

Planning for Data Management

Workshop instructor: Plato Smith, Ph.D.

Date: September 14, 2021

Time: 2:30 pm – 3:30 pm

Location: Zoom





Acknowledgements

- Flora W. Marynak (UF Informatics Institute (UFII) and UF Biodiversity Institute (UFBI))
- Alethea Geiger (UF Informatics Institute)
- Dr. George Michailidis (UF Informatics Institute)
- Chelsea Johnston (UF George A. Smathers Libraries)
- Dr. Laurie Taylor (UF George A. Smathers Libraries)
- Patrick Reakes (UF George A. Smathers Libraries)
- Dr. Matt Gitzendanner (UF Biology/FLMNH/Research Computing)
- Dr. Erik Deumens (UF Research Computing)
- Christopher P. Barnes (UF Clinical and Translational Science – IT (CTS-IT))
- Dr. Mike Allen (UF Institute of Food and Agricultural Sciences (IFAS))
- Dr. Bill Pine (UF Institute of Food and Agricultural Sciences (IFAS))
- Dr. Eakta Jain (UF Department of Computer & Information Science & Engineering)
- Dr. Forrest Masters (UF College of Engineering)
- Dr. Sobha Jaishankar (UF Office of Research)
- Dr. Stephanie Gray (UF Office of Research)



Context

“Responsible data management, and the resulting access to research data, can contribute to an improved public understanding of the university’s contributions to the public [**Greater Gator**] good.” – Erway, 2013

“A record [**research data**] if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted.” – Bush, 1945



Table of Content



1. Definitions
2. Planning ahead for your data management needs
3. Key components of a data management plan
4. Key stakeholders involved
5. Some UF infrastructure and resources
6. Examples of DMPs from successful awards by UF Researchers
7. DMPTool – Build your Data Management Plan
8. References



Definitions

- **Data formats** – “Packages of information that can be stored as data files or sent via network as data streams (aka bitstreams, byte streams).” – Library of Congress, 2017
 - **Examples of data formats:** Still Image, Sound, Moving Image, Textual, Web Archive, Datasets, Geospatial, Generic. (Library of Congress, 2019)
 - Browse alphabetical list of data formats via <https://tinyurl.com/yj833zvw>.



Definitions

- **Metadata** – “... is structured data [descriptive information] about anything that can be named, such as Web pages, books, journal articles, images, songs, products, processes, people (and their activities), **research data**, concepts, and services.” – DCMI, 2021
- **Research data** – “[T]he recorded factual material commonly accepted in the scientific community as necessary to validate research findings.” – OMB Circular A-110, 1999



Definitions

- **Types of research data:**

- **Computer code** – model or simulation source code; preserve the model and associated metadata with computational data arising from the model
- **Derived data** – resulting from processing or combining ‘raw’ or other data (where care may be required to respect the rights of owners of the raw data)
- **Experimental data** – scientific experiments and computational results, which may in principle be reproduced although it may in practice prove difficult or not cost effective
- **Observational data** – scientific phenomena at a specific time or location where the data will usually constitute a unique irreplaceable record
- **Reference data** – canonical or reference data relating for example to gene sequences, chemical structures or literary text

