



Data Management

Data management – what is it?

- Data management is a broad catch-all term used by different people in different contexts. It can be used to describe a variety of activities such as:
 - Data storage,
 - Data curation,
 - Data preservation,
 - Database design,
 - Data modeling and more
- Sometimes it can be used to refer to data management policy and sometimes to the practice of data management.

Data management for the researcher

- All those activities which a researcher can undertake
 - to organise and manage their data
 - to facilitate their own research, and
 - to provide a foundation for the longer-term sustainability of the data

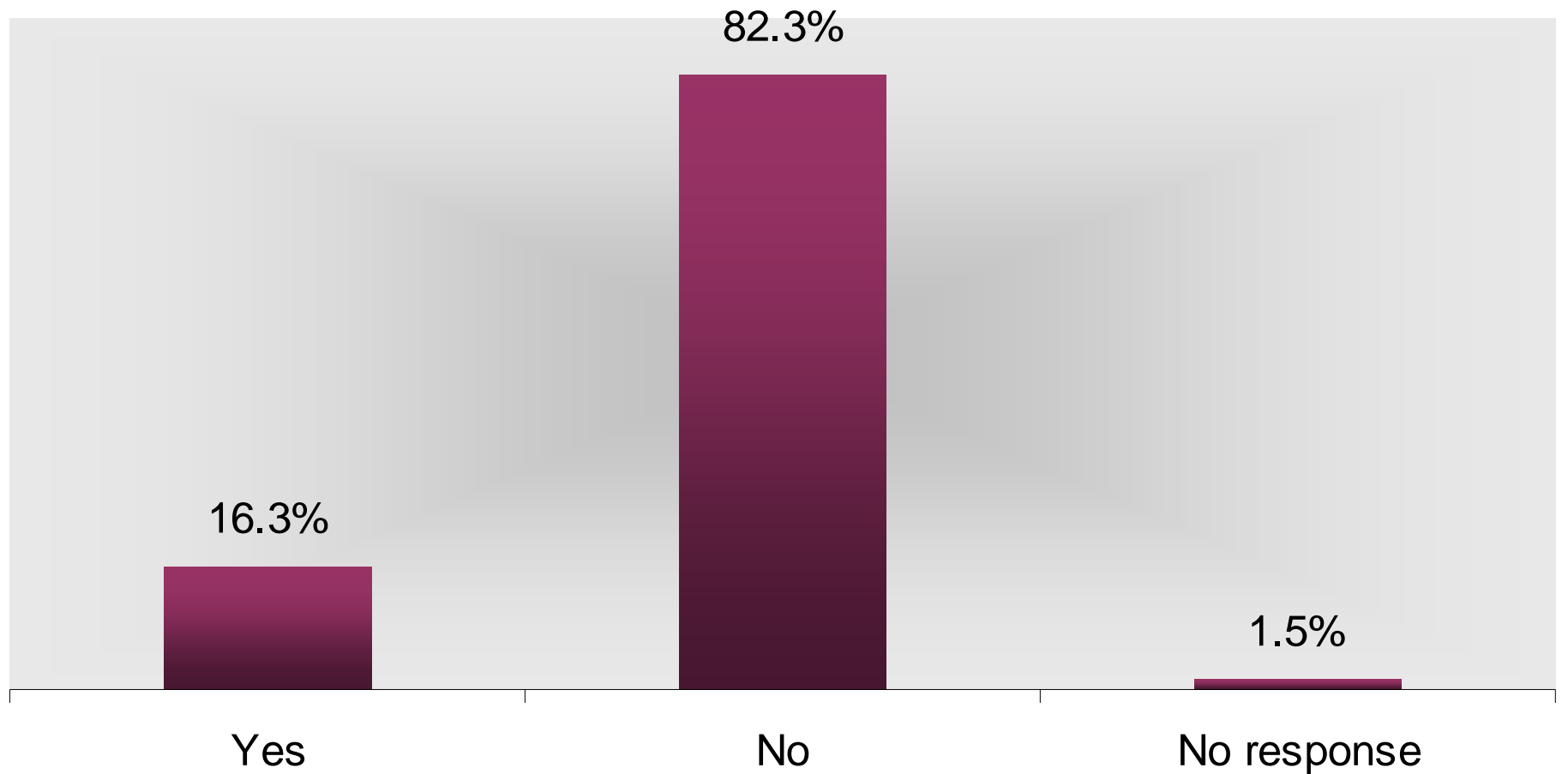
Archiving – different views

- The archivist's view
 - not all materials are kept in perpetuity, but may be subject to disposal according to defined protocols and schedules
- The sociologist's view
 - the collection, storage, and preservation of data
- The computer science view
 - A repository for information that the user wishes to retain, but without requiring immediate access

Why is data management important

- Code
 - ‘to justify the outcomes of the research and to defend them if they are challenged’
 - ‘good stewardship of public resources...’
- Data sharing and re-use
 - Data cannot be shared or re-used unless it has been well managed and described from the outset

Do you have a research data management plan?



Data management planning

- See the resource page on the ANDS website

<http://ands.org.au/resource/data-management-planning.html>



Managing Data from the Institutional Perspective

A useful model: OAIS

- Open Archival Information System
- Common vocabulary within institutions
- And across institutions
- Don't need to conform 100%
 - Example: ASSDA—Australian Social Science Data Archive

Institutional value of data

- Institutions measure metrics:
 - cost, ROI, depreciation
 - hardware, software, assets
- Data is a managed asset too
- Data has costs and can increase in value
- Can plan for and measure metrics
- May need a data advocate

Data privacy

- Foundational issue for data management
- May be ethical and legislative requirements
- Rule of thumb:
 - more privacy means more cost
 - more privacy means more development
- Example: ASSDA

Data licensing

- Data shared can become more valuable
- For open datasets consider Creative Commons
- Localized Australian and institutional versions
