Data Lifecycle Management

Hannah Norton and Rolando Garcia-Milian

UF Health Science Center Libraries

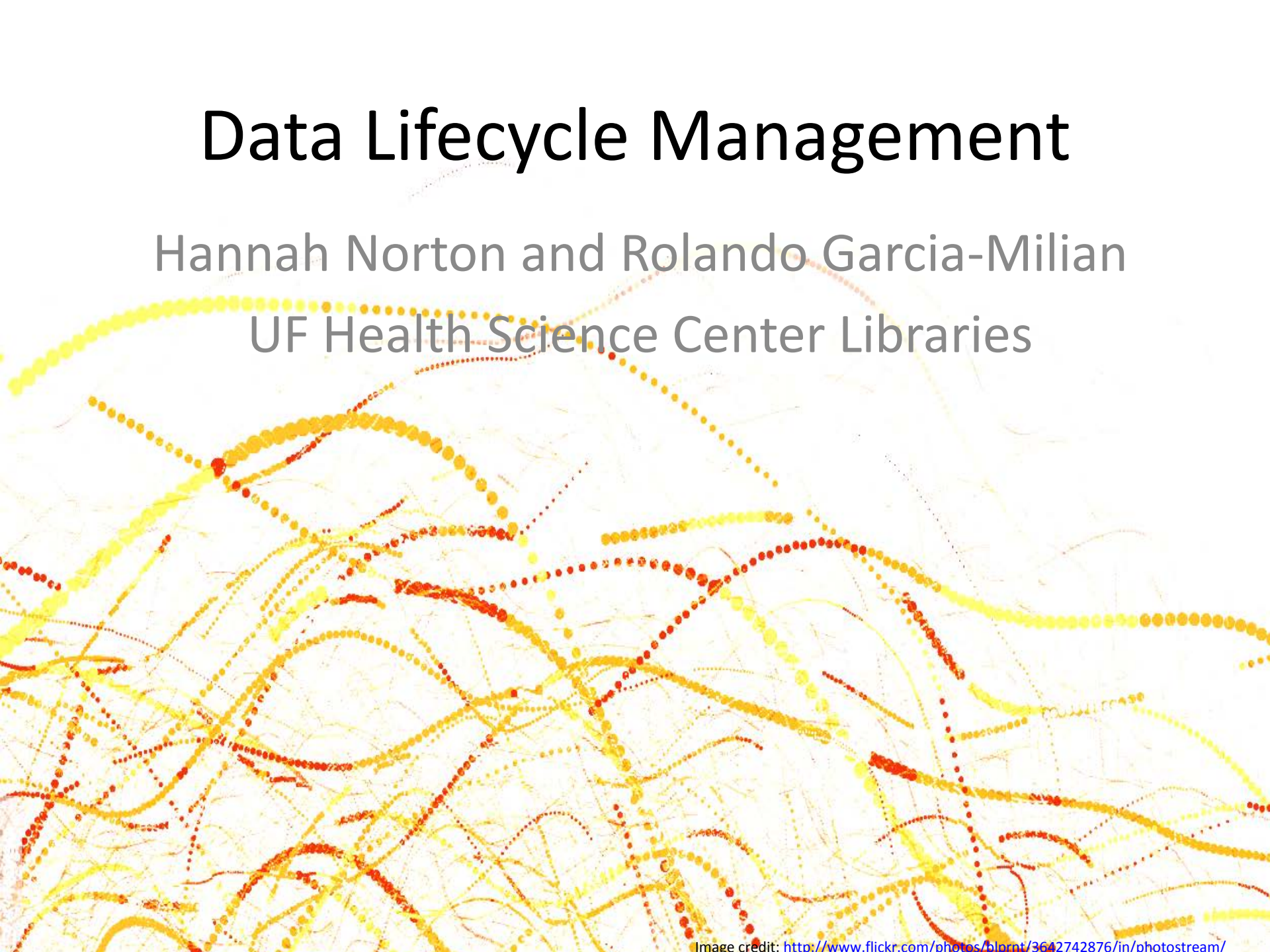
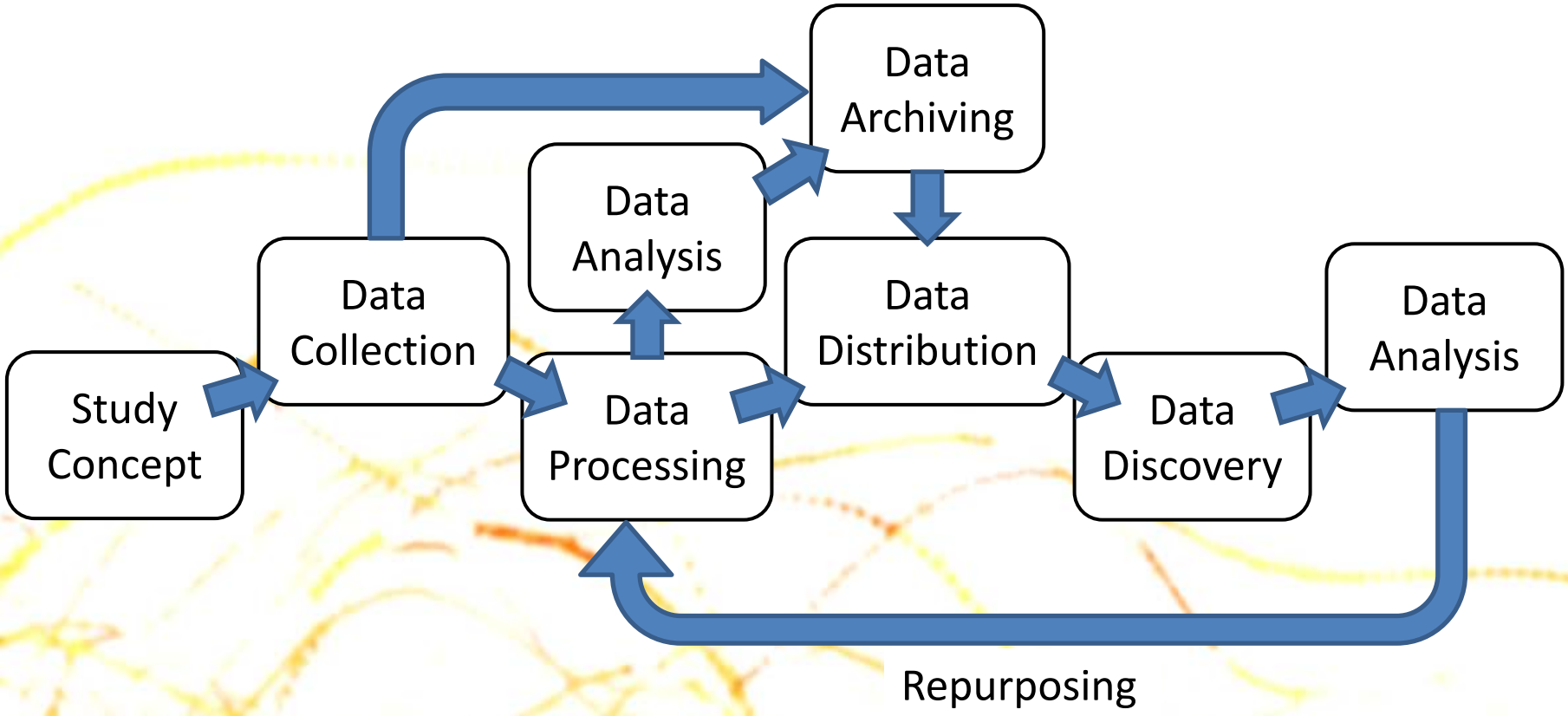


