UF | George A. Smathers Libraries

Developing Data Management Education, Support, and Training, v2

Plato Smith, Data Management Librarian
University of Florida
FSU – USF Symposium on Library and Information Science
1:15 pm – 1:45 pm, MSL, L136
April 19, 2018

Setting the Context

"I am the instructor for CAP5108, Research Methods for Human-centered Computing. Dr. Smith taught a guest lecture on ***data management****. It was a very informative lecture, starting with* ***why this is an important topic****, and* ***moving on to workflow*** *and things to keep in mind. Students were engaged and asked questions on topics as varied as* ***ownership of data*** *and* ***copyright/privacy****. All of these are topics that I was very happy to see them think about. I invited Dr. Smith to repeat this module next year, with suggestions on topics that could be highlighted in the form of handouts or* ***classroom worked examples****."* – UF Computer & Information Science & Engineering (CISE) faculty, Assistant Professor, 4/16/18

Setting the Context

"Attending a data management training workshop is a great experience for a graduate student in the social sciences. This is because these workshops help to expose us to key concepts relating to the management of data that we do not cover in our graduate coursework. I am now ***thinking more critically about*** *where and* ***how I store data***, ***how I manage metadata***, *and the ideals of* ***making data public*** *for the sake of open science. These types of trainings should be highly encouraged, if not part of coursework requirements, for graduate students who conduct research*." – UF College of Education, Curriculum and Instruction (specialization in Educational Technology), graduate student, 1/30/18

Table of Contents

Some key data management concepts and terms

- **Creative Commons** – articulates/enables licenses of creative works - attribution, ownership, use/reuse (e.g. https://creativecommons.org/licenses/; https://creativecommons.org/about/program-areas/open-science/)
- **DOI** – digital object identifier enables unique identification of a digital object (e.g. https://doi.org/10.5281/zenodo.1207215)
- **IR** – institutional repository is digital content management system of scholarly outputs unique to an organization (e.g. http://ufdc.ufl.edu/ufirg)
- **ORCiD** – open researcher and contributor id is a persistent unique identifier for authors to distinguish between researchers (e.g. https://orcid.org/0000-0003-1814-0151)

