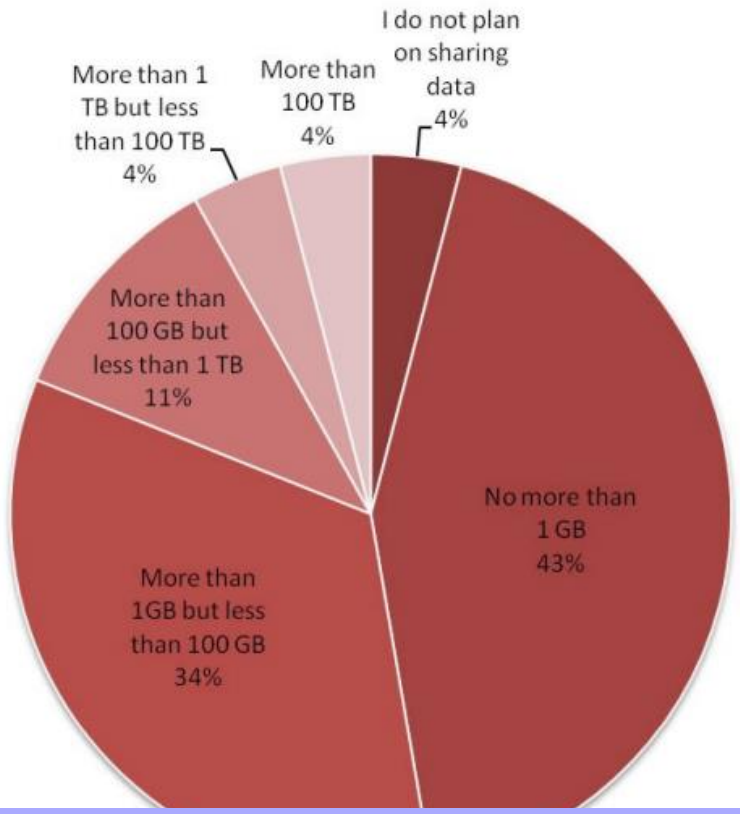


From: Timothy H. Vines et al. *How Does the Availability of Research Data Change With Time Since Publication?* Plenary session abstracts of the Seventh International Congress on Peer Review and Biomedical Publication, Sep. 8-10, 2013, Chicago, IL. (http://www.peerreviewcongress.org/abstracts_2013.html#31)

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

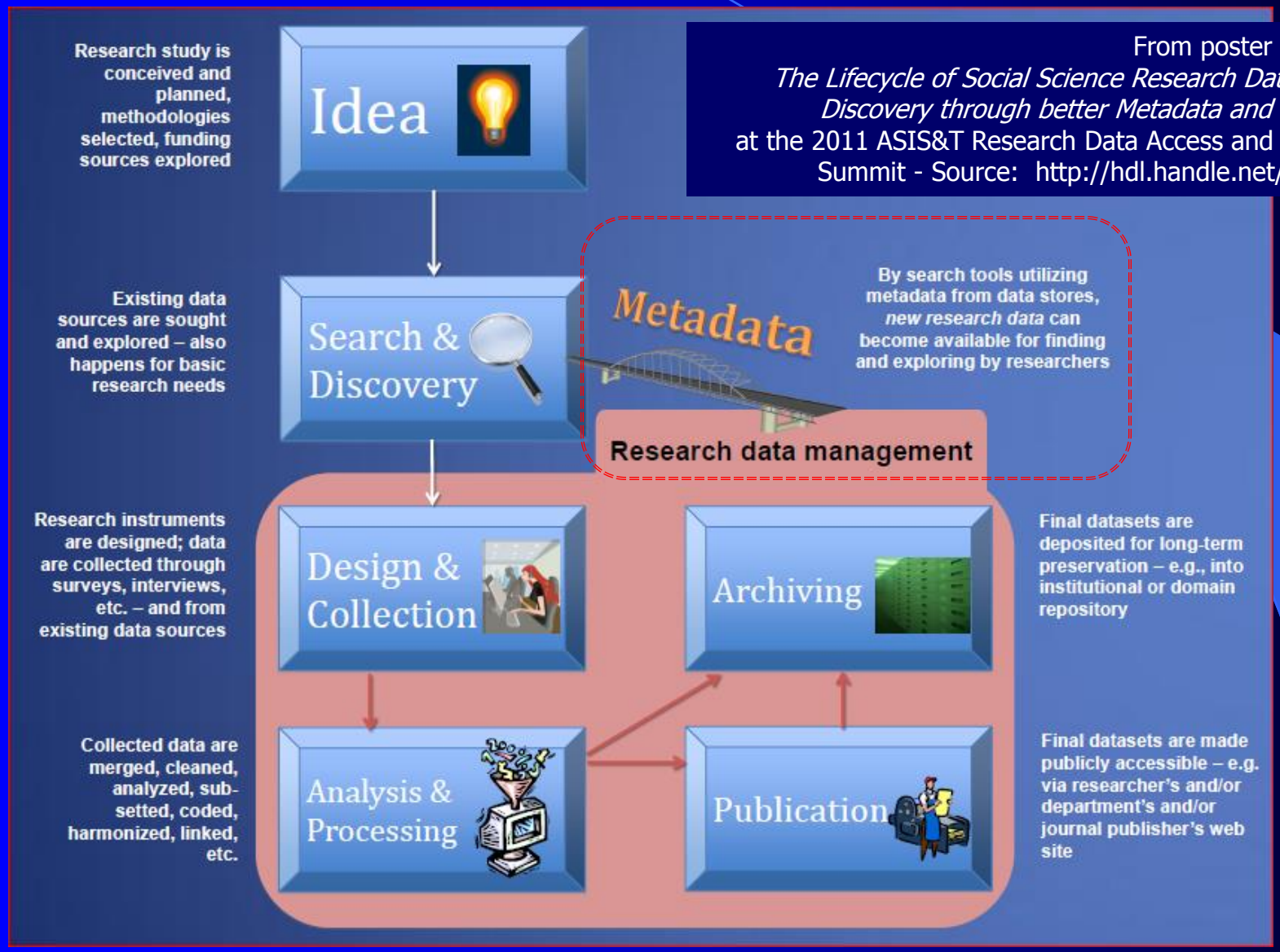
Figure 7: Responses to the question: "Given the NSF expectation to share data with other researchers, how much data would you intend to share?"



