HOW FAIR IS YOUR RESEARCH DATA?

Brief Review of FAIR Guiding Principles
to Foster Open Science and Research Data Management

Plato Smith, Data Management Librarian
April 18, 2019
9:35 am – 10:25 am, Weil 0279
CAP5108: Research Methods for Human-Centered Computing guest lecture on Data Management

Context

"**Open science** is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods." - FOSTER. (2019). Open Science Definition. https://bit.ly/2IJEnTO.

Researchers are now expected to make their funded research data **findable**, **accessible**, **interoperable**, and **reusable** (**FAIR**) with existing capacity, infrastructures, and resources. This workshop will introduce participants to the FAIR data principles, provide some examples of FAIR data, and sharing additional resources for training on making your research data FAIR.

NEW CORE! P42 Data Management and Analysis Core (DMAC) (Required) - 2018

- NIH Funding Mandate
  - *Participating Organization(s) – National Institute of Health (NIH)*
  - *Components of Participating Organizations – National Institute of Environmental Health Sciences (NIEHS)*
  - *Funding Opportunity Title – Superfund Hazardous Substance Research and Training Program (SRP) (P42 Clinical Trial Optional)*
  - *Funding Opportunity Announcement (FOA) Number – RFA-ES-18-002*
  - *New Core – Data Management and Analysis Core (DMAC)*

- NIH DMAC Purpose
  - *Support the management and integration of datasets across the Center, irrespective of dataset size*
  - *Establish, coordinate, and monitor processes for data analysis*
  - *Work with project/core leaders to ensure high data quality through lifecycle of data*
  - *Identify opportunities for integrating project/core-generated data with other existing datasets*
  - *Foster and enable interoperability of data between BMR and ESE projects to accelerate impact of Center's research*

NEW CORE! P42 Data Management and Analysis Core (DMAC) (Required) - 2018

- "**Data Management and Analysis**: The SRP acknowledges that integration of data from a broad range of scientific disciplines will be critical to understanding and breaking the link between exposures and disease. Data represent important research products and attention is needed to provide stewardship of the data and treat research as an asset. To support NIH data sharing policies (https://grants.nih.gov/policy/sharing.htm) and promote best principles **so data is Findable, Accessible, Interoperable, and Reusable** (FAIR), all applicants are required to have a Data Management and Analysis Core (DMAC)." **– NIH, 2018**

Table of Content

What are the FAIR Guiding Principles?

Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
  - *A1.1 The protocol is open, free, and universally implementable*
  - *A1.2 The protocol allows for an authentication and authorization procedure, where necessary*
- A2. Metadata are accessible, even when the data are no longer available

Source of reference: https://www.go-fair.org/fair-principles/

What are the FAIR Guiding Principles?

Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

Reusable

- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
  - *R1.1. (Meta)data are released with a clear and accessible data usage license*
  - *R1.2. (Meta)data are associated with detailed provenance*
  - *R1.3. (Meta)data meet domain-relevant community standards*

Source of reference: https://www.go-fair.org/fair-principles/

What is FAIR Data?