Image credit: http://www.flickr.com/photos/blprnt/3642742876/in/photostream/

Agenda

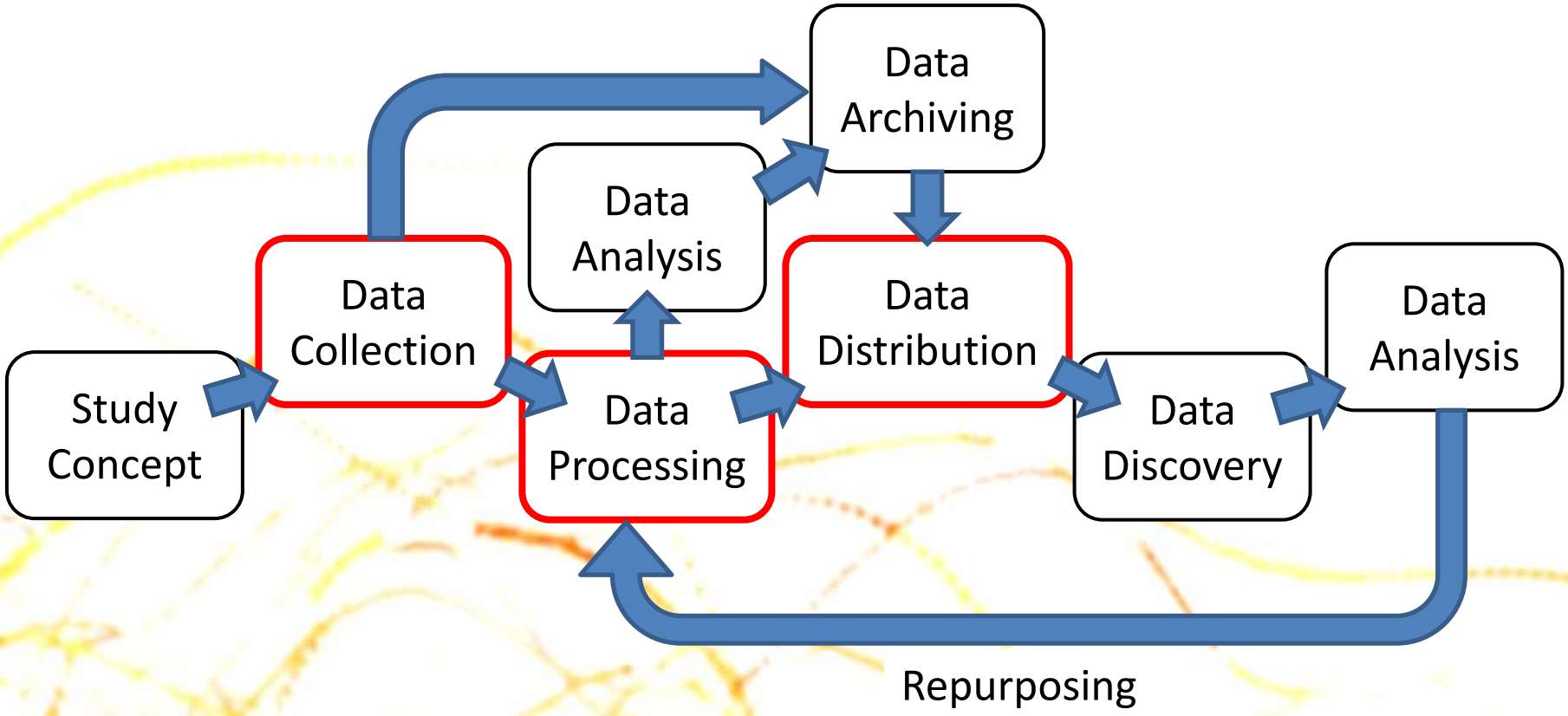
- The data lifecycle and data types
- Metadata and labeling your data
- Storage and preservation
- Data management planning
- Additional resources

Data Lifecycle*



* Based on Data Documentation Initiative (DDI) version 3.0 Combined Life Cycle Model

Data Lifecycle*

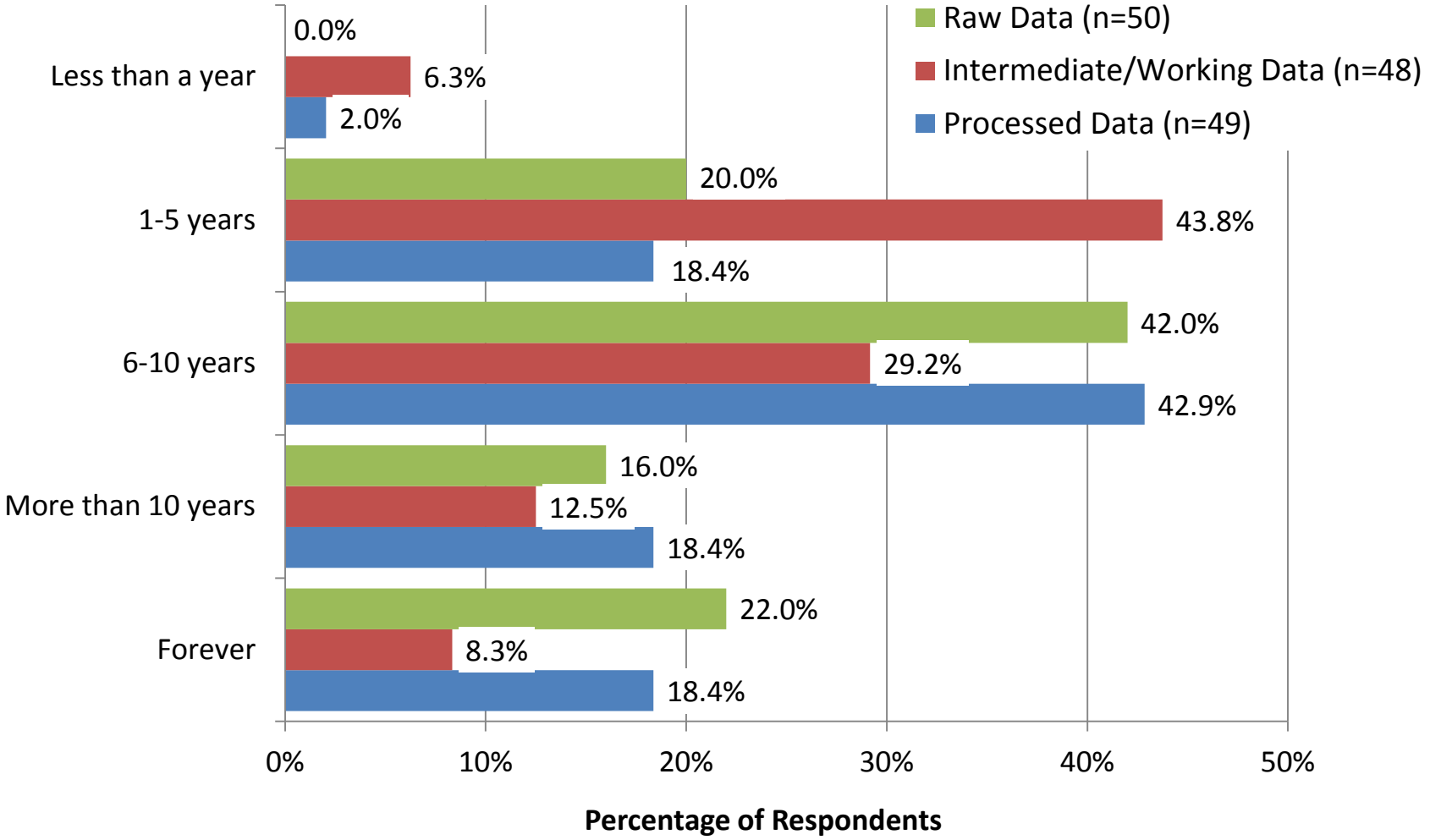


* Based on Data Documentation Initiative (DDI) version 3.0 Combined Life Cycle Model

Data generated throughout the lifecycle has different needs

- **Raw data** - some must be kept forever, others can be discarded after the project is complete
- **Intermediate data** for analyzing and processing - can be often be discarded at the end of the computation, but computational methods should be for reproducibility
- **Final data** - should be made available indefinitely to the community

How long do you need your data stored?



