UF | George A. Smathers Libraries

Developing Data Management Education, Support, and Training

Plato Smith, Data Management Librarian
University of Florida
CAP 5108 Guest Lecture on Data Management
4:05 pm – 4:55 pm, CSE E220
April 12, 2018

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

Linking stakeholders, liaisons, and students

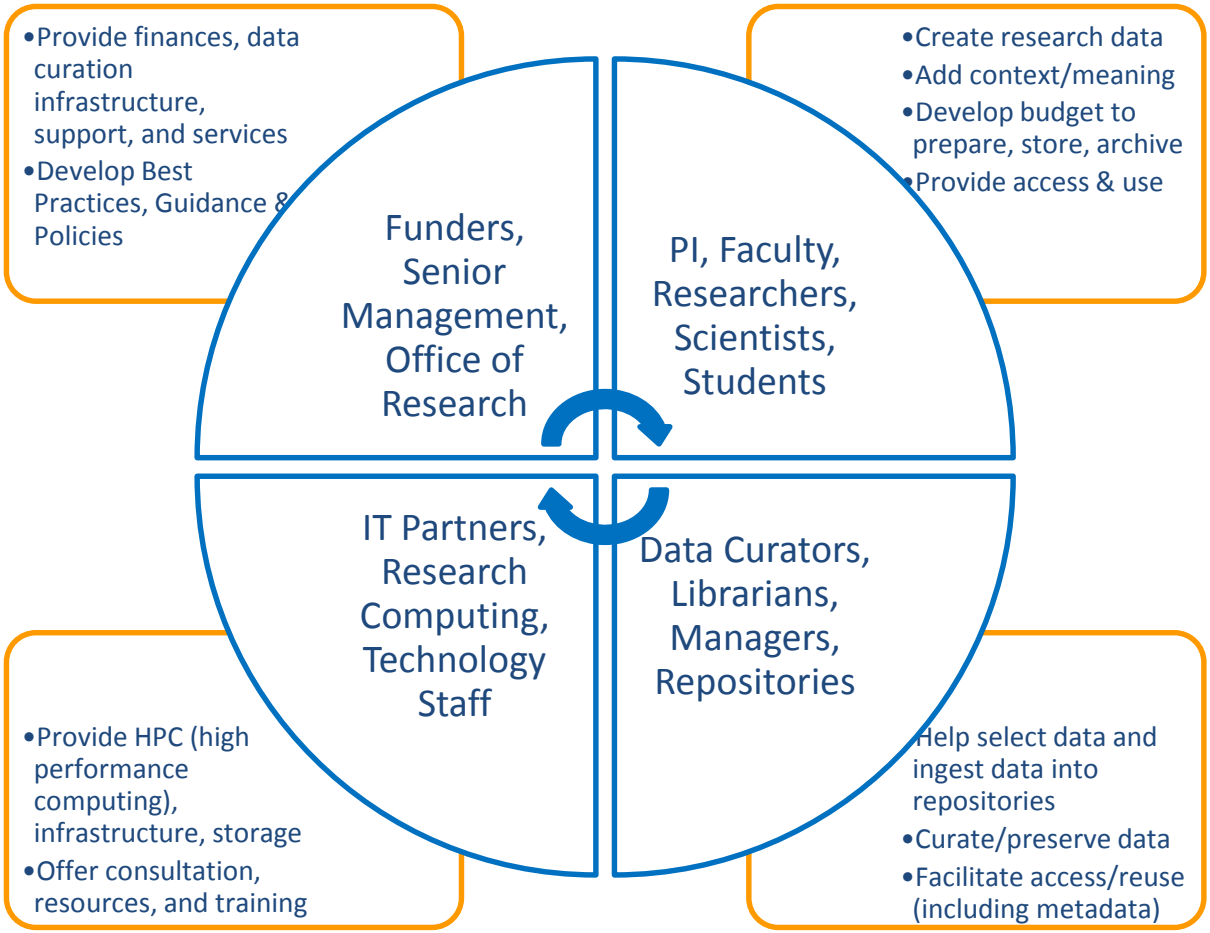

Fig. 1 Stakeholders and Data Management Responsibilities

Articulating a data management plan

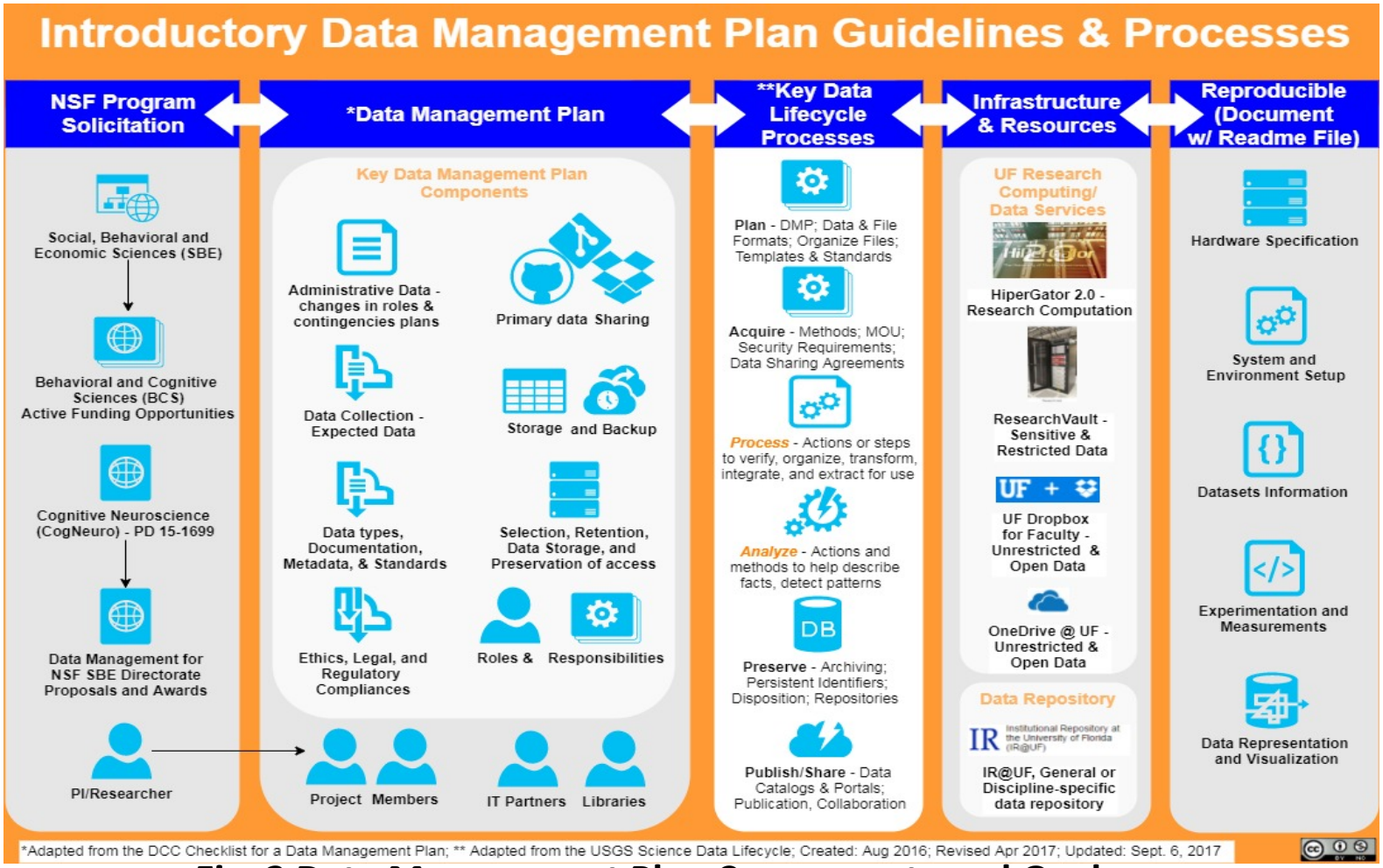

Fig. 2 Data Management Plan Components and Goals

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

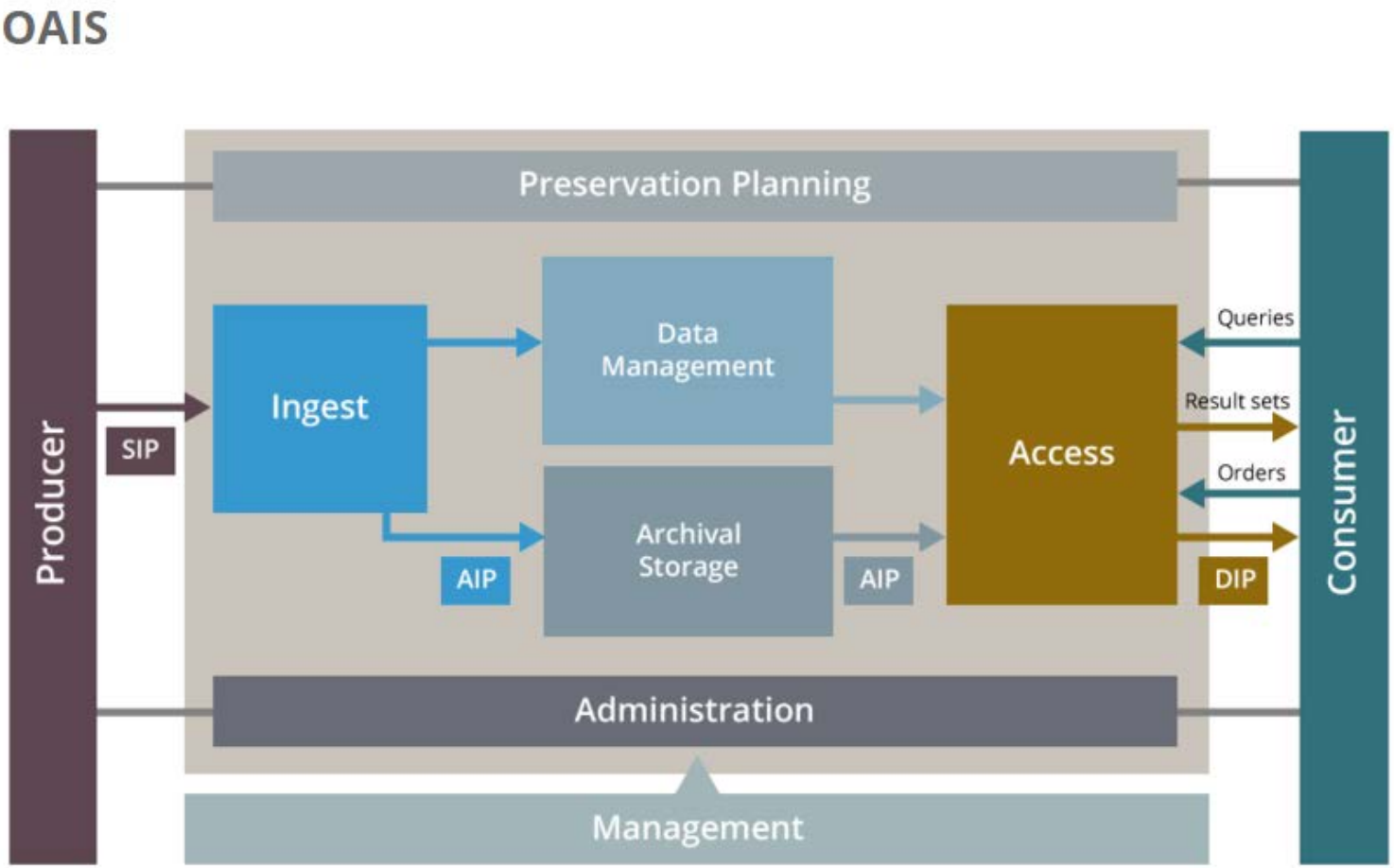

Fig. 3 Open Archival Information System (OAIS) Model (cessda, 2018)

UF George A. Smathers Libraries

Explaining some 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?

Explaining some 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 some 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 some 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)?