Linking stakeholders, liaisons, and students

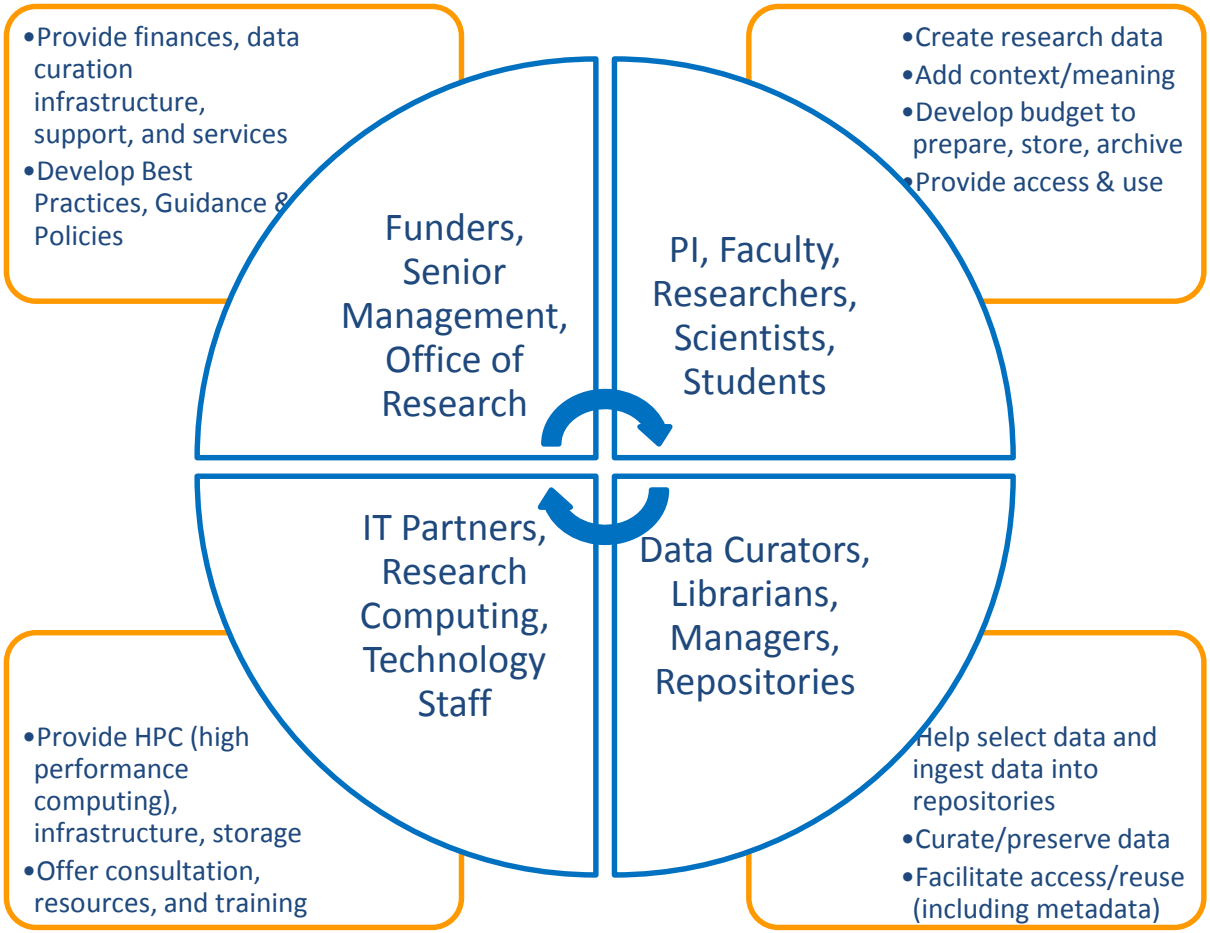

Fig. 1 Stakeholders and Data Management Responsibilities

Articulating a data management plan

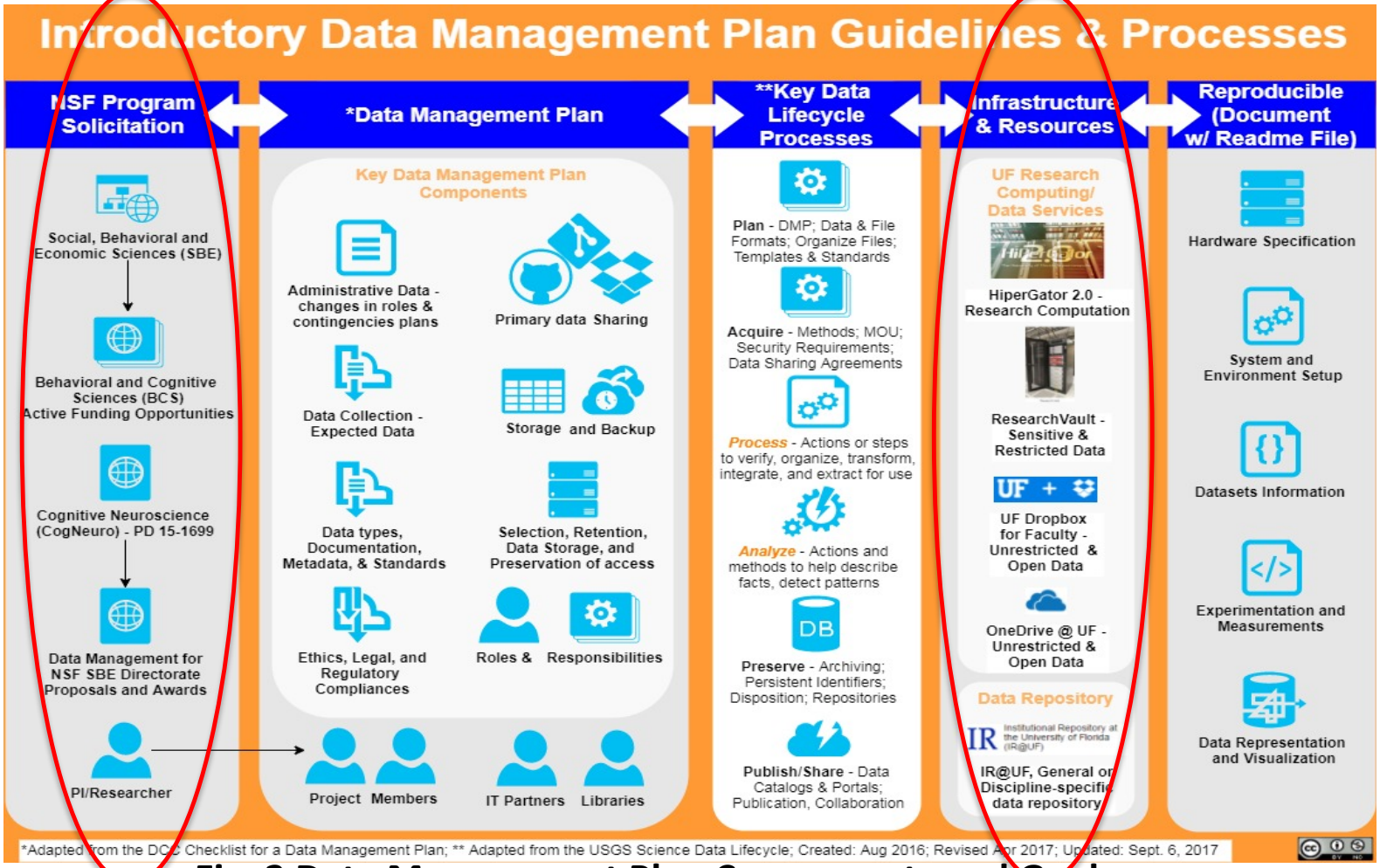

Fig. 2 Data Management Plan Components and Goals

Using the OAIS Model to Explain Concepts (CCSDS, 2002/2012)