Data types and reproducibility

- Reproducibility is a key parameter in determining the need for long-term preservation of data:
  - **Stable (S)**: Derives from simulations, reductions, measurements
  - **Ephemeral (E)**: Cannot be reproduced or reconstructed as it is time-sensitive
  - **Costly (C)**: Stable but costly to regenerate

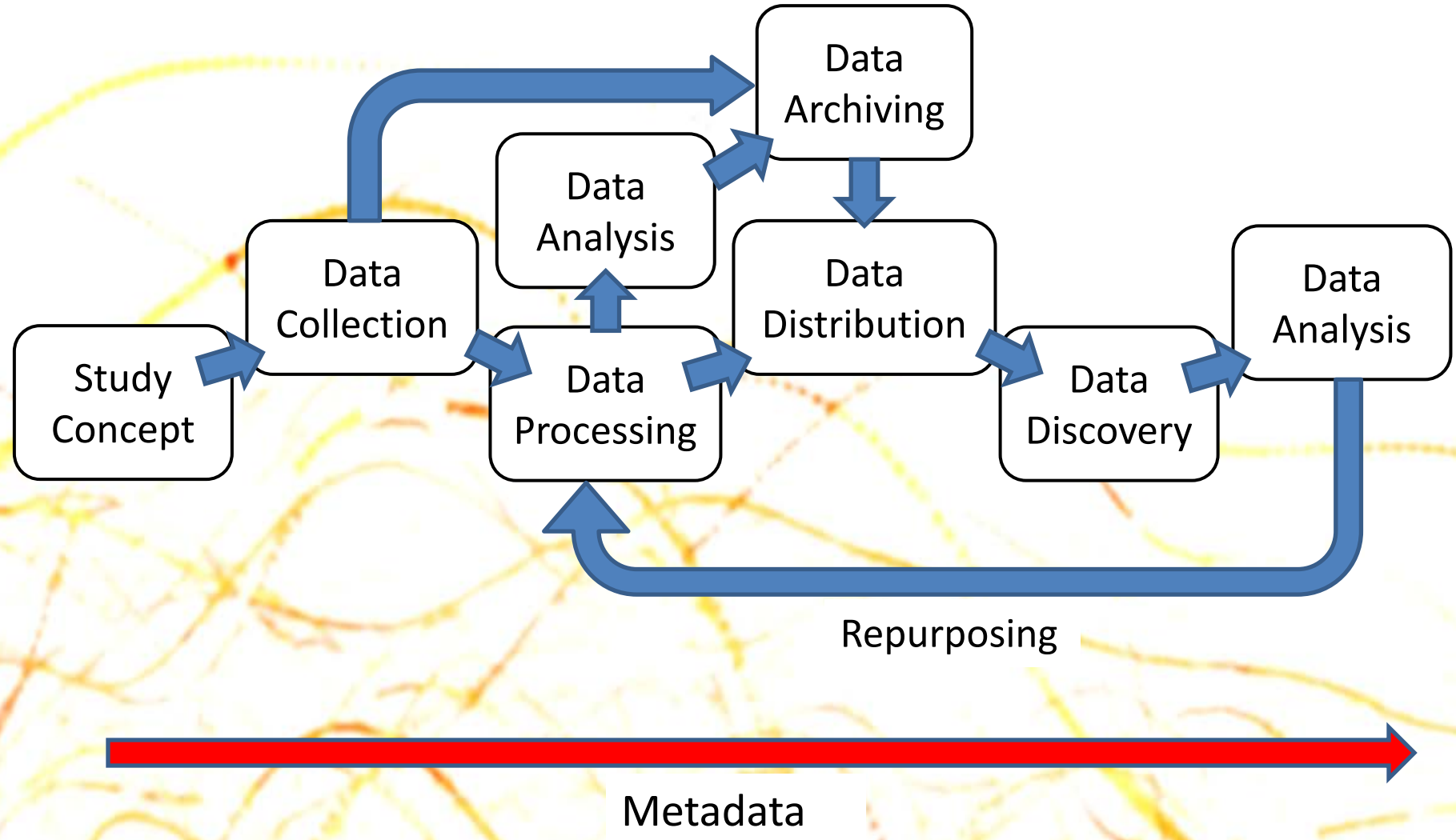
Modified from: http://tinyurl.com/724hezs

Data types and reproducibility

- Experimental data (S, E, C): from labs and equipment
- Observational data (E): captured in real time
- Derived data (S, E, C): after data mining and statistical processing
- Simulation data (S, E, C): data generated from modeling processes
- Software (S, E, C)