Definitions

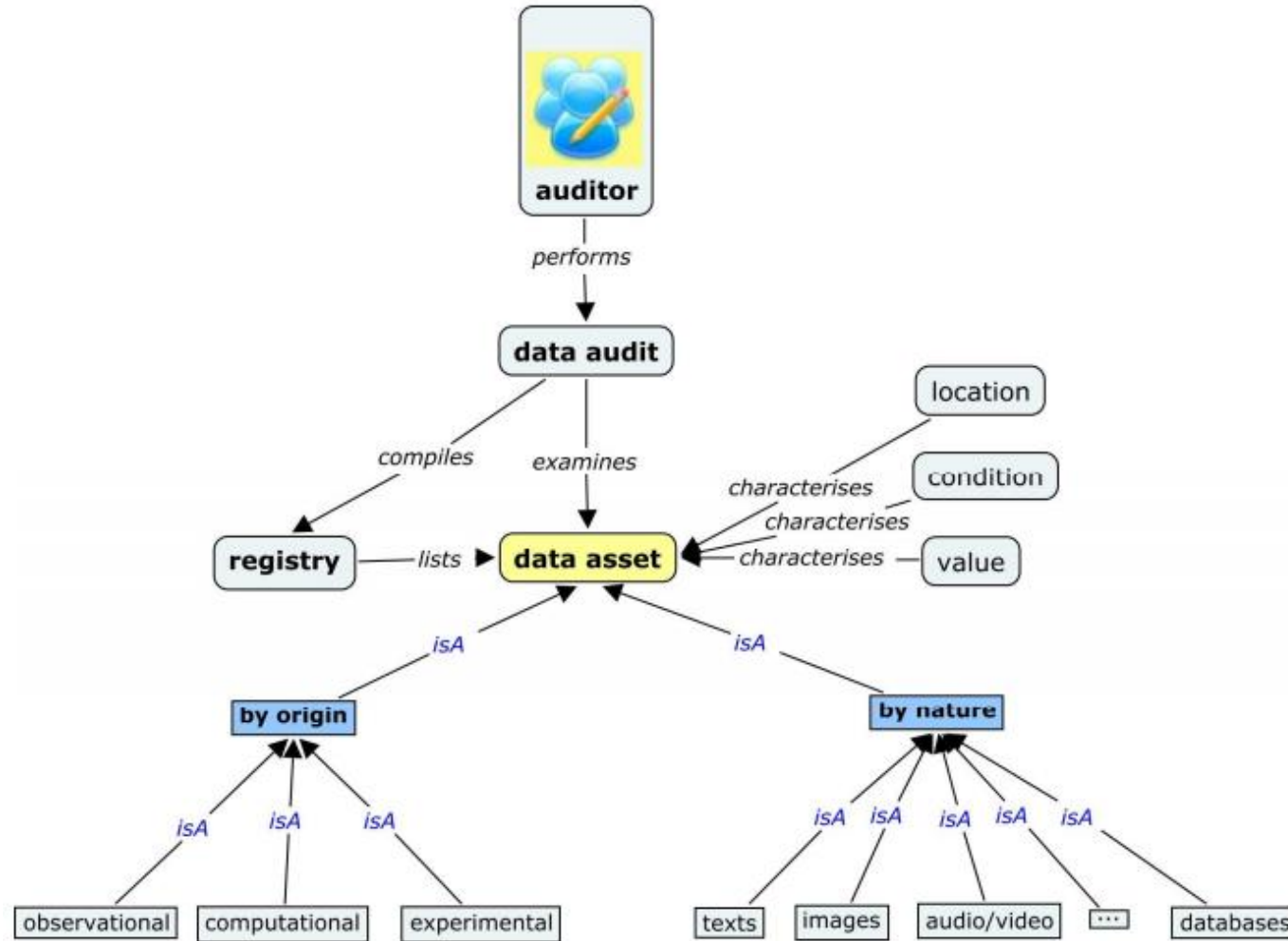


Fig. 1 – Concept map scope of Data Audit Framework (Jones et al, 2009)



Definitions

| Audit Form 3A: Data asset management (Core element set) | | |
|---|------------------------------|--|
| No | Parameter | Comment |
| 1 | ID | <i>A unique identification assigned by the audior or organisation to each data asset</i> |
| 2 | Data creator(s) | <i>Person, group or organisation responsible for the intellectual content of the data asset</i> |
| 3 | Title | <i>Official name of the data asset, with additional or alternative titles or acronyms if they exist</i> |
| 4 | Description | <i>A description of the information contained the data asset and its spatial, temporal or subject coverage</i> |
| 5 | Subject | <i>Information and keywords describing the subject matter of the data</i> |
| 6 | Creation date | <i>The date(s) on which the data was collected or created</i> |
| 7 | Purpose | <i>Reason why the asset was created, intended user communities or source of funding / original project title</i> |
| 8 | Source | <i>The source(s) of the information found in the data asset</i> |
| 9 | Updating frequency | <i>The frequency of updates to this dataset to indicate currency</i> |
| 10 | Type | <i>Description of the technical type of the data asset (e.g., database, photo collection, text corpus, etc.)</i> |
| 11 | Format | <i>Physical formats of data asset, including file format information</i> |
| 12 | Rights and restrictions | <i>Basic indication of the user's rights to view, copy, redistribute or republish all or part of the information held in the data asset. Access restrictions on the data itself or any metadata recording its existence should also be noted</i> |
| 13 | Usage frequency | <i>Estimated frequency of use and if known required speed of retrieval to determine IT infrastructure and storage needs</i> |
| 14 | Relation | <i>Description of relations the data asset has with other data assets and any any DOI ISSN or ISBN references for publications based on this data</i> |
| 15 | Back-up and archiving policy | <i>Number of copies of the data asset that are currently stored, frequency of back-up and archiving procedures</i> |
| 16 | Management to date | <i>History of maintenance and integrity of the data asset e.g. edit rights / security, and any curation or preservation activities performed</i> |