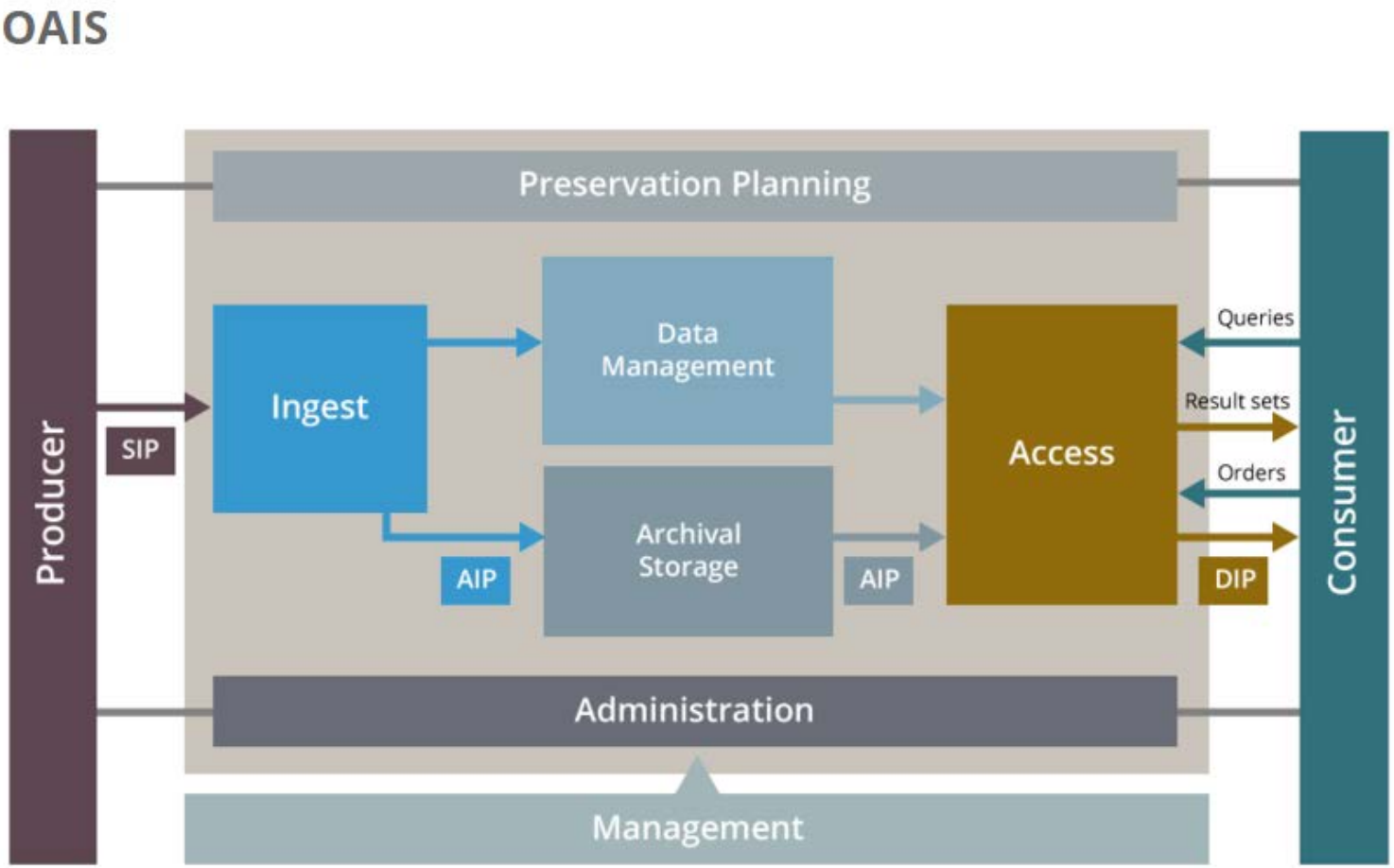

Fig. 3 Open Archival Information System (OAIS) Model (cessda, 2018)

Explaining key components of a data management plan (DCC, 2013)

Administrative Data

- ID (funder or institution)
- Funder
- Grant Reference #
- Project Name
- Project Description
- PI/Researcher
- Researcher ID (e.g. ORCID)
- Date of 1st version, last update, and related policies

Data Collection

- What data will you collect or create?
  - What type, format, and volume of data? (e.g. text, vcf, 30-50 Gigabyte per dataset)
- How will the data be collected or created?
  - What standards or methodologies will you use?
  - How will you structure and name your folders and files?