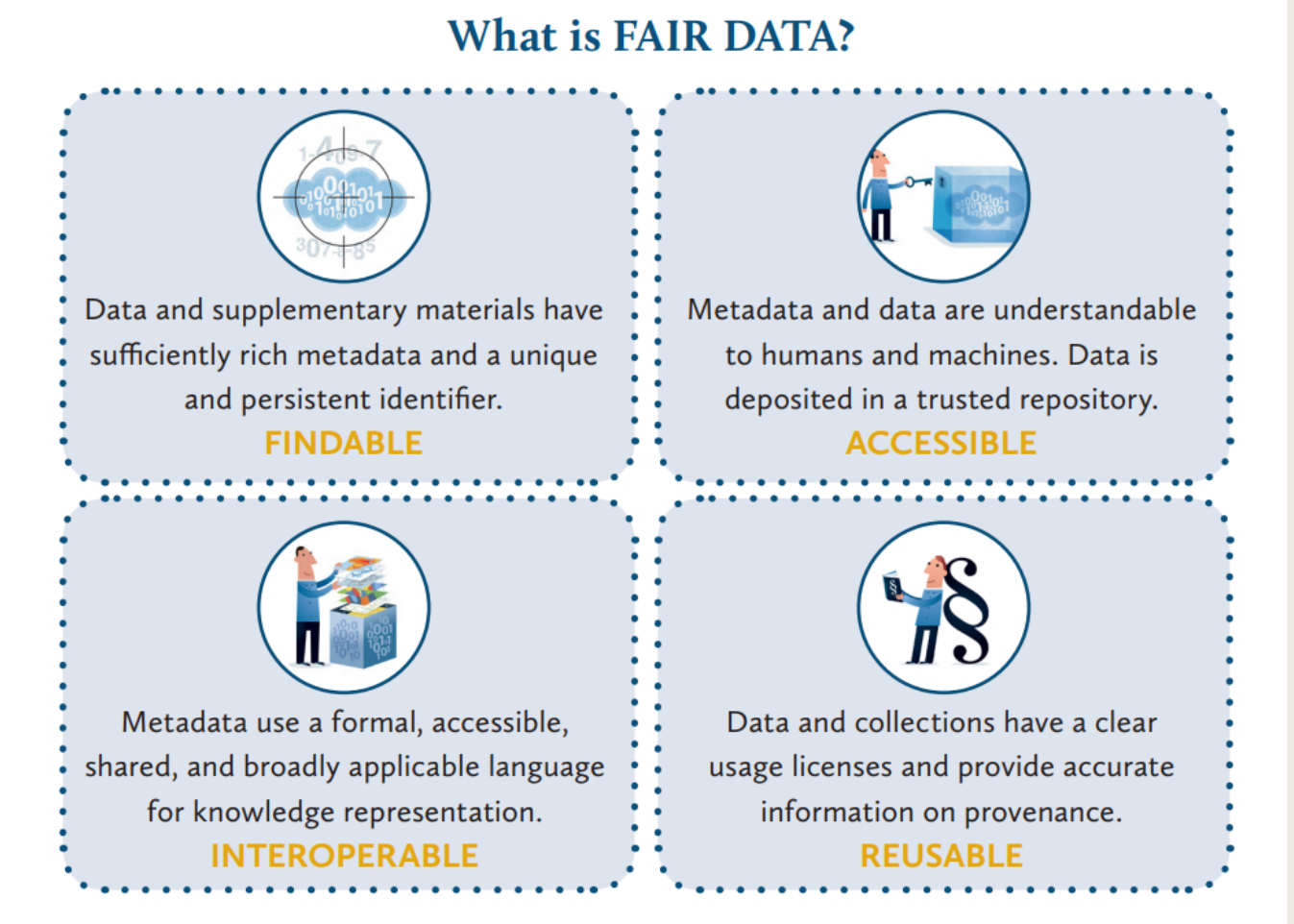

Figure 1. What is FAIR DATA? (Liber, 2017)

What is FAIR Data?