Fig. 2 – Minimum required information for each data asset (Jones et al, 2009)



Planning ahead for data management needs

1. Ensure you have adequate technological resources (e.g. storage space, support staff time)
2. Ensure your data will be robust and free from versioning errors and gaps in documentation
3. Ensure your data is backed up and safe from sudden lost or corruption
4. Ensure you can meet legal and ethical requirements
5. Ensure you are able to share finalized data publicly, if you and/or your funder requires
6. Ensure your data will remain accessible and comprehensible in the near, middle, and distant future.

***Source:** University of Cambridge – Crafting your data management plan: Planning ahead for your data management needs and activities will help ensure that: <https://tinyurl.com/yhpztu3o>.



Key components of a data management plan

(University of Cambridge – simple template)

1. What are the types of research data for your project? (e.g. microscopic images, video recordings, etc.)
2. What is your strategy for organizing your data? (e.g. How do organize your folders and name your files?) (i.e. [TIER Protocol 4.0](#))
3. What is your data backup strategy? (e.g. How frequently do you do your backups? How many independent locations?)
4. How do exchange files (and other information) with your collaborators? (i.e. Open Science Framework ([OSF](#)), [protocols.io](#))
5. **What are your plans for data sharing? Are your plans in-line with your funder's requirements?**
6. Are you working with commercial/sensitive/personal/patentable data? Will you be able to share these data? If not, why?

***Source:** University of Cambridge – Crafting your data management plan: Simple data management plan template. <https://www.data.cam.ac.uk/files/dataplan.docx>



Key components of a data management plan (Digital Curation Centre DMP Checklist)

1. Administrative Data

- ID** – a pertinent ID by the funder and/or institution
- Funder** – State research funder, if relevant
- Grant Reference Number**
[POST-AWARD DMPs ONLY]
- Project Description**
 - What is the nature of your research project?
 - What research questions are you addressing?
 - What purpose are they being collected or created?
- PI / Researcher**
- PI / Researcher ID** (e.g. [ORCID](#))
- Project Data Contact**

2. Data Collection

- What data will you collect or create?**
 - What type, format, and volume of data?
 - Do your chosen formats and software enable sharing and long-term access to the data?
 - Are there existing data that you can use?
- 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 (i.e. [TIER Protocol 4.0](#))?
 - How will you handle versioning?



Key components of a data management plan (Digital Curation Centre DMP Checklist)

3. 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 this documentation and metadata?
 - What metadata standards will you use and why? (i.e. [Seeing Standards: ...](#))

4. Ethics and Legal Compliance

- How will you manage any ethical issues?**
 - Have you gained consent for data preservation and sharing?
 - How will you protect the identity of participants, if required? (e.g. via anonymization, [HHS HIPAA Methods for de-identification](#), IRB)
 - How will sensitive data be handled to ensure it is stored and transferred securely? (i.e. [HiPerGatorRV](#), [REDCap](#))
- How will you manage copyright and Intellectual Property (IPR) issues?**
 - Who owns the data?
 - How will the data be licensed for reuse?
 - Are there any restrictions on the reuse of third-party data? (i.e. [Discipline data publication guides](#))
 - Will data sharing be postponed / restricted?



Key components of a data management plan (Digital Curation Centre DMP Checklist)

5. Storage and Backup

- How will the data be stored and backup during the research?**
 - Do you have sufficient storage or will you need to include charges for additional services? (i.e. HiPerGator)
 - How will data be backed up?
 - Who will be responsible for backup and recovery?
 - How will the data be recovered in the event of an incident?
- How will you manage access and security?**
 - What are the risks to data security and how will these be managed?
 - How will you control access to keep the data secure?
 - How will you ensure that collaborators can access your data securely?
 - If creating or collecting data in the field, how will you ensure its safe transfer into your main secured systems?