Modified from: http://tinyurl.com/724hezs

Metadata/annotation must be added throughout the lifecycle

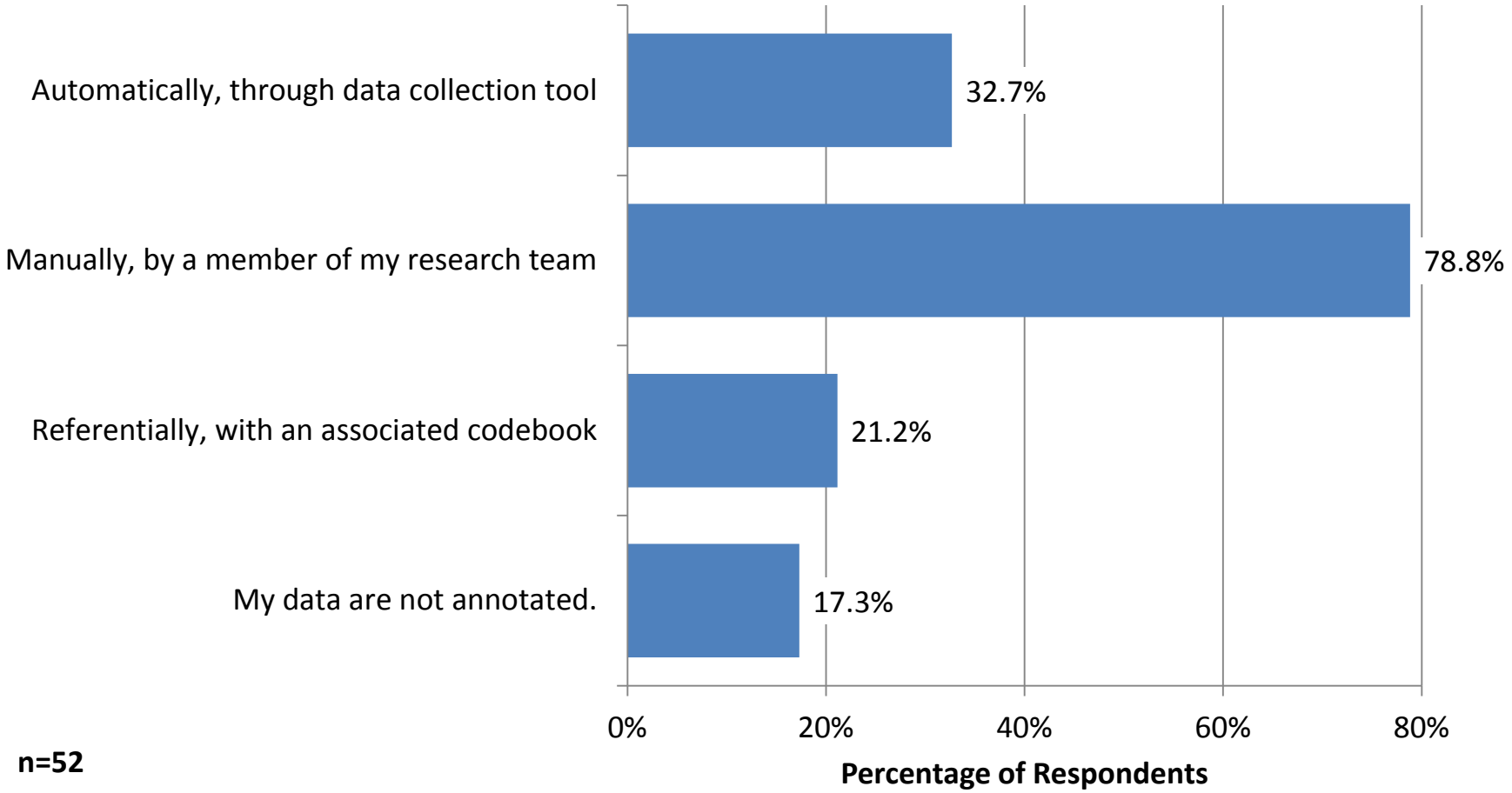


What exactly is metadata again?

- Descriptive information that helps you and others understand your data

- "Data about data" that acts as a surrogate for your data when you or others are trying to:
  - Find the data later
  - Know what the data is later
  - Share the data later

How are your data labeled or annotated?



n=52

Metadata across the disciplines

Basic information to keep:

- Descriptive
  - What is it about?
  - Title, time, author, keywords
  - Relations to other data objects
- Administrative
  - Ownership and use permissions
- Provenance
  - Where does it come from?
  - History of changes to the data, versions

More specific information varies by discipline

Standards

- Where possible, use standard data formats and metadata formats.
  - Saves you time, saves the data users time.
  - The tricky part is finding the right standard.
- Ontologies and controlled vocabularies can also help standardize the contents of your metadata and make it easier to understand.

Sample metadata standards

- Dublin Core
- Darwin Core
- METS (Metadata Encoding and Transmission Standards)
- FGDC (Federal Geographic Data Committee)
- DDI (Data Documentation Initiative)
- ABCD (Access to Biological Collections Data)
- AVMS (Astronomy Visualization Metadata Standard)
- CSDGM (Content Standard for Digital Geospatial Metadata)

<META NAME="DC.Title" CONTENT=" Fayetteville Street Business District, Raleigh, N.C., circa 1947">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Photography">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Streets">
<META NAME="DC.Subject" SCHEME="lctgm" CONTENT="Commercial streets">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Motion picture theaters">
<META NAME="DC.Description" CONTENT="A circa 1947 street scene depicting Fayetteville Street in Raleigh, North Carolina, and ending at the State Capitol building in the distance. The street is quite busy with cars, and demonstrates the parking along the street at the time of the photograph. The photograph shows the landscape and tall buildings that lined Fayetteville Street at that time.">
<META NAME="DC.Publisher" SCHEME="lcnaf" CONTENT="North Carolina State Archives. Audio Visual and Iconographic Materials.">
<META NAME="DC.Date.Created" SCHEME="iso8601" CONTENT="1945/1949">
<META NAME="DC.Type" SCHEME="DCMIType" CONTENT="Image">
<META NAME="DC.Format.Medium" SCHEME="gmgpc" CONTENT="Aerial views">
<META NAME="DC.Format.Extent" CONTENT="1 item">
<META NAME="DC.Coverage.Spatial" SCHEME="lcsh" CONTENT="Raleigh (N.C.)">
<META NAME="DC.Rights" CONTENT="The North Carolina State Archives owns copyright to this document.">
<META NAME="DC.Source" CONTENT="Carolina Power & Light Company">

From: http://www.ncecho.org/dig/ncdc2007.shtml#6

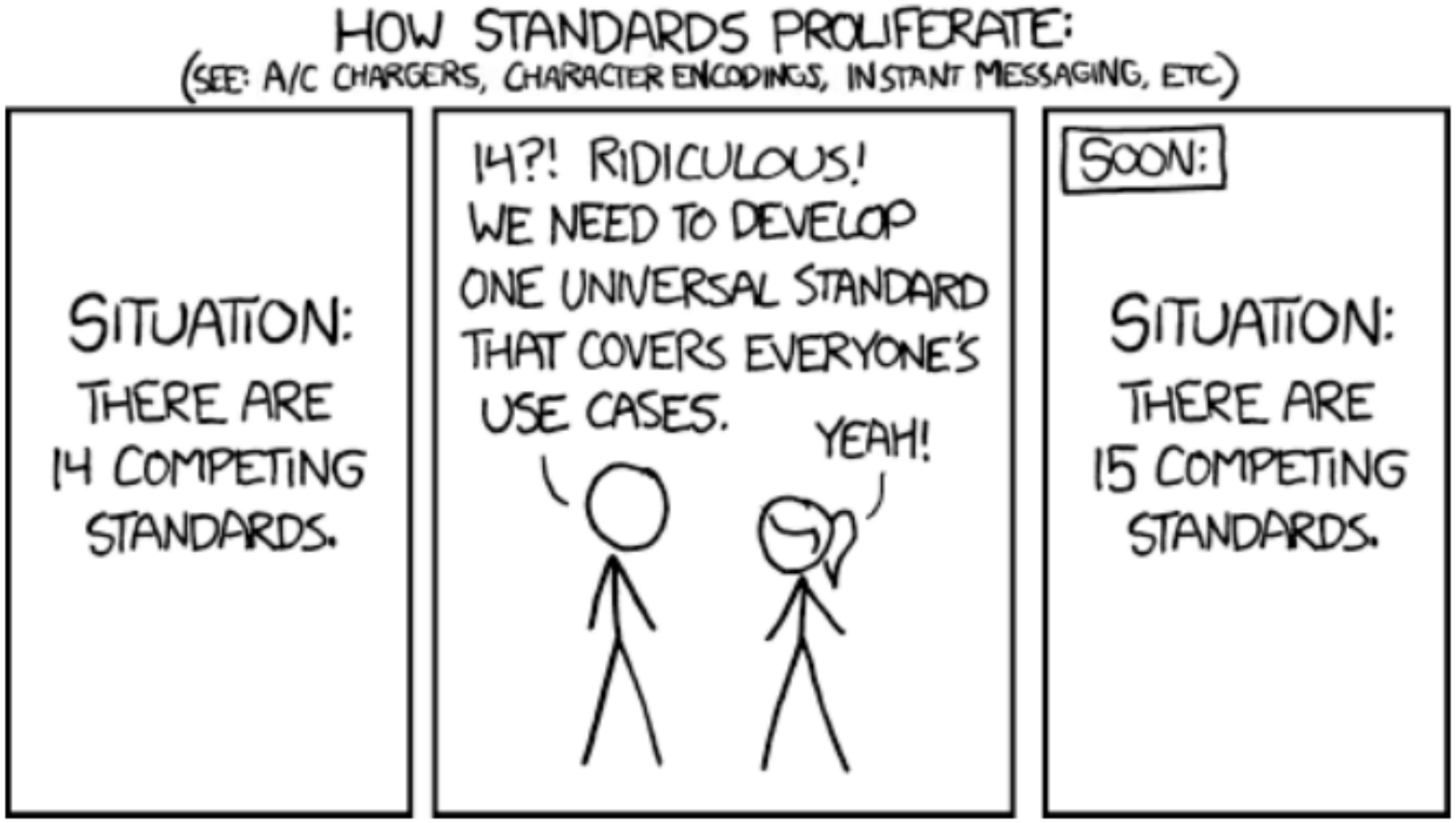
<META NAME="DC.Title" CONTENT=" Fayetteville Street Business District, Raleigh, N.C., circa 1947">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Photography">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Streets">
<META NAME="DC.Subject" SCHEME="lctgm" CONTENT="Commercial streets">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Motion picture theaters">
<META NAME="DC.Description" CONTENT="A circa 1947 street scene depicting Fayetteville Street in Raleigh, North Carolina, and ending at the State Capitol building in the distance. The street is quite busy with cars, and demonstrates the parking along the street at the time of the photograph. The photograph shows the landscape and tall buildings that lined Fayetteville Street at that time.">
<META NAME="DC.Publisher" SCHEME="lcnaf" CONTENT="North Carolina State Archives. Audio Visual and Iconographic Materials.">
<META NAME="DC.Date.Created" SCHEME="iso8601" CONTENT="1945/1949">
<META NAME="DC.Type" SCHEME="DCMIType" CONTENT="Image">
<META NAME="DC.Format.Medium" SCHEME="gmgpc" CONTENT="Aerial views">
<META NAME="DC.Format.Extent" CONTENT="1 item">
<META NAME="DC.Coverage.Spatial" SCHEME="lcsh" CONTENT="Raleigh (N.C.)">
<META NAME="DC.Rights" CONTENT="The North Carolina State Archives owns copyright to this document.">
<META NAME="DC.Source" CONTENT="Carolina Power & Light Company">