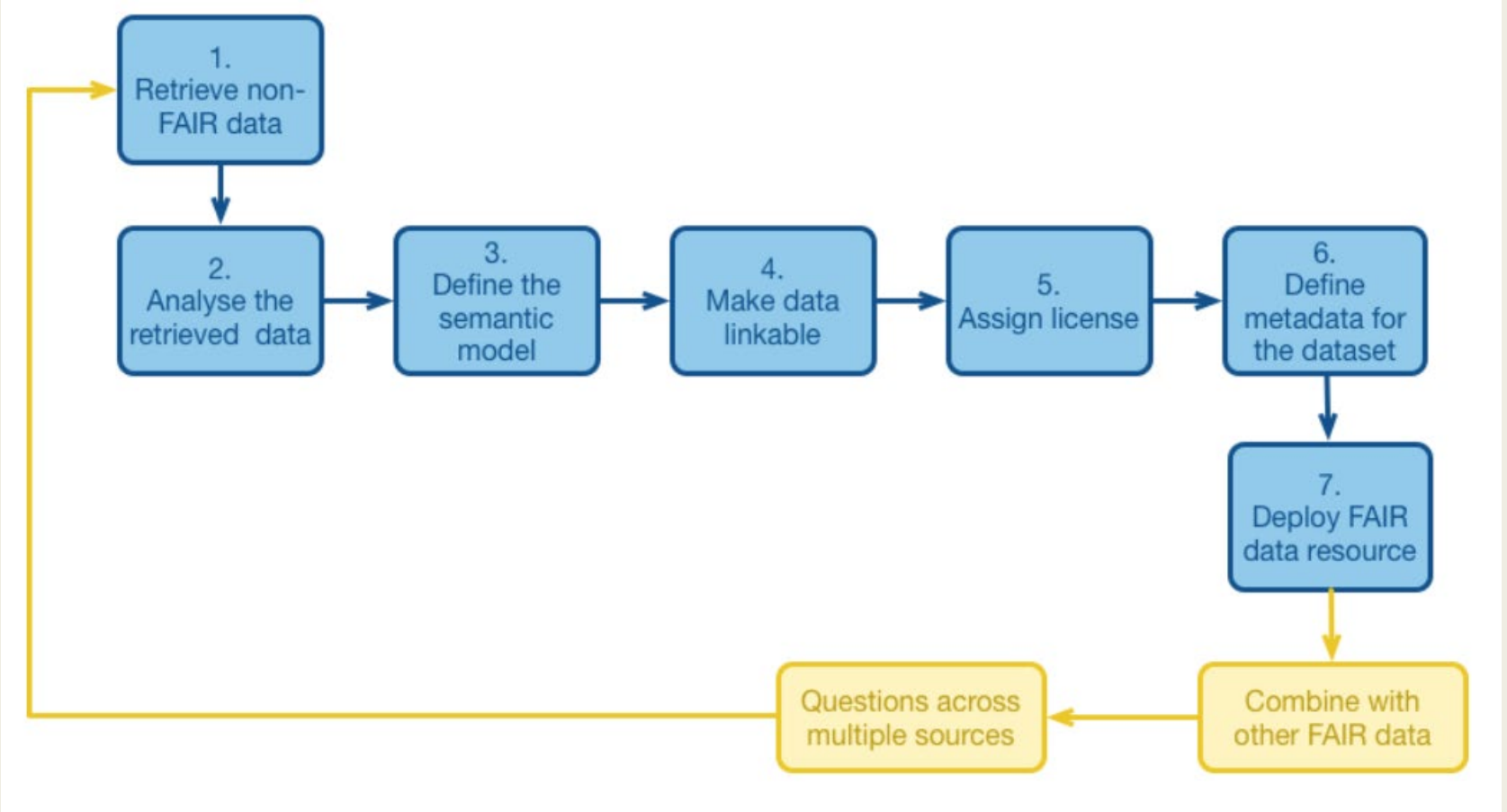

Figure 2. FAIRification Process? (GO FAIR, n.d.)

What are some examples of FAIR Data?

Principle	Explanation	Example in human genetics and genomics
Findability	Datasets should be described, identified and registered or indexed in a clear and unequivocal manner.	BBMRI-ERIC Directory
Accessibility	Datasets should be accessible through a clearly defined access procedure, ideally using automated means. Metadata should always remain accessible.	European genome–phenome archive
Interoperability	Data and metadata are conceptualized, expressed and structured using common, published standards.	Genomic Data Toolkit
Reusability	Characteristics of data and their provenance are described in detail according to domain- relevant community standards, with clear and accessible conditions for use.	BRCA exchange

Table 1. The meaning of the FAIR principles, based on Wilkinson et al. and Mons et al (Boeckhout, et al, 2018)