Key components of a data management plan (Digital Curation Centre DMP Checklist)

6. Selection and Preservation

- Which data should be retained, shared, and/or preserved?
 - What data must be retained/destroyed for contractual, legal, or regulatory purposes?
 - How will you decide what data to keep?
 - What are the foreseeable research uses for the data?
 - How long will the data be retained and preserved?
- What is the long-term preservation plan for the dataset?
 - Where e.g. in which repository or archive will the data be held?
 - What costs if any will your selected data repository or archive charge?
 - Have you costed in time and effort to prepare the data for sharing / preservation?

7. Data Sharing

- How will you share the data?
 - How will potential users find out about your data?
 - Will you share data via a repository, handle requests directly or use another mechanism
- Are any restrictions on data sharing required?
 - What action will you take to overcome or minimize restrictions?



Key components of a data management plan (Digital Curation Centre DMP Checklist)

8. Responsibilities

- Who will be responsible for data management?**
 - Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?
 - Who will be responsible for each data management activity?
 - How will responsibilities be split across partner sites in collaborative research projects?
- What resources will you require to deliver your plan?
 - Is additional specialist expertise (or training for existing staff) required?
 - Do you require hardware or software which is additional or exceptional to existing institutional provision?
 - Will charges be applied to data repositories?

“Data management practices... should be addressed before any data are collected by taking into consideration four important issues:

- Ownership
- Collection
- Storage, and
- Sharing” – Steneck, 2007



Key components of a data management plan

(Field Trials of Health Interventions:... (Smith et. al, 2015))

1. Introduction to Data Management

2. Before starting to collect data

1. Hardware
2. Software
3. Personnel
4. Data oversight
5. Summary

3. Planning the data flow

1. Database design
2. Data cleaning and integrity
3. Programming issues
4. Standard operating procedures
5. Version control
6. Confidentiality
7. Training
8. Pilot testing and database testing

4. Data collection systems

1. Questionnaires
2. Electronic data capture
3. Laboratory data
4. Clinical data
5. Longitudinal data collection
6. Quality control
7. Future trends

5. Managing data

1. Data entry
2. Data checks
3. Data cleaning
4. Variable naming and coding
5. Data lock



Key components of a data management plan (Field Trials of Health Interventions:... (Smith et. al, 2015))

6. Archiving

1. Interim backups
2. Metadata
3. Data sharing policy
4. Archiving hard copies

7. Preparing for analysis

1. Data dictionary
2. Creating new variables
3. Coding and recoding
4. Merging and linking data

The Research Data Management Lifecycle

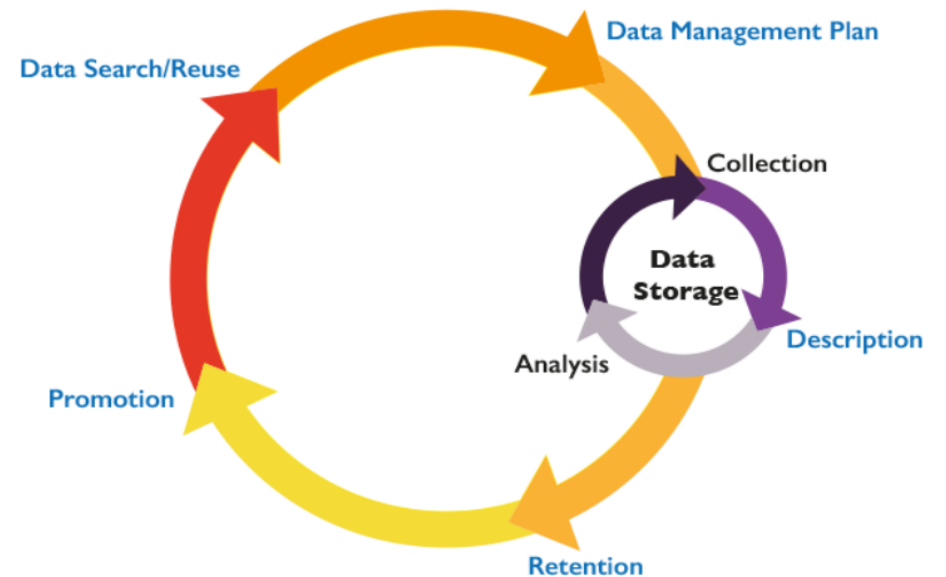


Figure 1.0 – Research Data Management Lifecycle

Fig. 3 – University of New South Wales Research Data Governance & Materials Handling Policy, Research Data Management Lifecycle, Appendix 1 (UNSW, 2019)