<META NAME="DC.Title" CONTENT=" Fayetteville Street Business District, Raleigh, N.C., circa 1947">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Photography">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Streets">
<META NAME="DC.Subject" SCHEME="lctgm" CONTENT="Commercial streets">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Motion picture theaters">
<META NAME="DC.Description" CONTENT="A circa 1947 street scene depicting Fayetteville Street in Raleigh, North Carolina, and ending at the State Capitol building in the distance. The street is quite busy with cars, and demonstrates the parking along the street at the time of the photograph. The photograph shows the landscape and tall buildings that lined Fayetteville Street at that time.">
<META NAME="DC.Publisher" SCHEME="lcnaf" CONTENT="North Carolina State Archives. Audio Visual and Iconographic Materials.">
<META NAME="DC.Date.Created" SCHEME="iso8601" CONTENT="1945/1949">
<META NAME="DC.Type" SCHEME="DCMIType" CONTENT="Image">
<META NAME="DC.Format.Medium" SCHEME="gmgpc" CONTENT="Aerial views">
<META NAME="DC.Format.Extent" CONTENT="1 item">
<META NAME="DC.Coverage.Spatial" SCHEME="lcsh" CONTENT="Raleigh (N.C.)">
<META NAME="DC.Rights" CONTENT="The North Carolina State Archives owns copyright to this document.">
<META NAME="DC.Source" CONTENT="Carolina Power & Light Company">

Sample ontologies/ controlled vocabularies

- LCSH (Library of Congress Subject Headings)
- MeSH (Medical Subject Headings)
- GeneOntology
- Plant Ontology
- International Standard Classification of Education
- NASA Thesaurus
- Multilingual Thesaurus of the Geosciences
- SNOMED Clinical Terms

<META NAME="DC.Title" CONTENT=" Fayetteville Street Business District, Raleigh, N.C., circa 1947">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Photography">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Streets">
<META NAME="DC.Subject" SCHEME="lctgm" CONTENT="Commercial streets">
<META NAME="DC.Subject" SCHEME="lcsh" CONTENT="Motion picture theaters">
<META NAME="DC.Description" CONTENT="A circa 1947 street scene depicting Fayetteville Street in Raleigh, North Carolina, and ending at the State Capitol building in the distance. The street is quite busy with cars, and demonstrates the parking along the street at the time of the photograph. The photograph shows the landscape and tall buildings that lined Fayetteville Street at that time.">
<META NAME="DC.Publisher" SCHEME="lcnaf" CONTENT="North Carolina State Archives. Audio Visual and Iconographic Materials.">
<META NAME="DC.Date.Created" SCHEME="iso8601" CONTENT="1945/1949">
<META NAME="DC.Type" SCHEME="DCMIType" CONTENT="Image">
<META NAME="DC.Format.Medium" SCHEME="gmgpc" CONTENT="Aerial views">
<META NAME="DC.Format.Extent" CONTENT="1 item">
<META NAME="DC.Coverage.Spatial" SCHEME="lcsh" CONTENT="Raleigh (N.C.)">
<META NAME="DC.Rights" CONTENT="The North Carolina State Archives owns copyright to this document.">
<META NAME="DC.Source" CONTENT="Carolina Power & Light Company">