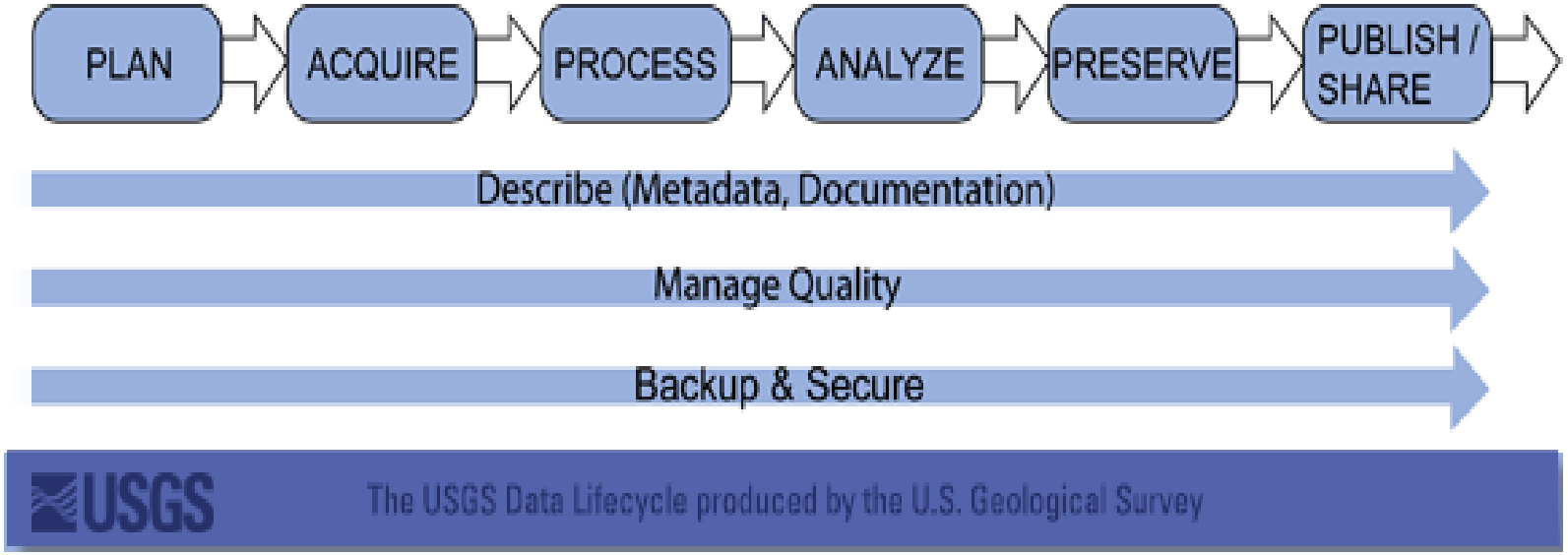

Fig. 4 USGS Data Lifecycle Model (USGS, 2013)

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?