Key stakeholders involved

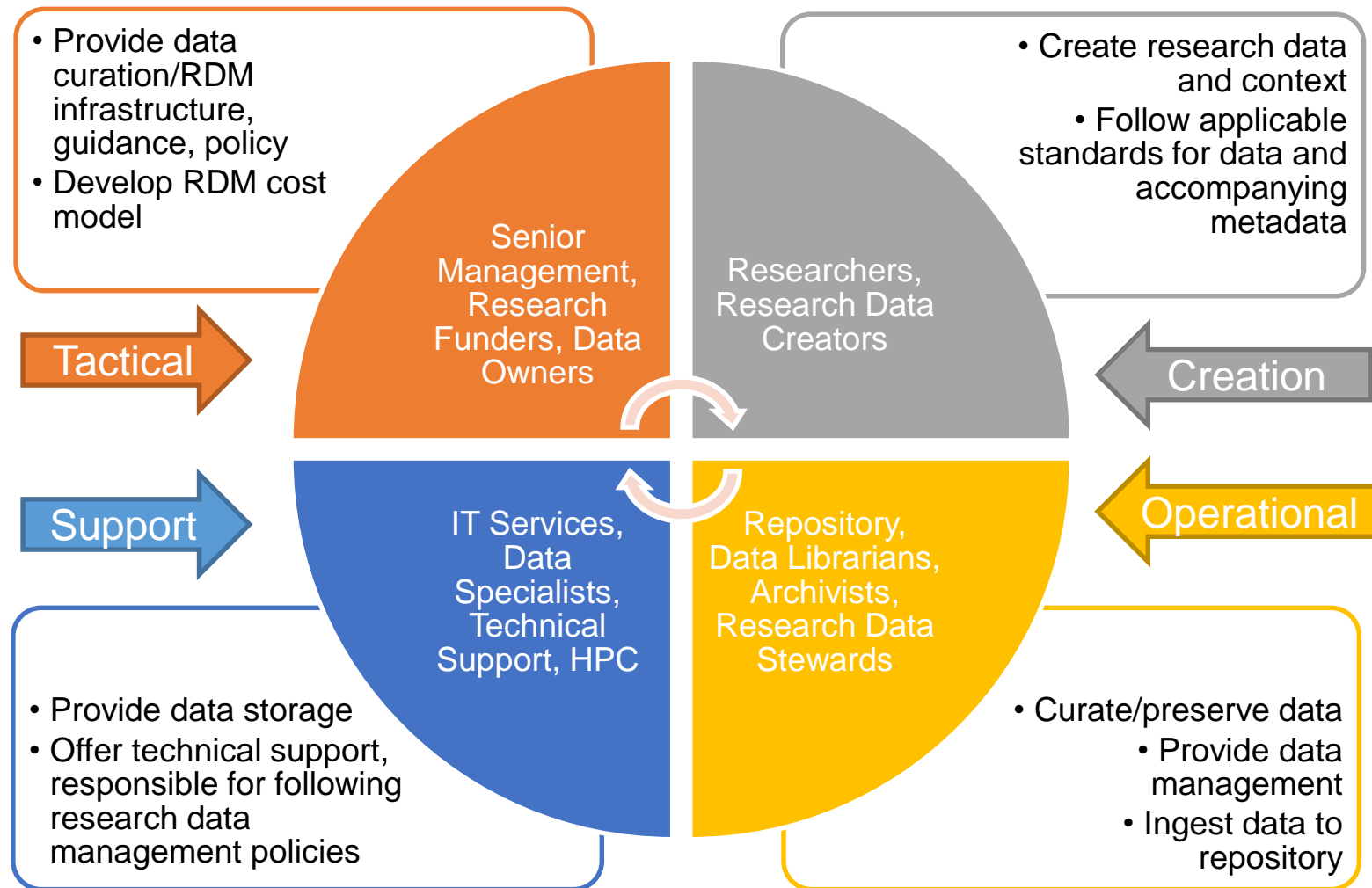


Fig. 4 – RDM Roles and Responsibilities infographic

Key stakeholders involved (cont.)