Note: this is just the amount of research data being *shared with others* = subset of what's collected and needs to be archived?

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

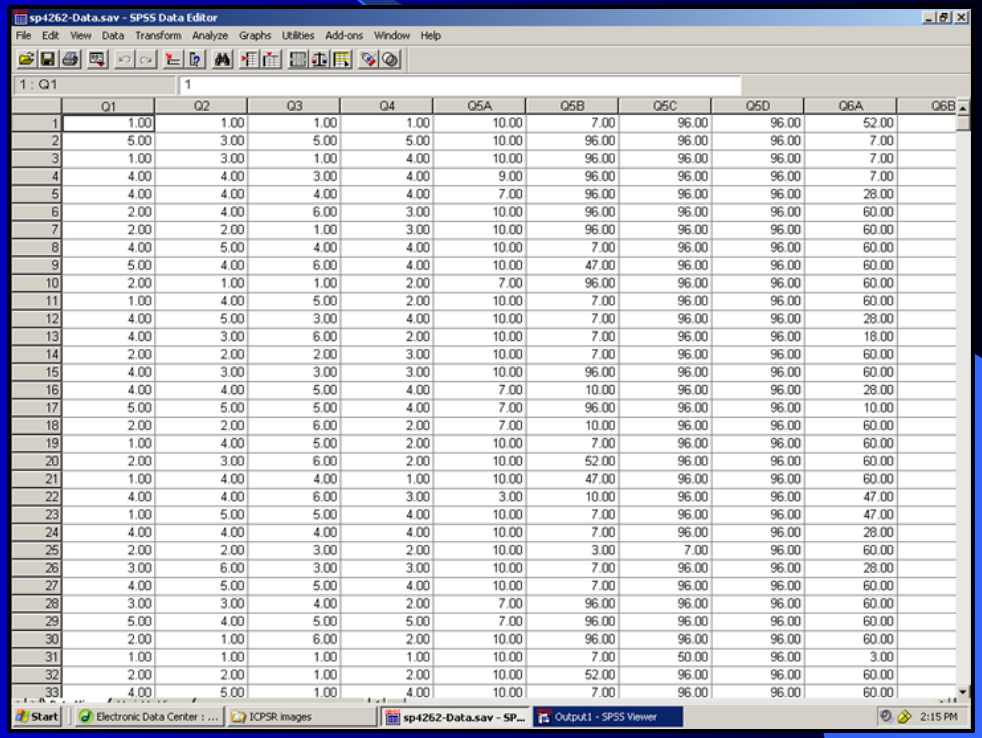
… but there's also the *metadata*



From poster presentation
The Lifecycle of Social Science Research Data: Improved Discovery through better Metadata and Search Tools
at the 2011 ASIS&T Research Data Access and Preservation Summit - Source: http://hdl.handle.net/1813/22464

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 full-text 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.



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