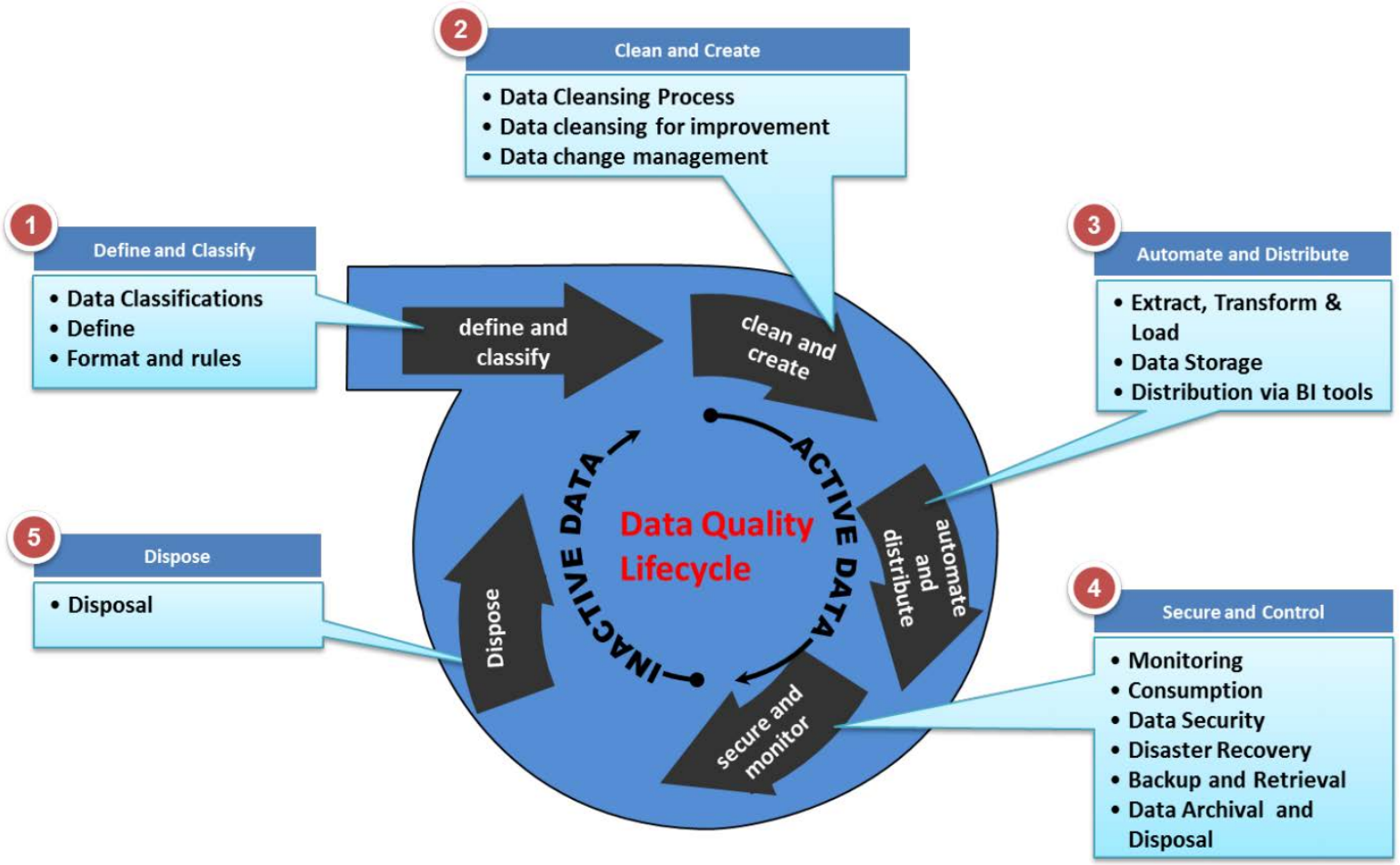

Fig. 5 UNSW Data Lifecycle Model (UNSW, 2017)

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

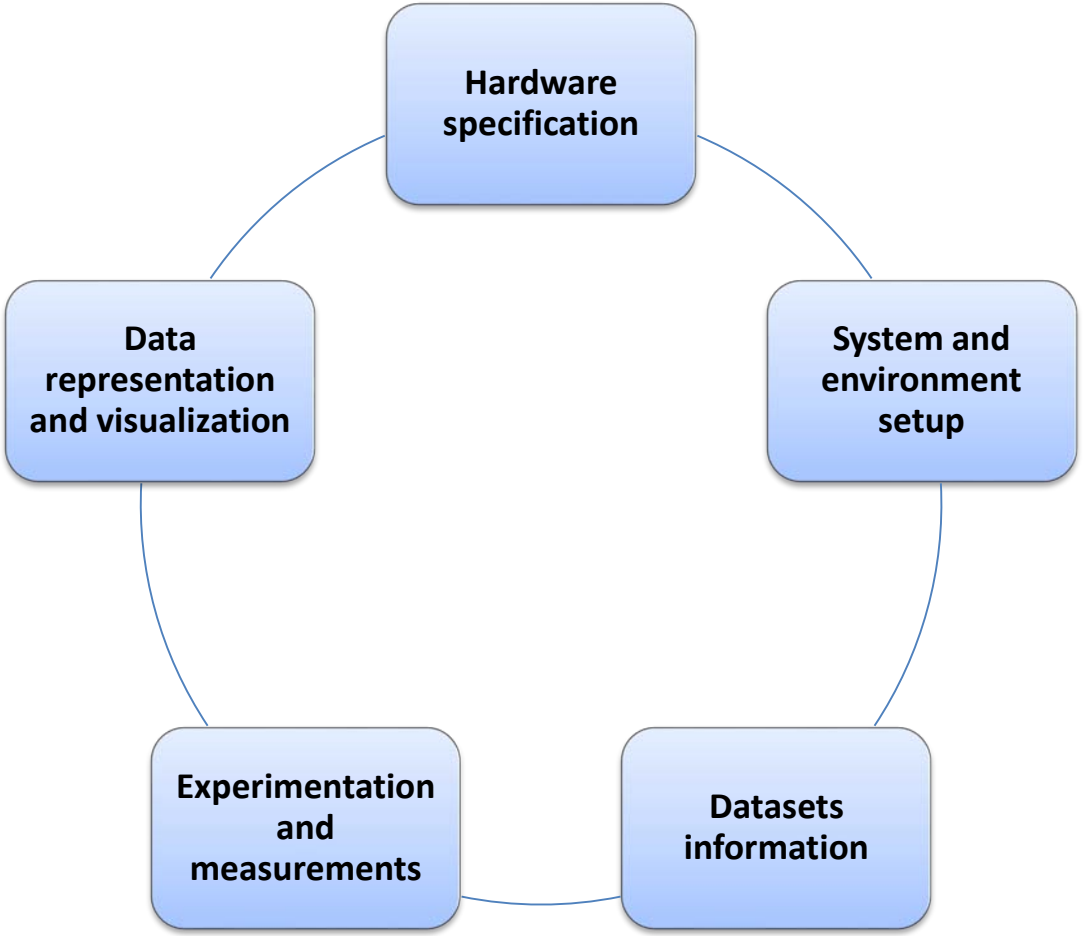

Fig. 6 Reproducibility Template Components (DASlab, Harvard SEAS, 2017)

UF George A. Smathers Libraries

What are some key research data workflow processes – Modeling Study?