- Example: APSA
 - Australasian Pollen and Spore Atlas

Data volume

- *Bigger is different*
- 1GB, 1TB, 1PB qualitatively different
- Implications for:
 - Hardware, software, workflows
- Rule of thumb:
 - x 10 for text, XML, image, audio, video
- Example: MACHO
 - High profile astronomy project and dataset

Data quality

- Institutions implement quality
 - Software, services, research, processes
- Data can have measureable quality targets
- Common success factor:
 - people who care
- Support through automation
- Example: PARADISEC
 - <http://paradisec.org.au>

YDHTDIAY

- You Don't Have To Do It All Yourself (!)
- Data management brings many workflows
 - digitization, conversion, dissemination
- Consider external services:
 - many cloud based
 - elephant in the room: Google
- Example: HCCDA
 - Historical Census and Colonial Data Archive

Automate, automate, automate

- Data management bring many workflows
- Beware manual or semi-manual workflows
- Supports the people-who-care
- Supports data quality
- Close the loop by upgrading automation
- Example: PARADISEC

Data collaborators

- Institutions have known data collaborators:
 - researchers, government, other institutions
- But also consider automated collaborators:
 - web services
 - search engines
 - data federations
- May mean creating new exchange formats
- Example: APSA

It's all data

- Successful data projects increase demand
- Demand for more and different datasets
- Rule of thumb:
 - more data variety means more cost
 - more data variety means more development
- Consider:
 - documents, audio, video, instrument readings
 - GIS, diagrams, log-files, financial data, programs
 - bio-markers, asset registers, images, XML, ...

Names and ID's

- Tedious-but-essential management practice
- Overlaps with institutional metadata policy
- Having any scheme better than none
- May be mandated for you (e.g. DOI)
- ANDS source of national expertise
- Example: HCCDA

Migration strategies

- Technology evolves:
 - hardware, databases, frameworks, networks
- Need an exit strategy for each migration
- Beware proprietary formats and tools
- Example: HCCDA

Digitization

- Institutions often have legacy “analog” archive
 - documents, maps, charts, photos, film, tape
- Answer: digitization
 - can be simple to implement
 - can also be highly complex
- Plan for tomorrow’s technology
 - don’t sacrifice quality for a \$100 hard drive
- Example: PARADISEC