UF | George A. Smathers Libraries

Explaining key components of a data management plan (DCC, 2013)

Documentation and Metadata

- What documentation and metadata will accompany the data?
  - What information is needed for the data to be read and interpreted in the future?
  - How will you capture/create the documentation and metadata?
  - What metadata standards will you use and why?

Ethical, Legal, and Regulatory Compliances

- How will you manage any ethical issues?
  - Have you obtained consent for data preservation and sharing?
- How will you manage copyright and Intellectual Property Rights (IPR) issues?
  - Who owns the data?
  - How will the data be licensed for reuse?

UF | George A. Smathers Libraries

Explaining key components of a data management plan (DCC, 2013)

Storage and Backup

- How will the data be stored and backed up during research (e.g. FDA, Tivoli)?
  - Do you have sufficient storage or will you need to include charges for additional services?
- How will you manage access and security?
  - What are the risks to data security and how will these be managed?

Selection & Preservation

- Which data should be retained, shared, and/or preserved?
  - What data must be retained/destroyed for contractual, legal, or regulatory purposes?
- What is the long-term preservation plan for the dataset?
  - Where will you store and archive your data (e.g. which repository – re3data)?

UF | George A. Smathers Libraries

Explaining key components of a data management plan (DCC, 2013)

Data Sharing

- How will you share the data?
  - How will potential users find out about your data?
- Is there any restriction on data sharing required?
  - What action will you take to overcome or minimize restriction?

Responsibilities & Resources

- Who will be responsible for data management?
  - Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?
- What resources will you require to deliver your plan?
  - Is additional specialist expertise (or training for existing staff) required?

What are some key data lifecycle processes (USGS, 2013)?

Plan for the data

- Full-lifecycle data management articulation
- Steps to identify and secure resources and utilize infrastructure for data acquisition

Acquire the data

- Collect new data
- Convert/transform legacy data
- Share /exchange data
- Purchase data

What are some key data lifecycle processes (USGS, 2013)?

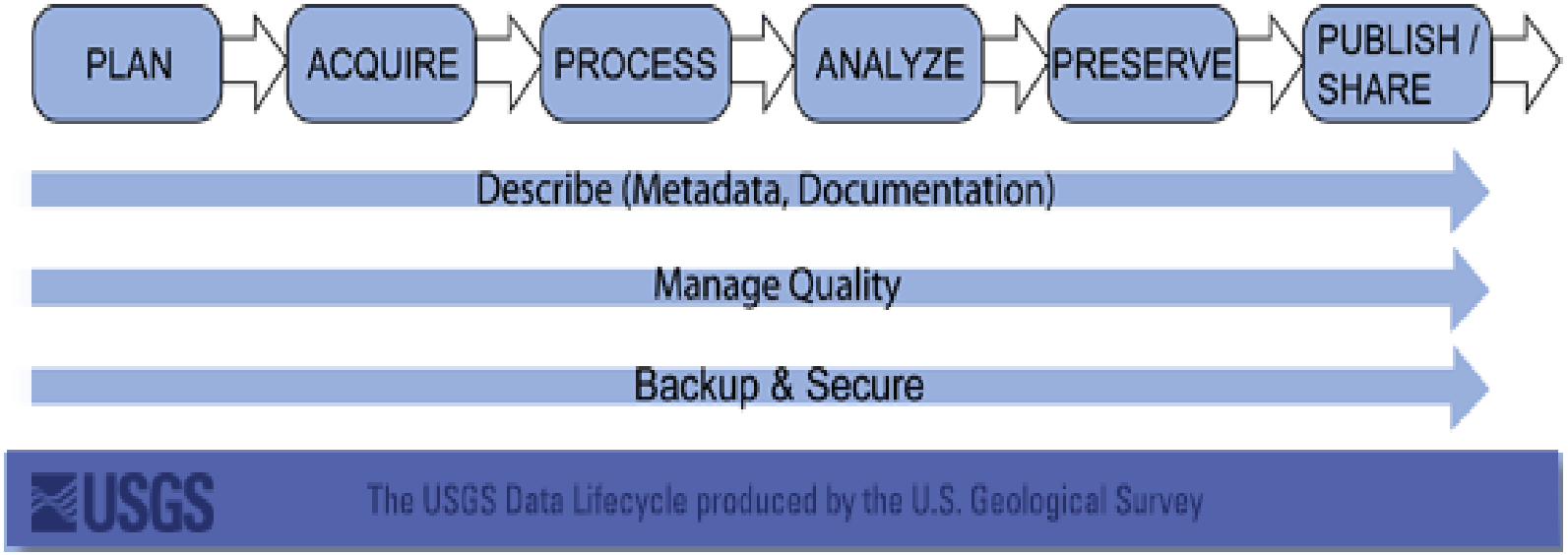

Fig. 4 USGS Data Lifecycle Model (USGS, 2013)

What are some key data lifecycle processes (USGS, 2013)?