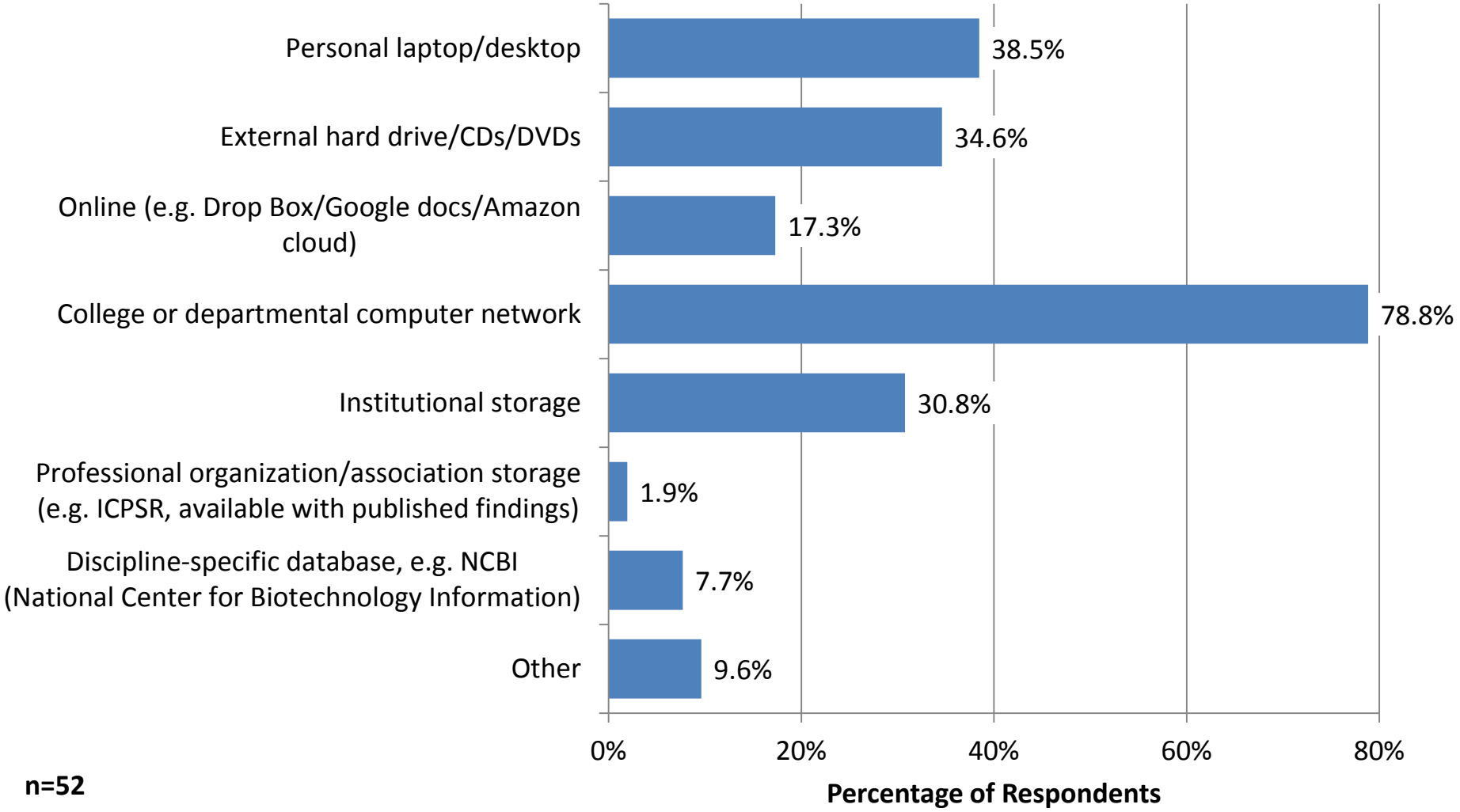


Source: http://xkcd.com/927/

Finding a home for your data

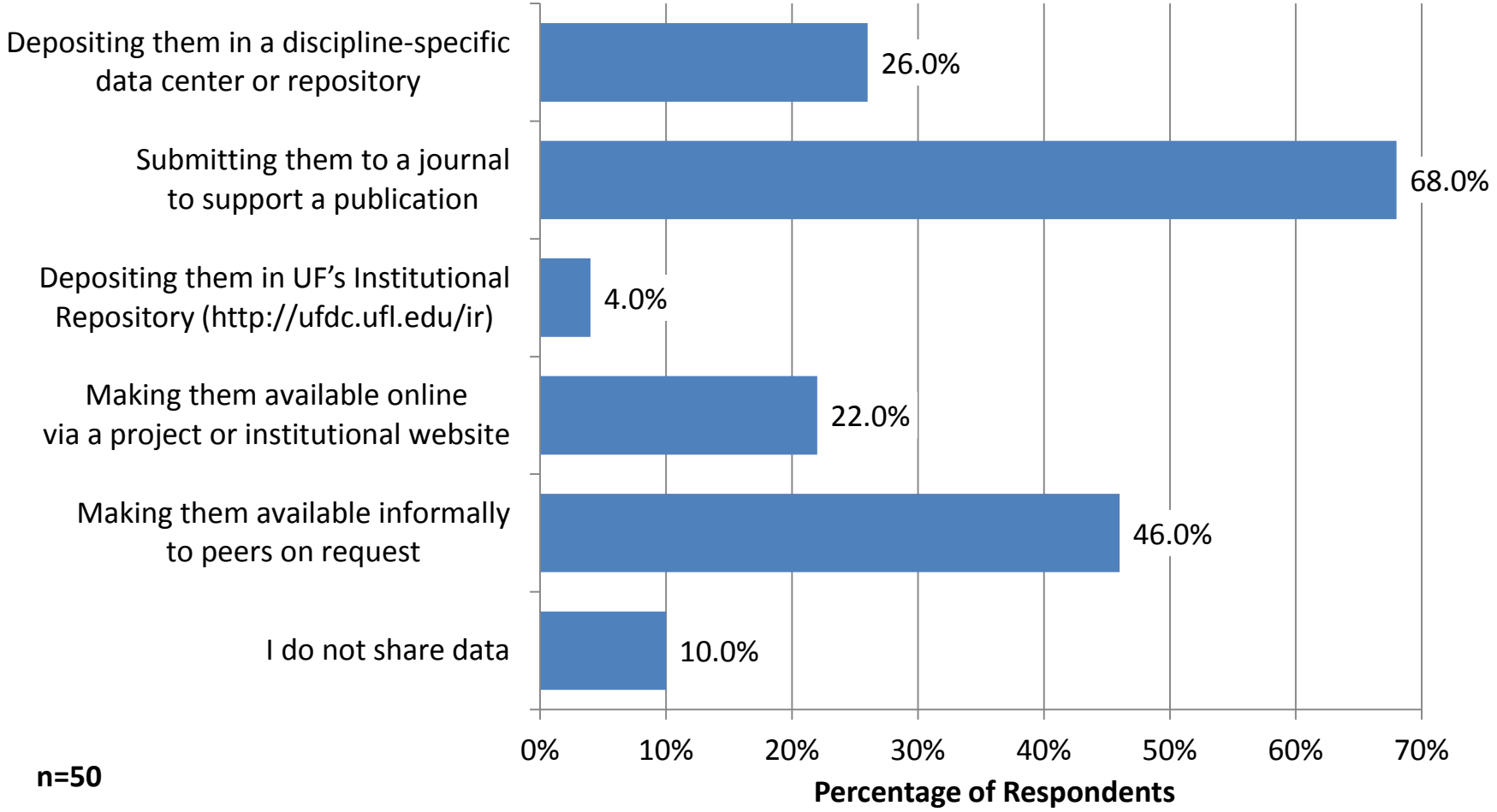
- Data storage, both short-term and long-term, can take place in 3 types of places:
  - Locally, within the lab or research environment
  - Within the institution
  - Within a national/discipline-based repository

How do you store your data?



n=52

How are you sharing or planning to share your data?



n=50

Repositories

Advantages of an institutional repository

- Linked to your institution – intellectual capital of the institution in one place
- You can put all your datasets together
- Some guarantee of support from the university
- Some domain repositories may “go out of business” once their funding ends

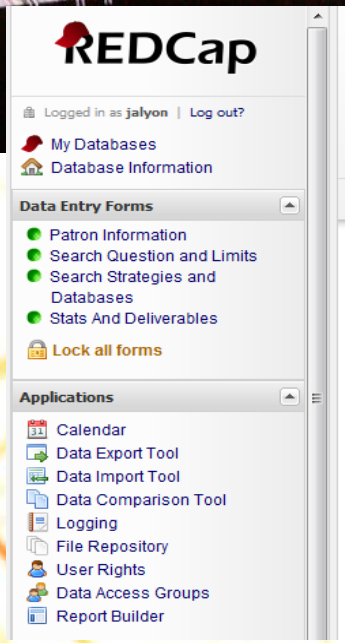
Advantages of a domain repository

- Your data will stored with similar datasets
- Researchers will find your data easily
- The repository will understand what your data needs in terms of storage, archiving and preservation
- Computational tools may be developed to crunch a critical mass of data of a certain kind