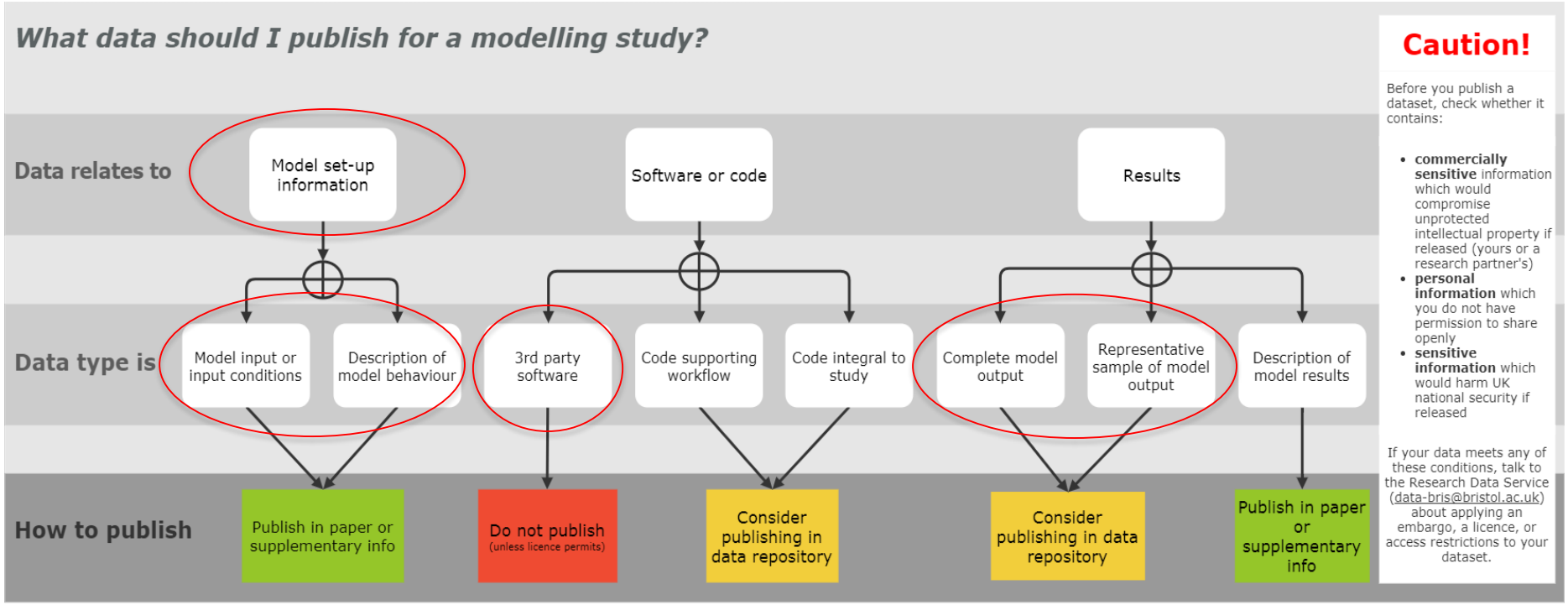

Fig. 7 Modeling Study data publication guide (Beckles, 2018)