Process the data

- Verify, organize, transform, and extract data in an appropriate output for subsequent use

Analyze the data

- Perform actions and method that describe facts, detect patterns, develop explanations, and test hypothesis

What are some key data lifecycle processes (USGS, 2013)?

Preserve the data

- Perform actions and procedures to keep data for specific period of time for future use (e.g. data retention strategy)

Publish/Share the data

- Process to prepare data for dissemination, public access, and reuse (includes documentation and metadata to facilitate aggregation, dissemination, and representation)

What are some key data lifecycle processes?

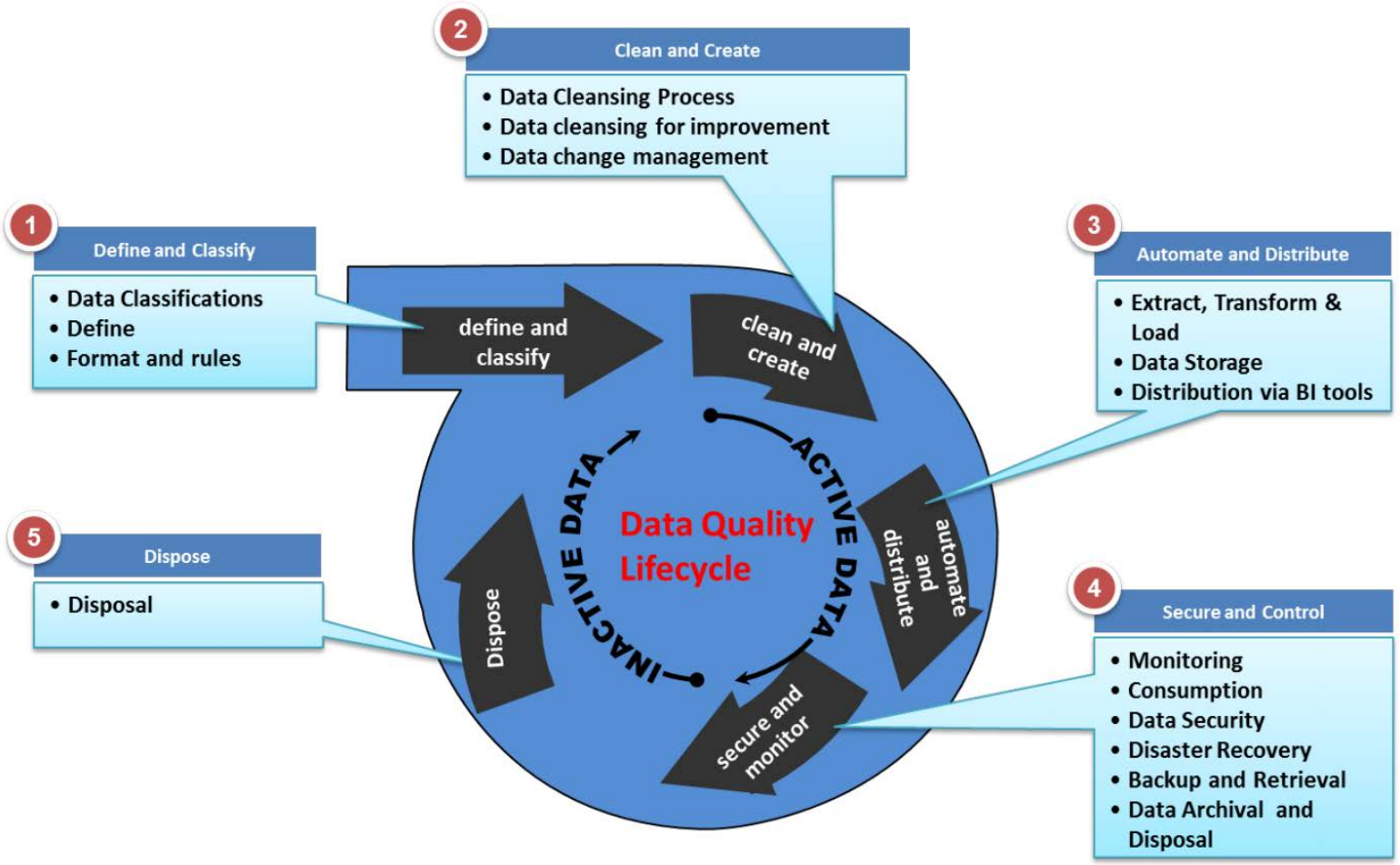

Fig. 5 UNSW Data Lifecycle Model (UNSW, 2017)

What are some key reproducible data processes (ACM SIGMOD, 2017/2018)?