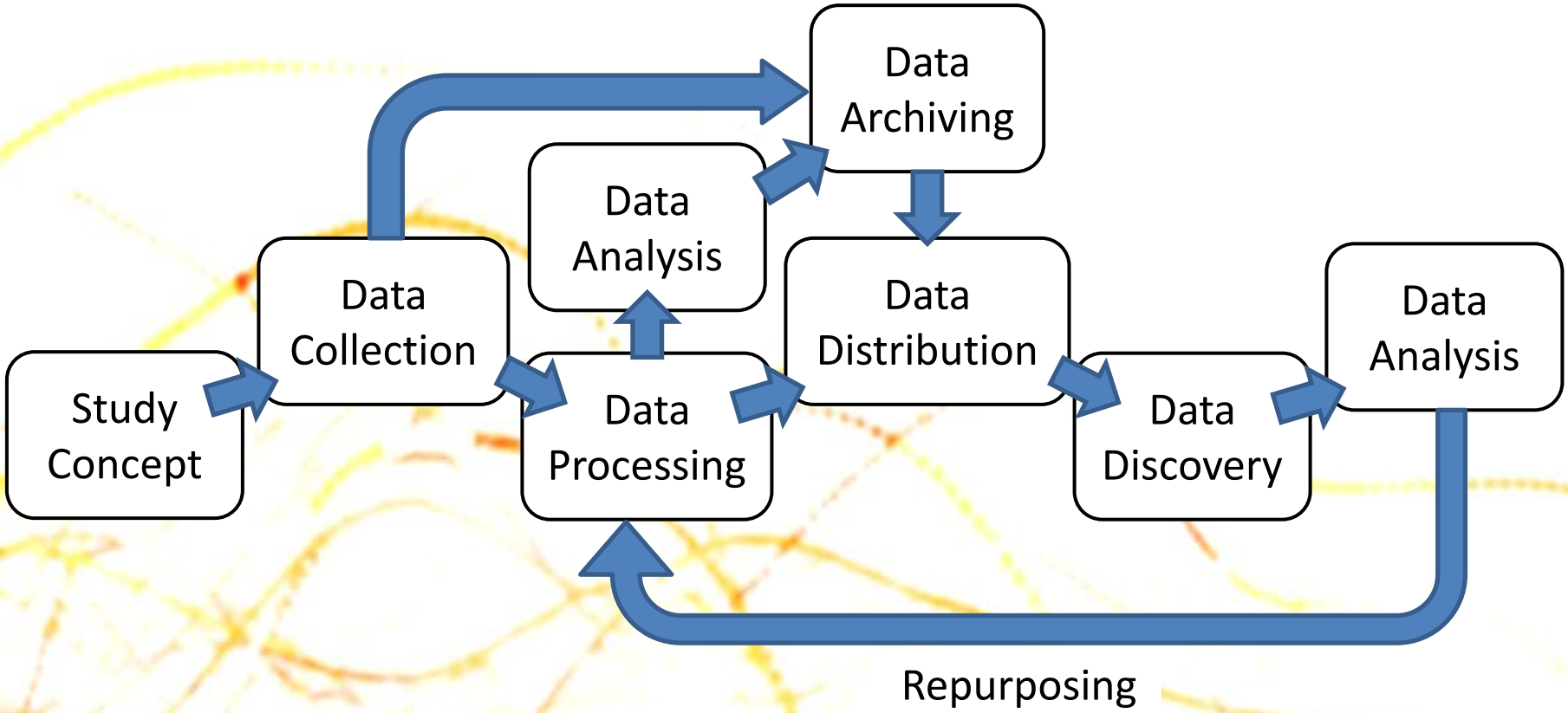
Adapted from: http://libraries.mit.edu/guides/subjects/data-management/Managing%20Research%20Data%20101.pdf

What about non-digital data?

Consider migrating to digital…



Data Management Plans describe the whole data lifecycle.



What is a data management plan (DMP)?

- A clear description of how you plan to address data management issues in your research.
- A way to communicate your data management efforts to members of your team and others (especially funders).

A data management plan gives a concise description of the who, what, where, and when of your data throughout its life cycle.

Why do you need a DMP?

For all the same reasons you should consider following data management best practices…

- To ensure that your valuable data resources will be accessible in the future to members of your team and the broader research community.
- To make your life easier – by planning ahead and documenting your data throughout its life cycle, you can save time and focus on your research.
- To increase the visibility of your research.
- To satisfy funders' requirements.

Components of a DMP

- Project description
- Data collection:
  - Types of data
  - Data and metadata standards to be used
- Legal and ethical issues:
  - Privacy and confidentiality
  - Intellectual property rights
- Policies for data sharing and re-use
- Data preservation (long-term)
- Who is responsible for data management

Funders' data requirements

- National Science Foundation: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp
- National Institutes of Health: http://grants.nih.gov/grants/policy/data_sharing/
- Centers for Disease Control: http://www.cdc.gov/od/foia/policies/sharing.htm
- NASA Earth Science: http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/
- Environmental Protection Agency: http://www.epa.gov/quality/informationguidelines/documents/EPA_InfoQualityGuidelines.pdf

DMP Templates and Tools

Templates can give you a place to start, as long as you customize them for your project.

- HPC Center links: http://www.hpc.ufl.edu/proposals/
- DMPTool
  Guidance and Resources for your Data Management Plan
  https://dmp.cdlib.org/

Data Management and Access Plan

I. Products of the Research

(a) Observational data may include images, spectra and/or photometric time series of astronomical objects, calibration data (e.g., flat fields, observations of reference stars) and associated metadata necessary for the proposed research.
(b) Results of data/statistical analyses (e.g., parameter estimates and associated uncertainties) are derived from observations/simulations and summarize the results relevant to the proposed research.
(c) Simulated data used for publications is compared to observations to aid in the interpretation of results.
(d) Software used for publications will be developed to reduce observations, model observations of astronomical objects and/or perform statistical analyses.
(e) Curriculum materials may contain web pages, handouts, PowerPoint/KeyNote /OpenOffice slides, multimedia presentations and assessment materials associated with the proposed broader impact activities.

Preliminary data, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues and physical samples are not included in this plan as set forth by the US Office of Management and Budget.

II. Data Formats

(a) We plan to store both raw and final reduced observational data in standard FITS files, with standard metadata (e.g., time, position, instrument settings) in FITS headers and supplemental metadata in plain ASCII (e.g., observing logs with observing conditions and other notes).
(b) Results of statistical analyses are typically recorded in standard ASCII based formats (e.g., plain ASCII, LaTeX table). Large results (e.g., posterior parameter distributions) may be compressed using standard open-source compression software (e.g., gzip) and/or stored in a binary format, provided that software is provided to extract binary data into ASCII form.
(c) Simulated data will be stored in either plain ASCII formats or in a binary format, provided that software is provided to extract data into ASCII form.

Visualizations of simulated data may be stored in standard graphics formats (e.g., eps, jpg, mov, .mpg, and/or .wmv). The Institutional Repository at UF (IR@UF) "will migrate items to new formats as necessary."
(d) Software source code and associated documentation will be stored in plain ASCII files and may be packaged using standard open-source tools (e.g., tar, gzip). A version control system (i.e., cvs, git, or subversion) will be used to support collaboration, version control and recovery of previous versions. Documentation will be embedded in source code, in separate ASCII files (e.g., plain ASCII, Asciidoc, html, LaTeX) and/or in formatted files (e.g., html generated via Doxygen and/or pdf generated from LaTeX source).
(e) Curriculum materials will be stored in standard formats (e.g, html, pdf, swf, ppt, odp) for which IR@UF provides at least basic level of preservation support.