UF | George A. Smathers Libraries

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

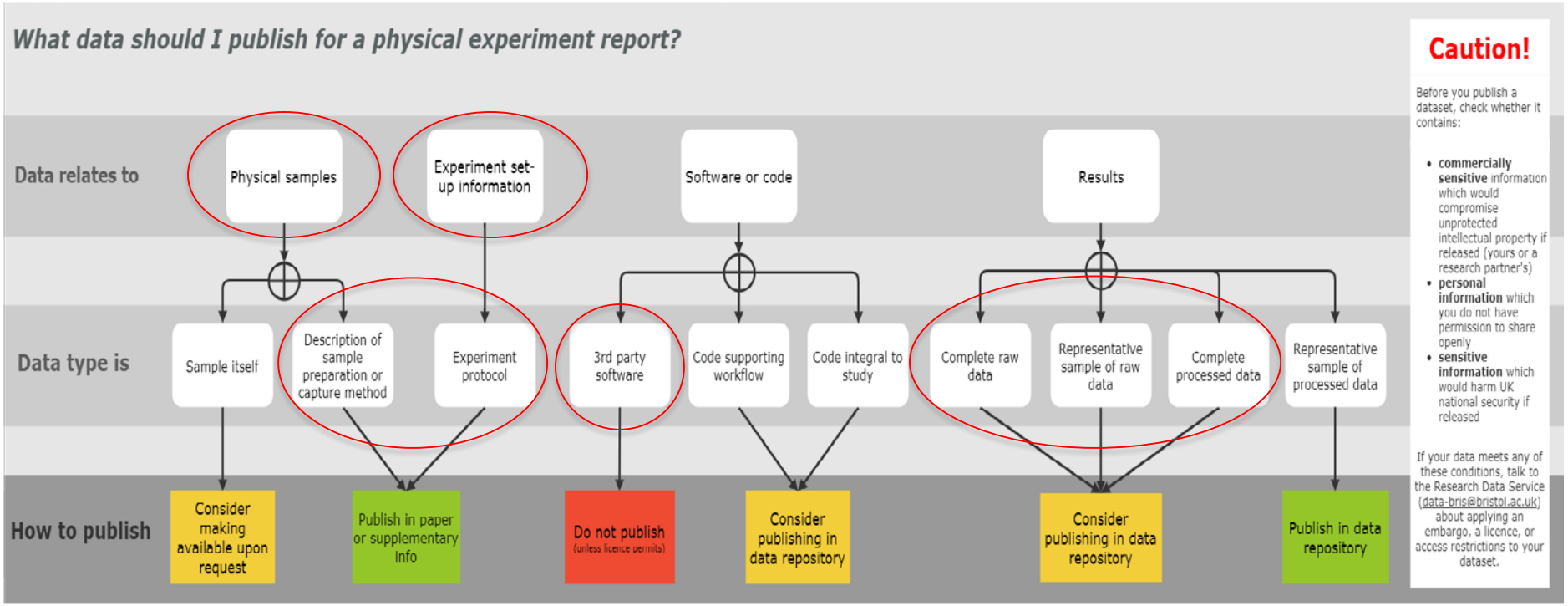

Fig. 8 Physical Experiment Report data publication guide (Beckles, 2018)

UF | George A. Smathers Libraries

What are some key research data workflow processes – Review?

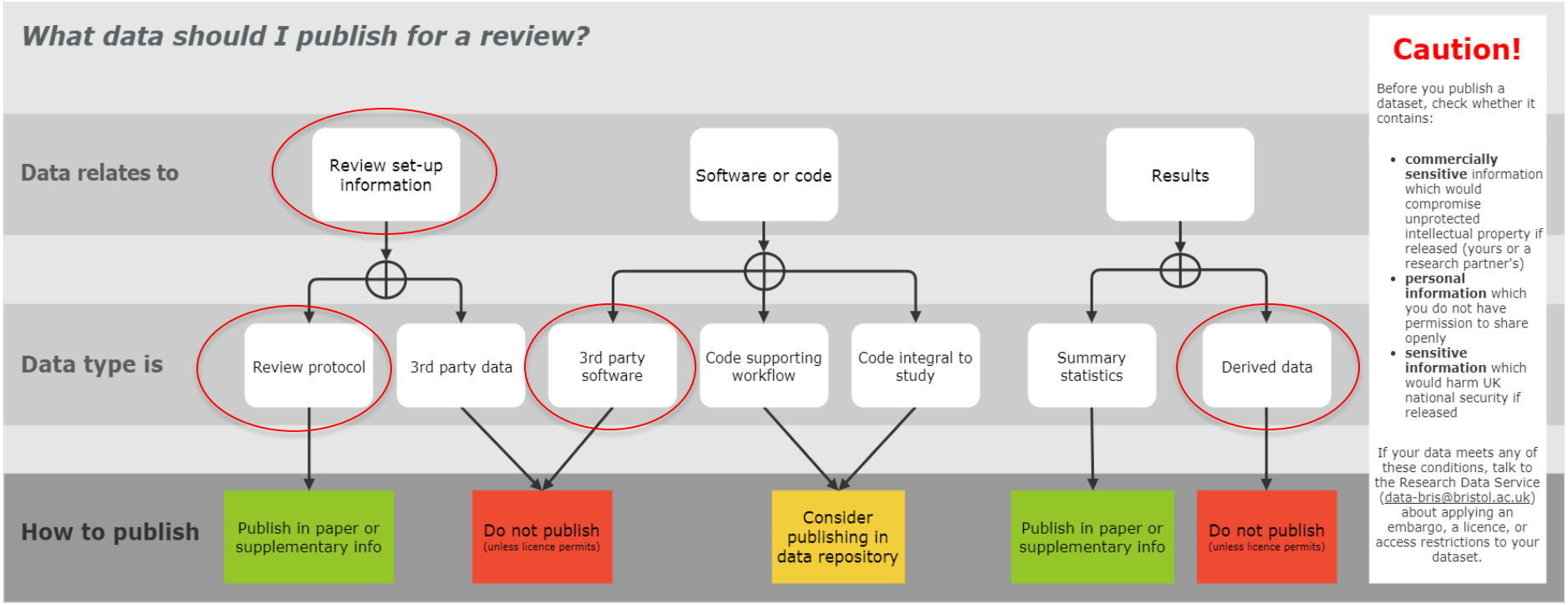

Fig. 9 Review data publication guide (Beckles, 2018)

UF | George A. Smathers Libraries

What are some DMP examples and the DMPTool?