FOSTER Open Science Taxonomy

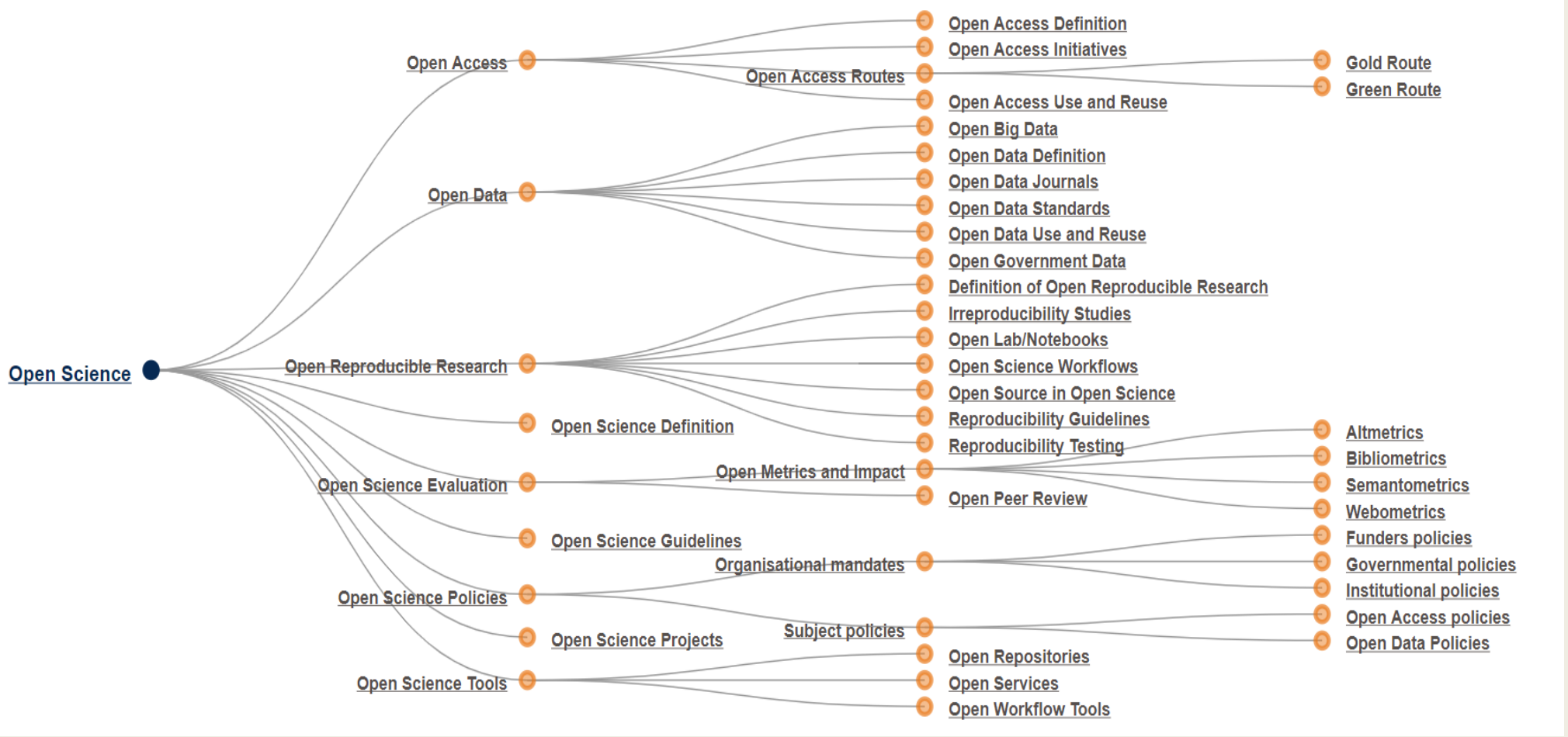

Figure 3. FOSTER Open Science Taxonomy (https://www.fosteropenscience.eu/resources)

FOSTER Open Science Taxonomy

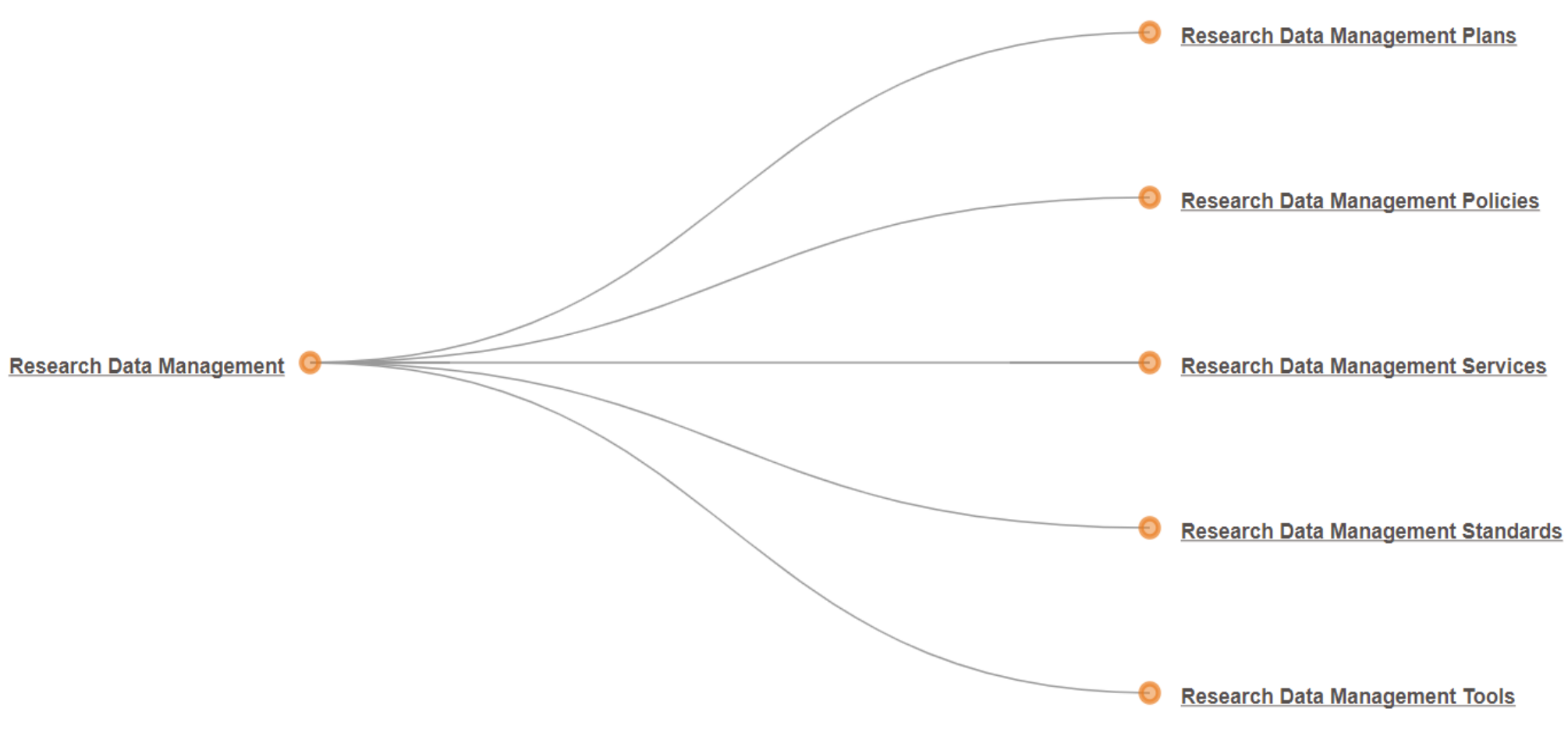

Figure 4. FOSTER Research Data Management Taxonomy (https://www.fosteropenscience.eu/resources)