III. Access to Data and Data Sharing Practices and Policies

The primary results of the proposed research, including the results of associated statistical analyses, will be disseminated primarily through publication in journals (including online-only supplements for extended tables, animations, etc.), the arXiv preprint server, conference presentations and student theses (long-term open-access via IR@UF). We will work to provide electronic data derived from this project upon request, in a timely fashion and on a nondiscriminatory basis. Depending on the size of the request, data may be provided via email, the UF Astronomy web/FTP servers, IR@UF or the cloud (e.g., dropbox.com). Data will be preserved for at least three years beyond the award period, as required by NSF guidelines. We note details specific to certain types of data below:
(a) Observatories generally provide access to raw observational data via web/ftp after the observatory proprietary period (e.g., 1 year for GTC, 18 month default for Keck).
(a-d) The raw observational data, final reduced observational data, results of statistical analyses, simulated data used for publications and software used for publications will be available for data sharing upon request. We reserve the right to maintain a propriety period for observational data we collect as part of the

proposed research, but voluntarily limit that period to the lesser of the time until publication or 12 months after the data become available to us. We will respect the restrictions imposed by data owners for any data obtained by unfunded collaborators or as part of collaborations (e.g., SDSS-III).
(e) Curriculum materials will be made available via the UF Astronomy department web site, IR@UF and/or the national Multimedia Educational Resource for Learning and Online Teaching (MERLOT) repository where they will be peer-reviewed. To increase visibility, we will request a link to these materials on the Science Information for Florida Teachers Guide to Everything website.

IV. Policies for Re-Use, Re-Distribution, and Production of Derivatives

Published data will be available in print or electronically from publishers, subject to subscription/printing charges and copyrights. We will work to provide other data with as few restrictions as possible. For example, preprints will be posted to arXiv, so as to ensure results are generally accessible without regard to journal subscriptions. Source code will be made available under the GNU General Public License (GPL). Other data provided via UF websites will include a request to cite the most relevant publication(s) and notice of any copyright restrictions (e.g., how to obtain permission to reuse figures published in ApJ).

As the proposed research does not involve the acquisition of either animal or human subjects data, we do anticipate any privacy or ethical issues associated with the data. We do not anticipate that there will be any significant intellectual property issues involved with the acquisition of the data. In the event that discoveries or inventions are made in direct connection with these data, access to the data will be granted upon request once appropriate invention disclosures and/or provisional patent filings are made.

V. Archiving of Data

Electronic data will be preserved using multiple on-site copies, with all servers using RAID hard drive arrays. Final versions of software source code and curriculum materials will be stored in home directories for which UF astronomy system administrators provide secure off-site backup.

On campus resources

Your data partners:

- Research Computing/HPC Center
- UF Libraries
  - Institutional Repository

Other data-related resources:

- Division of Sponsored Research
- Integrated Data Repository
- REDCap
- Office of Technology Licensing
- Information Security Office
- Intellectual Property Policy

http://guides.uflib.ufl.edu/datamanagement

Data Management at UF

Last Updated: Apr 23, 2012 | URL: http://guides.uflib.ufl.edu/datamanagement | Print Guide | Email Alerts

Home | Importance of Proper Data Management | Metadata | Formatting | Storage | Security | Copyright | Sharing

Home | Comments(0) | Print Page | Search: This Guide Search

Data Management Basics

See guide tabs for additional information on these data management topics:

- Importance of proper data management
- Metadata
- Formatting
- Storage
- Security
- Copyright
- Sharing

Comments (0)

Data-related Resources at UF

- High Performance Computing Center
- Institutional Repository (IR@UF)
- Integrated Data Repository
- George A. Smathers Libraries and Health Science Center Libraries

Comments (0)

Data Management Plans

Data management plans (DMPs) ensure that you have defined how you will collect and label, protect, store, and share your data.

Having a DMP in place at the beginning of your research process means you won't have to worry about it as much later! DMP's are good research practice, but you might also want to have one because your funder requires you to.

Basic components of a DMP:

- Data collection: Types of data; Data and metadata standards to be used
- Data protection: Privacy, confidentiality, intellectual property rights
- Data sharing and re-use
- Data preservation (long-term)

Guides to writing a DMP:

- Writing an NSF Data Management Plan (MIT)
- Data Management Plan Self-Assessment Questionnaire (Purdue)

Comments (0)

Other Data Mangement Planning Resources

Funders' Data-related Policies:

- NSF Data Sharing and Data Management Plan Requirements
- NIH Data Sharing Policy
- National Endowment for the Humanities
- Department of Energy
- NASA

Other guides:

Your Librarian

Hannah Norton

nortonh

checking status of nortonh

Type **here** and hit enter to send an offline message.

edit nickname: meeboguest15651!

get meebo

Contact Info

352-273-8412

nortonh@ufl.edu

Feel free to contact us:

- Hannah Norton

  nortonh@ufl.edu

  273-8412
- Rolando Garcia-Milian,

  rolando.milian@ufl.edu

  273-8440

UF Health Science Center Libraries

References

- Data Documentation Initiative (DDI) version 3.0 Combined Life Cycle Model: http://www.ddialliance.org/what
- Alex Ball. (2012). *Review of Data Management Lifecycle Models* (version 1.0). REDm-MED Project Document redm1rep120110ab10. Bath, UK: University of Bath. Available at http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf
- Deumens E, Taylor LNF, Schipper RA, Botero C, Garcia-Milian R, Norton HF, Tennant MR, Acord SK, Barnes CP. (2011). "Research Data Lifecycle Management: Tools and guidelines", position paper, Workshop on Research Data Lifecycle Management. Available at http://www.columbia.edu/~rb2568/rdlm/Deumens_UF_RDLM2011.pdf
- A. Collie. (2005). "NSB Long-Lived Data Collections: Enabling Research and Education in the 21st Century." Available at http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf
- Texas Advanced Computing Center. (2012). "Writing a Data Management Plan: A guide for the perplexed." Available at http://www.tacc.utexas.edu/c/document_library/get_file?uuid=e9774145-9801-4049-b324-a1b0d6e635ca&groupId=13601
- University of Wisconsin Research Data Services. Data Plan Essentials: http://researchdata.wisc.edu/make-a-plan/data-plans/

References

- MIT Libraries. Data Management and Publishing: http://libraries.mit.edu/guides/subjects/data-management/index.html
- JA Lyon, N Ferree, H Norton, MR Tennant. “Electronic capture and analysis of librarian-mediated literature searches in the health sciences”, contributed presentation, 6th Evidence Based Library and Information Practice conference, Sheffield, U.K., June 28, 2011.
- Dublin Core Metadata Initiative: http://dublincore.org/
- Darwin Core Biodiversity Information Standards: http://rs.tdwg.org/dwc/
- Metadata Encoding & Transmission Standard: http://www.loc.gov/standards/mets/
- Federal Geographic Data Committee: http://www.fgdc.gov/
- ABCD Schema – Task Group on Access to Biological Collection Data: http://www.bgbm.org/tdwg/codata/schema/
- Astronomy Visualization Metadata Standard: http://www.jodcast.net/avm/microformat.html
- Content Standard for Digital Geospatial Metadata: http://www.fgdc.gov/metadata/csdgm/