DMP Examples – UF researchers

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

DMPTool

- A free tool to create DMP
  - https://dmptool.org/

References

- ACM SIGMOD. (2017/2018). Reproducibility. Retrieved April 11, 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 11, 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 11, 2018 from http://tinyurl.com/9frmcu6.
- NSF. (2011). Dissemination and Sharing of Research Results. Retrieved April 11, 2018 from http://www.nsf.gov/bfa/dias/policy/dmp.jsp.
- UNSW. (2017). Data Governance Policy. Appendix 1 – Data Management Life Cycle. Retrieved April 11, 2018 from http://tinyurl.com/y7dct55o.
- USGS. (2013). USGS Data Management. Data Lifecycle Overview. Retrieved April 11, 2018 from http://tinyurl.com/ycc6y8sx.

Reproducibility Template Draft

Note: This draft is a revised version of SIGMOD Template 2016, most sections are the reorganized with some summarized information from the SIGMOD Guidelines: http://db-reproducibility.seas.harvard.edu/

The goal is to enable reproducibility of the raw data and relevant plots that the authors used to draw their conclusions. Authors should provide a complete set of scripts to **(1)** install the system, **(2)** produce the data, **(3)** run experiments, and **(4)** produce the resulting graphs along with a detailed Readme file that describes the process systematically for reproducibility by a reviewer or other researchers.

Commented [ly1]: This part is added and introduction is from SIGMOD Guideline

A) Hardware Specification

Commented [ly2]: This part is the original part C" hardware" in SIGMOD template . In draft, I moved it to the beginning because it's the prerequisite needing to be considered even before the software/environment setup.

[Here you should include any details and comments about the used hardware in order to be able to accommodate the reproducibility effort. Any information about non-standard hardware should also be included. You should also include at least the following info:]

1) Processor (architecture, type, and number of processors/sockets)
2) Caches (number of levels, and size of each level)
3) Memory (size and speed)
4) Secondary Storage (type: SSD/HDD/other, size, performance: random/sequential read or write)
5) Network (if applicable: type and bandwidth)

B) System and Environment Setup

Commented [ly3]: This part is the original part A "Source code info" in SIGMOD template. A new name is given.

[System setup is one of the most challenging aspects when repeating experiments. System setup will be easier to conduct if it is automatic rather than manual. You should test that the system they distribute can actually be installed in a new environment. The documentation should detail every step in system setup]

Commented [ly4]: Added based on SIGMOD guideline to give an idea why is part is very important and detailed steps

1) Operation System (e.g., the required compiler must be run with a specific version of the OS)

2) Configuration for the environment if needed (e.g., environment variables, paths)

Commented [ly5]: Added based on SIGMOD guideline—System part

3) Programming Language: [C/C++/java/...]

4) Additional Programming Language info: [optional, e.g., java version]

5) Packages/Libraries Needed: [an as thorough as possible list of software packages needed]

6) Compiler Info: [full details of compiler and version]

7) Procedures to test if the system is configured correctly

[Ideally, there is a script called: ./prepareSoftware.sh]

Commented [ly6]: Added based on SIGMOD guideline—System part

C) Datasets Info

1) Repository: [url]
2) Data generators: [url]

D) Experimentation and Measurements

1) Scripts and how-tos to generate all necessary data or locate datasets

   [Ideally, there is a script called: ./prepareData.sh]

2) Scripts and how-tos for all experiments executed and measurements are taken

   [Ideally, there is a script called: ./runExperiments.sh]

3) Scripts for a clean-up phase where the system is prepared to avoid interference with the next round of experiments.

Commented [ly7]: This part is added based on Guideline—Experiments part,
The third phase

E) Data Representation and Visualization

Commented [ly8]: This whole part is added based on the Guideline –Graphs and Plots part

[For each graph in the paper, you should describe how the graph is obtained from the experimental measurements.]

1) Tools that are used to generate the graphs (e.g., Gnuplot or Matplotlib)
2) Scripts (or spreadsheets) how to generate the graphs.

==========