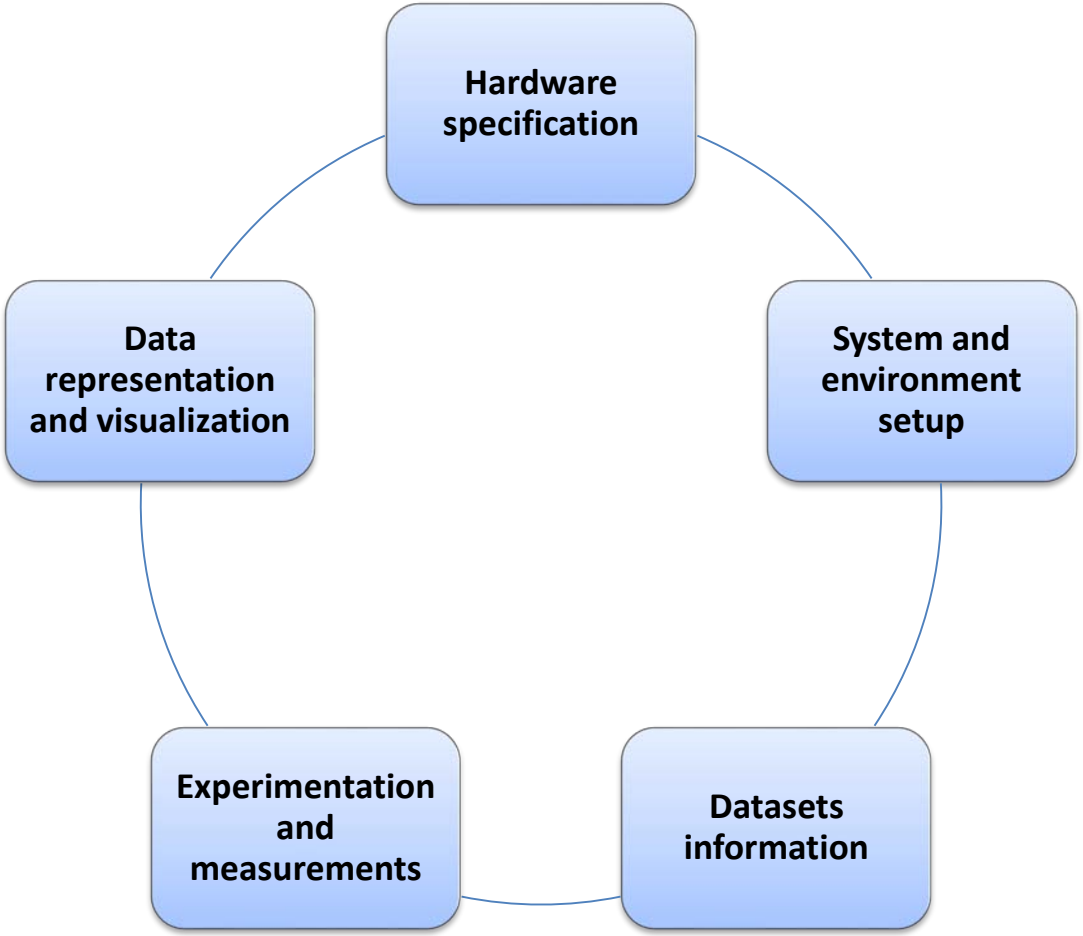

Fig. 6 Reproducibility Template Components (DASlab, Harvard SEAS, 2017)

UF | George A. Smathers Libraries

What are some key research data workflow processes – Physical Experiment Report?