Developing a Data Management Plan

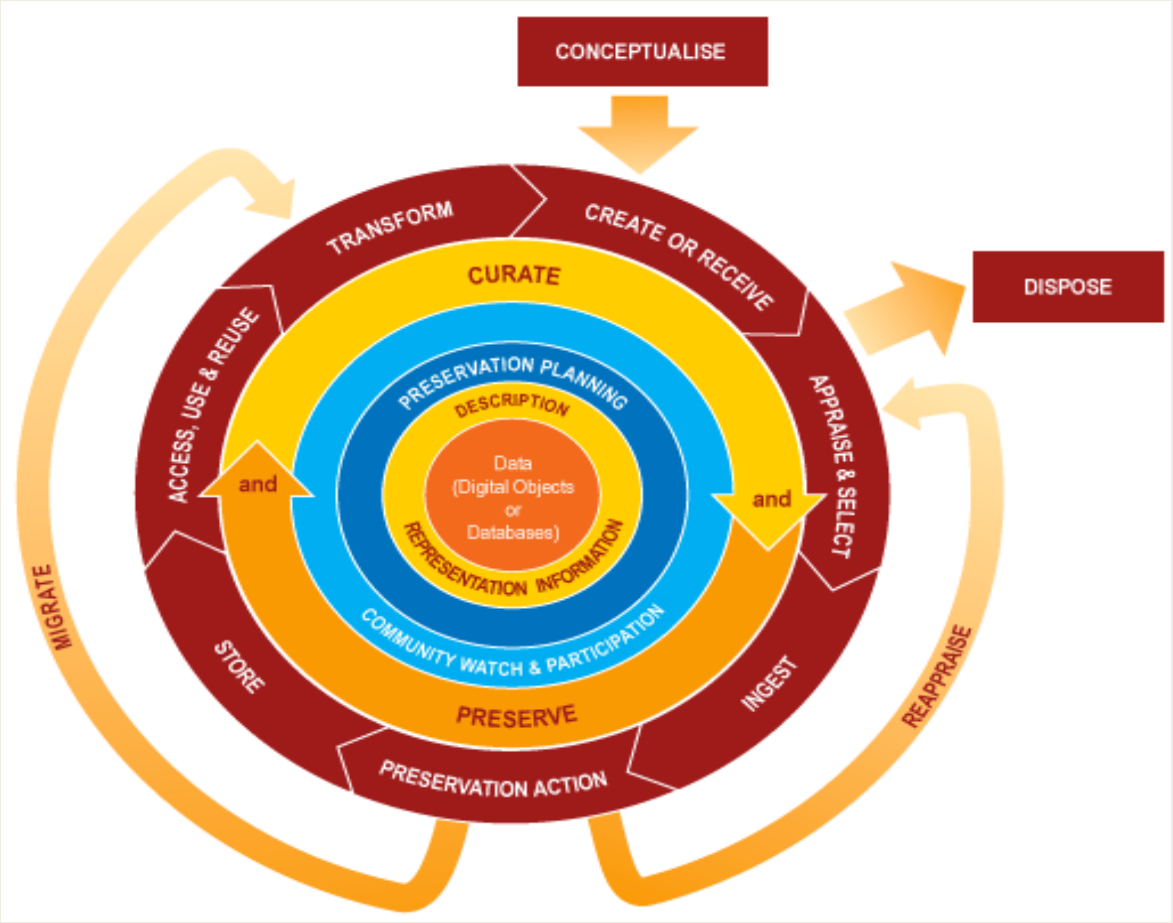

Figure 5. DCC Curation Lifecycle Model (DCC, 2007)

Developing a Data Management Plan

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?*
- 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?*

Source of reference: DCC Checklist for a Data Management Plan (DCC, 2013)

Developing a Data Management Plan

Step 1 - Organize Data and Products (file names and folder organization):

- ❑ Define folder and file names and structure - and use them!
- ❑ Limit the depth of sub-folders to no more than 2 (see example on right)
- ❑ Use meaningful names that include basic information (e.g. date, measurement, collection, etc.)
- ❑ Make the name unique
- ❑ Avoid Spaces
- ❑ Use ASCII Characters only
- ❑ Document, share, and evaluate
- ❑ Separate classes of products: raw data, derived data, graphics, code, documents, etc.
- ❑ Make Backups!