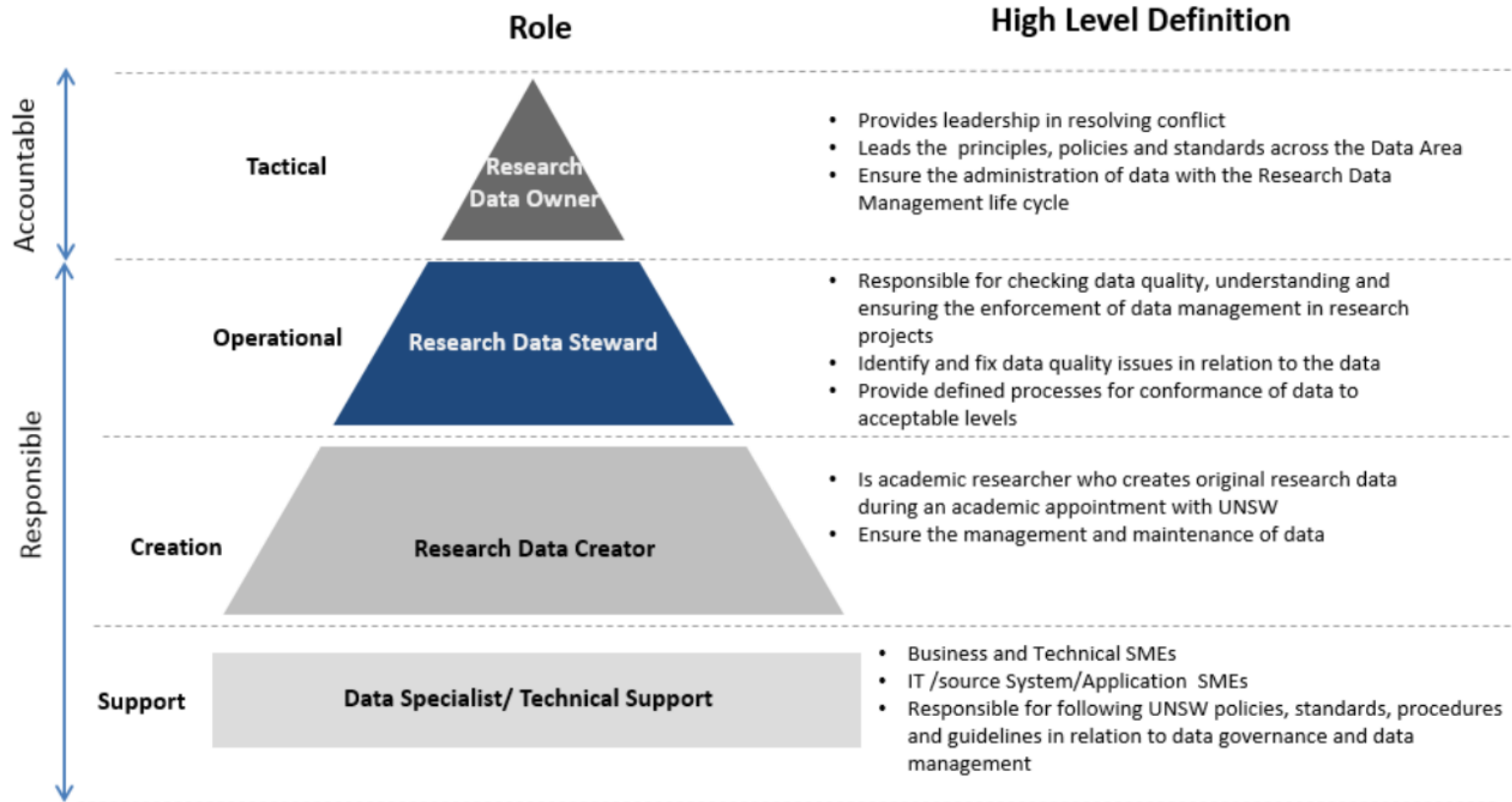


Fig. 5 – Research Data Governance Roles and Responsibilities (UNSW, 2019)

Some UF infrastructure and resources



HiPerGator RV

- Restricted Data
 - [NIST-800-53](#) Moderate
 - [NIST-800-171](#)
 - [FISMA](#) (Federal Information Security Modernization Act)



HiPerGator Storage

- Blue Storage
 - Parallel filesystem
 - High-performance
 - \$625/TB for 5-years
- Orange Storage
 - Parallel filesystem
 - Lower performance
 - \$125/TB for 5-years

Data Repository



UF Clinical and Translational Science Institute

- Data, Informatics and IT
 - REDCap
 - Integrated Data Repository
 - Informatics Consulting

Fig. 6 – University of Florida Research Computing and University of Florida Clinical and Translational Science Institute resources



Examples of DMPs from successful awards by UF Researchers

1. American Brain Tumor Association (ABTA) (2021) (offline)
 2. University of Florida, Institute of Food and Agricultural Sciences (UF/IFAS) in cooperation with Florida Fish and Wildlife Research Institute and Texas Parks & Wildlife Department - <https://zenodo.org/record/5500391#.YTt7pY5KiUk> (2017)
- UF Research Data Management Plan (DMP) examples repository (pilot) <https://sandbox.zenodo.org/communities/uf-research-dmps/?page=1&size=20>

DMPTool – Build your Data Management Plan

Learn Sign in Language

DMPTool
Build your Data Management Plan




Notice: Signed out successfully.