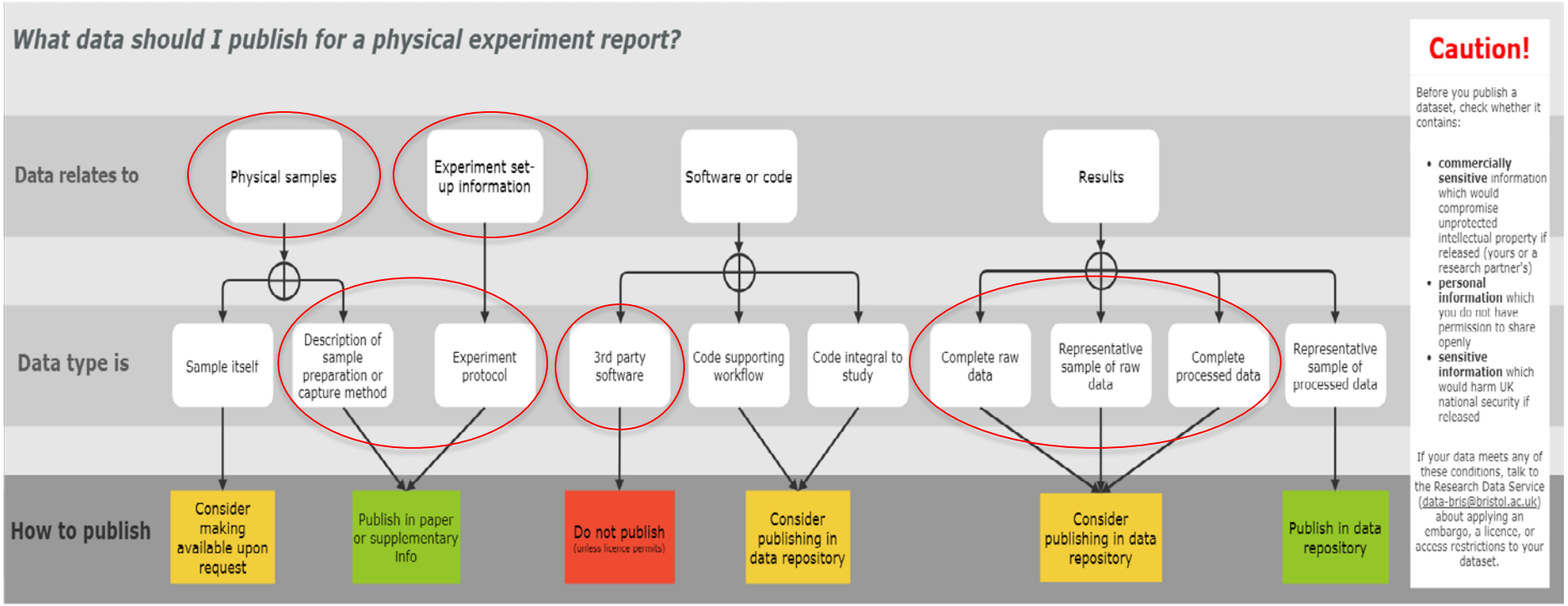

Fig. 7 Physical Experiment Report data publication guide (Beckles, 2018)

UF | George A. Smathers Libraries

What are some DMP examples and the DMPTool?

DMP Examples – UF researchers

- USDA-NIFA - $5m (2018)
- UF/IFAS NCBS - $1.2m (2017)
  - http://ufdc.ufl.edu/AA00014835/00111?search=dmctf
- UF/SFRC - $480k (2016)
  - http://ufdc.ufl.edu/AA00014835/00088?search=dmctf

DMPTool and Resources

- A free tool to create DMP
  - https://dmptool.org/
- UVA Libraries RDS – Data Types & File Formats
  - http://data.library.virginia.edu/data-management/plan/format-types/
- University of Minnesota Libraries Managing Your Data – Storing Data Securely
  - https://www.lib.umn.edu/datamanagement/storedata
- MSU Libraries – Grants and Related Resources
  - http://staff.lib.msu.edu/harris23/grants/index.htm

UF | George A. Smathers Libraries

Developing campus collaborations (e.g. UFII/UF RC/UF Carpentries/UF DSI)

UF Carpentries

- Started 2015
- Teach fundamental research data science skills (e.g. SQL)
- 1st Annual Data Symposium collaborator
  - https://uf-carpentry.github.io/2018-03-20-UFDataSymposium/
- UF Carpentry - https://github.com/UF-Carpentry

UF DSI (Data Science & Informatics)

- Student organization
- Teach data science skills and tools (e.g. R, Python)
- 1st Annual Data Symposium contributor/participator
  - https://zenodo.org/record/1207215#.Wtif_S7wZhE
- UF DSI - http://www.dsiufl.org/