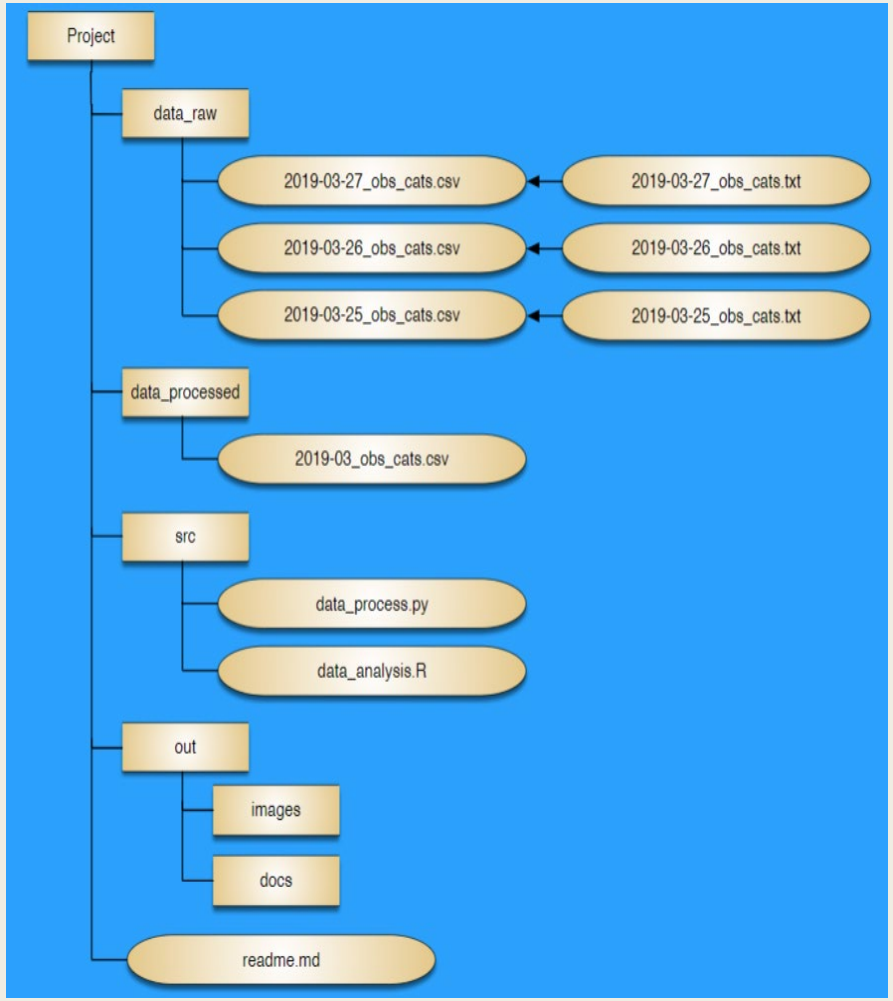

Figure 6. Data Management Skills Training Resources, Organizing Data & Products (Benedict, 2019)

Developing a Data Management Plan

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 gained 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?*

Source of reference: DCC Checklist for a Data Management Plan (DCC, 2013)

Developing a Data Management Plan

Storage and Backup

- How will the data be stored and backed up during research (e.g. HiperGator)?
  - *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 who will manage these?*

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 and in which repository or archive will the data be held (e.g. NCBI, NCEI)?*

Source of reference: DCC Checklist for a Data Management Plan (DCC, 2013)

Developing a Data Management Plan

Data Sharing

- How will you share the data?
  - *How will potential users find out about your data?*
- Are any restriction on data sharing required?
  - *What action will you make to overcome or minimize restrictions?*

Responsibilities & Resources

- Who will be responsible for data management?
- What resources will you require to deliver your plan?
  - *Is additional specialist expertise (or training for existing staff) required?*

Source of reference: DCC Checklist for a Data Management Plan (DCC, 2013)

Developing a Data Management Plan

Objective	Output name	Output description	Output (type, format)
Obj. 1	Synthesized datasets	Habitat; Fisheries independent; Fisheries dependent	Habitat (derived, geospatial), Fisheries (derived, tabular)
Obj. 2	Hierarchical analyses of spatial recruitment and angler effort	Reports; Instructions for analyses; Data analyses code; Geospatial images	Reports and Instructions (text, PDF/XML); Code (text, .txt); Geospatial (TIFF and GIS)
Obj. 3	Socio-ecological regional system model analyses	Reports; Instructions for analyses; Data analyses code	Reports and Instruction (text, PDF/XML); Code (text, .txt)
Obj. 4	Restoration management strategy evaluation (MSE)	Simulation results; Reports; Instructions for analyses; Data analyses code	Simulation (simulated data, CSV); Reports and Instructions (text, PDF/XML); Code (text, .txt)

Table 2: Description of project data output and products in a DMP (Lorenzen, et al., 2016)