Welcome to the DMPTool

Create data management plans that meet institutional and funder requirements.

Get started

DMPTool by the Numbers

| | | |
|--|--|--|
|  64,133 Users |  62,142 Plans More |  312 Participating Institutions More |
|--|--|--|

Top Templates

- Digital Curation Centre
- Template USP - Baseado no DCC
- Template USP - Mínimo
- Digital Curation Centre (português)
- NSF-SBE: Social, Behavioral, Economic Sciences

[More](#)

DMPTool News

DMP IDs and the DMPTool: Announcing DMPTool v. 3.1

[Go to the blog](#)

Fig. 7 – DMPTool Build your Data Management Plan - <https://dmptool.org/>



References

- Bush, V. (1945). As We May Think. The Atlantic. <https://tinyurl.com/y67tnv2n>.
- 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>.
- DCC. (2021). Example DMPs and guidance. <https://tinyurl.com/yjalz6og>.
- DMPTool. (2021). Data Management Planning tool. <https://dmptool.org/>.
- DCMI. (2021). Dublin Core Metadata Initiative. Metadata Basics. <https://www.dublincore.org/resources/metadata-basics/>.
- Erway, R. (2013). Starting the Conversation: University-wide Research Data Management Policy. <https://tinyurl.com/tjzlrk8>.
- Jones, S., Ross, S., & Ruusalepp, R. (2009). Data Audit Framework Methodology. HATII, University of Glasgow. https://www.data-audit.eu/DAF_Methodology.pdf.
- Library of Congress. (2017). Sustainability of Digital Formats: Planning for Library of Congress Collections. Formats, Evaluation Factors, and Relationships. <https://tinyurl.com/yfryn259>.
- Library of Congress. (2019). Sustainability of Digital Formats: Planning for Library of Congress Collections. Format Descriptions. <https://tinyurl.com/yzsvapdl>.
- Lowndes, J., Best, B., Scarborough, C. et al. Our path to better science in less time using open data science tools. Nat Ecol Evol 1, 0160 (2017). <https://doi.org/10.1038/s41559-017-0160>.



References

- Office of Management and Budget (OMB). (1999). Circular A-110 Revised 11/19/93 As Further Amended 9/30/99. https://obamawhitehouse.archives.gov/omb/circulars_a110/.
- Smith, P. G., Morrow, R. H., & Ross, D. A. (2015). Field Trials of Health Interventions: A Toolbox (3 ed.). Chapter 20. Data Management. <https://tinyurl.com/yhjryt8d>.
- Steneck, N. H. (2007). ORI Introduction to the Responsible Conduct for Research. Chapter 6. Data Management Practices. <https://tinyurl.com/ydmpw949>.
- UF IRM. (2021). Fast Path Solutions. <https://irm.ufl.edu/fast-path-solutions/>.
- UF IRM. (2021). UF Data Guide, <https://irm.ufl.edu/uf-data-guide/>.
- UF Research. (2021). UF Research Navigating the Research Lifecycle. <https://research.ufl.edu/research-lifecycle.html>.
- UFRC. (2021). UF Research Computing. <https://www.rc.ufl.edu/>.
- UNSW. (2019). University of New South Wales. Research Data Governance & Materials Handling Policy. <https://tinyurl.com/yjfn4z9q>.
- USGS. (2021). Data Sharing Agreements. <https://tinyurl.com/rdxjspe>.
- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.



Thank you

Who can help you with data management planning at the University of Florida?

1. UF Office of Research
2. Your department or college IT staff
3. UF Academic Research Consulting & Services - <https://arcs.uflib.ufl.edu/>
4. Subject and departmental librarians – Subject/Area Specialists - <https://uflib.ufl.edu/specialists/>
5. Your funder

Contact information

UF Libraries Data Management Librarian

plato.smith@ufl.edu