References

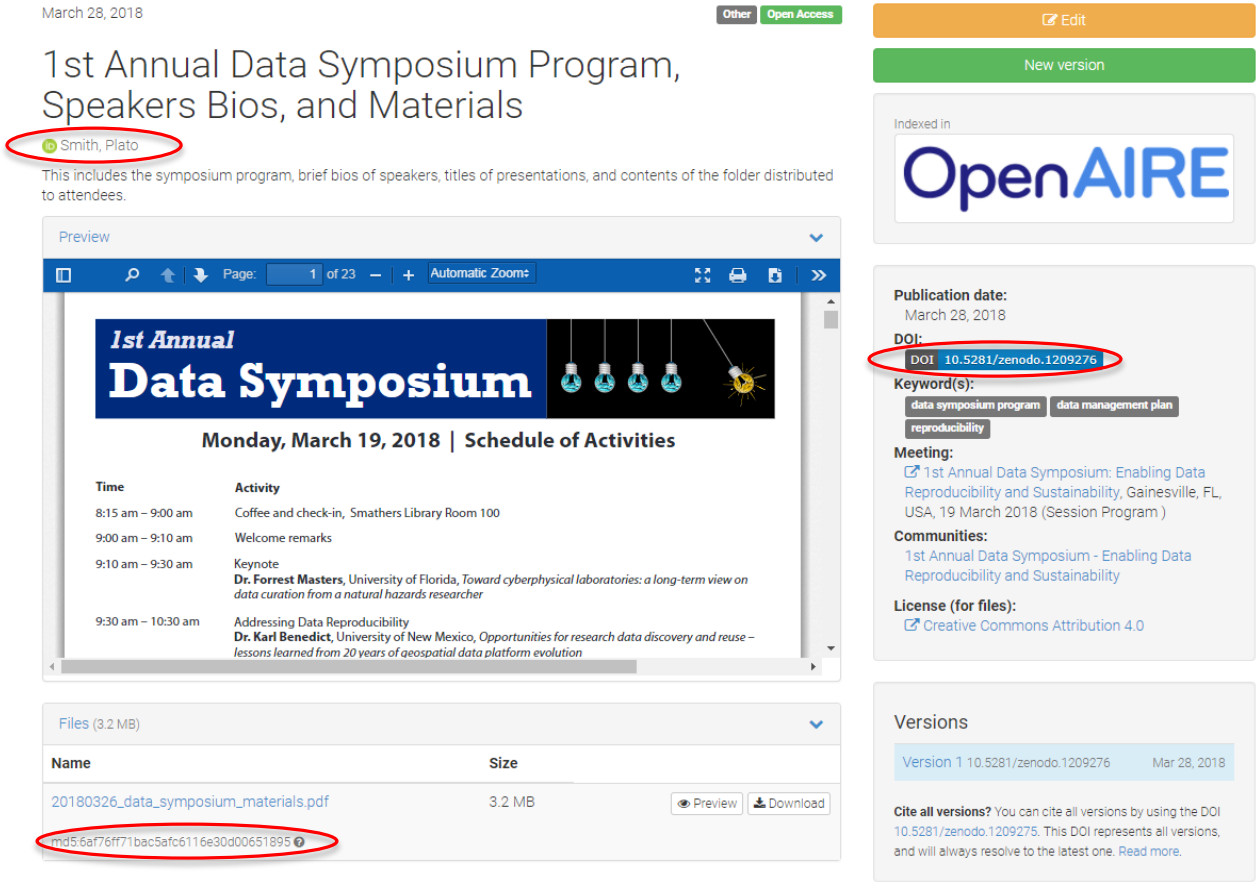

Fig. 8 Learn using zenodo test instance - https://sandbox.zenodo.org/

References

'Research Data Science' is defined by CODATA-RDA as an ensemble of (1) Open Science principles and practices (FAIR) and **research data management and curation skills**, (2) the use of a range of data platforms and infrastructures, (3) large scale analysis, (4) statistics, (5) visualization and modeling techniques, (6) software development and annotation, and (7) more. - http://www.codata.org/working-groups/research-data-science-summer-schools

References

- ACM SIGMOD. (2017/2018). Reproducibility. Retrieved April 19, 2018 from http://db-reproducibility.seas.harvard.edu/.
- Beckles, Zosia (2018): Supporting data for Beckles, Z., Gray, S., Hiom, D., Merrett, K., Snow, K., & Steer, D. (2018) 'Disciplinary data publication guides'. figshare. Fileset. DOI: https://doi.org/10.6084/m9.figshare.5803266.v4.
- cessda. (2018). Consortium of European Social Sciences Data Archive (cessda). Tutorial: OAIS. Retrieved April 19, 2018 from http://tinyurl.com/y7bvyp57.
- DASLab, Harvard SEAS. (2017). Data Systems Laboratory, Harvard School of Engineering and Applied Sciences. http://daslab.seas.harvard.edu/.
- DCC. (2013). Checklist for a Data Management Plan. V.4.0. Edinburgh: Digital Curation Centre. Available online: http://tinyurl.com/pjrmh9n.
- JISC, University of Glasgow – HATII, & DCC. (2009). Data Asset Framework: Implementation Guide. Retrieved April 19, 2018 from http://tinyurl.com/9frmcu6.
- NSF. (2011). Dissemination and Sharing of Research Results. Retrieved April 19, 2018 from http://www.nsf.gov/bfa/dias/policy/dmp.jsp.
- UNSW. (2017). Data Governance Policy. Appendix 1 – Data Management Life Cycle. Retrieved April 19, 2018 from http://tinyurl.com/y7dct55o.
- USGS. (2013). USGS Data Management. Data Lifecycle Overview. Retrieved April 19, 2018 from http://tinyurl.com/ycc6y8sx.

Thank you

Questions

Contact

plato[dot]smith[at]ufl[dot]edu