Developing a Data Management Plan

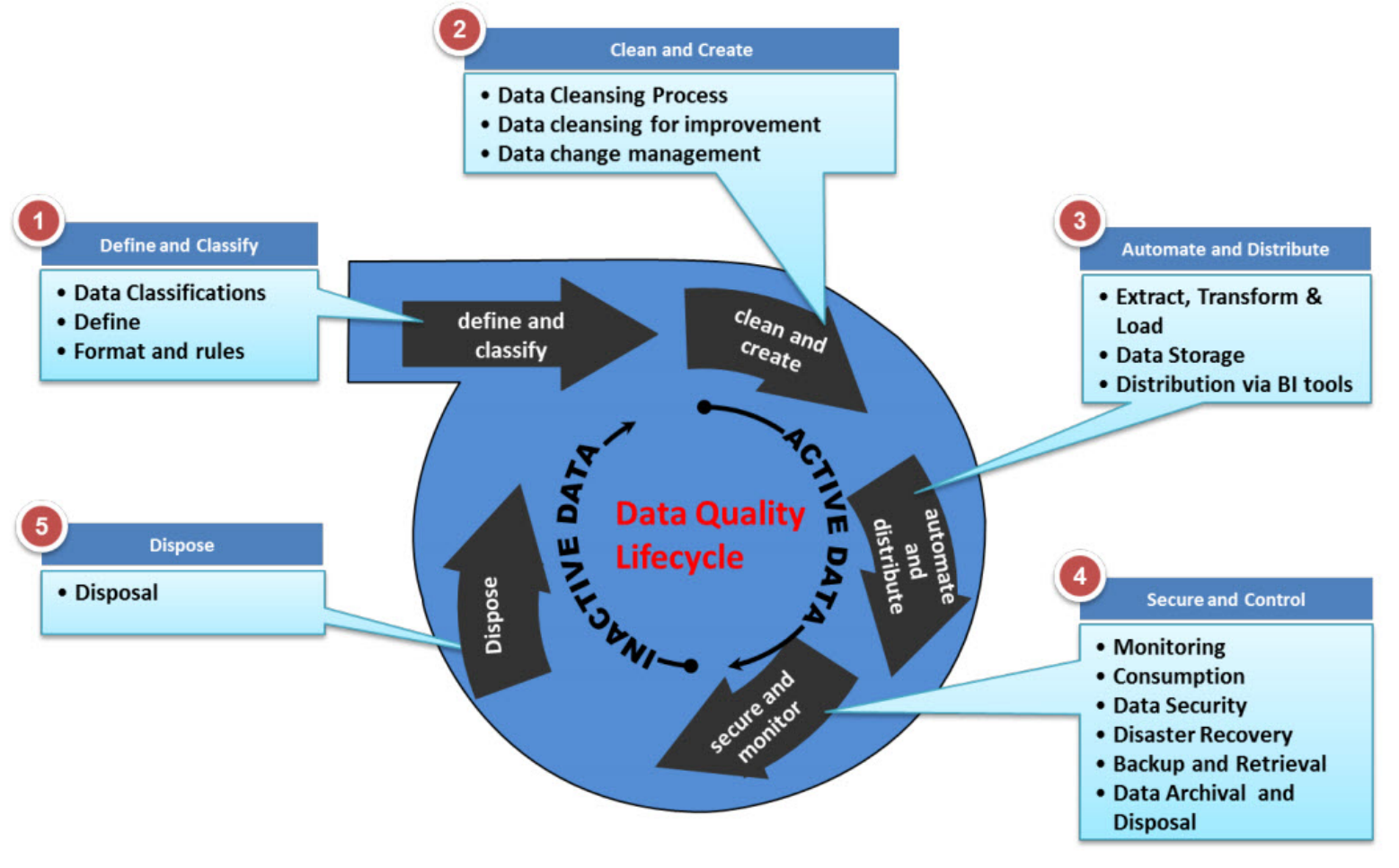

Figure 7. Data Management Lifecycle Model (UNSW, 2017)

Developing a Data Management Plan

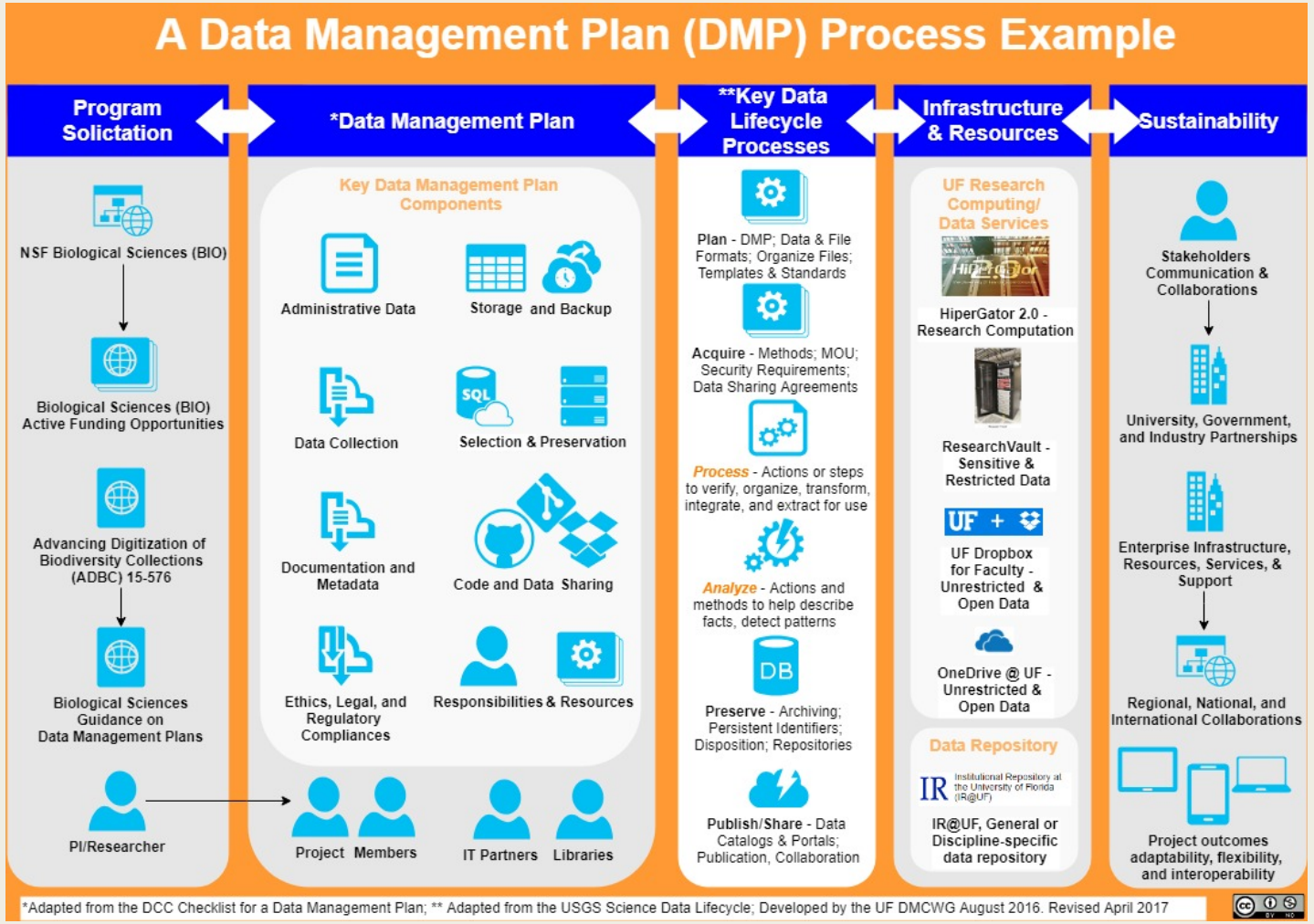

Figure 8. Data Management Plan (DMP) Process Example

References

- Benedict, K. (2019). Data Management Skills & Training Resources. Presentation to the INSC590 Problems: Information Science – Data Management graduate course at the University of Tennessee, Knoxville School of Information Sciences. March 27, 2019.
- Boeckhout, M. Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: fair enough? *European Journal of Human Genetics* 26, pp. 931-936. https://www.nature.com/articles/s41431-018-0160-0.
- DCC. (2007). Digital Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model.
- Foster. (2019). Open Science. Taxonomy. https://www.fosteropenscience.eu/foster#taxonomy.
- GO FAIR. (n.d.). FAIR Principles. https://www.go-fair.org/fair-principles/.
- LIBER. (2017). Implementing FAIR Data Principles: The Role of Libraries. https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf.
- Lorenzen, K., Camp, E., & Dutka-Gianelli, J. (2016). Synthesizing spatial dynamics of recreational fish and fisheries to inform restoration strategies: red drum in the Gulf of Mexico. Revised Data Management Plan. http://ufdc.ufl.edu/AA00014835/00088.
- Mons, et al. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use* 37, pp. 49-56. https://content.iospress.com/articles/information-services-and-use/isu824.
- NIH. (2018). RFA-ES-18-002, Data Management and Analysis. https://grants.nih.gov/grants/guide/rfa-files/RFA-ES-18-002.html.
- UNSW. (2017). Data Governance Policy, Data Management Lifecycle Model, Appendix 1. Accessed online April 17, 2019 from https://www.gs.unsw.edu.au/policy/documents/datagovernancepolicy.pdf.
- Wilkinson, M. D, et al. (2016). The FAIR Guiding Principles for Scientific data management and stewardship. *Scientific Data* 3, Article number: 160018. https://www.nature.com/articles/sdata201618.

Resources

- CESSDA Training Working Group. (2017 – 2018). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide.
- DataONE Community Engagement & Outreach Working Group. (2017). Data Management Skillbuilding Hub. Accessed through the Data Management Skillbuilding at https://dataoneorg.github.io/Education/.
- DCC. (2013). Checklist for a Data Management Plan. V.4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans.
- DMPTool. (2019). https://dmptool.org/.
- Franck, G. and Brinken H. (2019). FOSTER: Research Data Management Data Management Plans. https://www.fosteropenscience.eu/node/2663.

Thank you

Questions

Contact:

Plato Smith, Data Management Librarian, plato[dot]smith[dot][at]ufl[dot]edu

Data Management and Curation Working Group

DATAMGMT-L@LISTS.UFL.EDU