Selecting a Data Repository

Presenter: Plato Smith, Ph.D., Data Management Librarian

Date: October 12, 2021

Time: 2:30 pm – 3:30 pm

Location: Zoom

UF | George A. Smathers Libraries
UNIVERSITY of FLORIDA

Context

"Data should be submitted to discipline-specific, community-recognized repositories where possible. Where a suitable discipline-specific resource does not exist, data should be submitted to a generalist repository."
– Scientific Data (2021)

Table of Content

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

- **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

- **Open data** – “is publicly available data that is structured in a way that enables the data to be fully discoverable and usable by end users. It is public, accessible, fully documented, reusable, complete, timely, and updated or managed following release.” – USAID FAQ, n.d.

Definitions

- **FAIR Data Principles** are 15 elements across the four categories of **Findable (4)**, **Accessible (4)**, **Interoperable (3)**, and **Reusable (4) (FAIR)** to facilitate knowledge discovery by assisting humans and machines in discovering and accessing integrations and analysis of scientific data, associated algorithms, and workflows (FORCE11, 2020).

  - Launched at a Lorentz workshop [Jointly designing a data FAIRPORT] in The Netherlands in 2014
  - FAIR principles were published in 2016
  - FAIR Principles: https://www.go-fair.org/fair-principles/
  - The FAIR Guiding Principles for scientific data management and stewardship - *Sci Data* article (Wilkinson et. al, 2016)

Definitions

- "A **data repository** is a centralized place to store and maintain data. A repository can consist of one or more databases of files which can be distributes over a network. Data repositories are often managed by data curation personnel who ensure that files are managed and preserved for the long-term." – United States Geological Survey (USGS) (n.d.)

Source: USGS Data Management Repositories -
https://www.usgs.gov/products/data-and-tools/data-management/repositories

Definitions

- **Institutional repository** – The Institutional Repository at the University of Florida (IR@UF) brings together the scholarly, research, and creative works of our academic community.

- **General repository** – General repositories can be helpful for researchers looking to collaborate among institutions, or who want to maintain their scholarship independently (and without an institutional affiliation).

- **Subject-specific repository** – Subject-specific repositories are an option if you want your work to be discoverable by a particular discipline.

Source: UF Digital Partnerships & Strategies -
https://digitalpartnerships.uflib.ufl.edu/repository-publishing/repositories/

What are data submission guidelines?

- **Data submission guidelines – key components**
  1. Review data for quality
  2. Codebook
  3. Use open file formats, whenever possible (i.e. AVIF, CSV, HTML, JPEG, JSON, Markdown, NetCDF) See: List of open formats
  4. Organize files
  5. Write high quality metadata
  6. Documentation

- **Ex 1. The Knowledge Network for Biocomplexity** (KNB) is an international repository intended to facilitate ecological and environmental research.
- **Ex 2. USAID Development Data Library** (DDL) is your gateway to USAID-funded, machine readable data.
- **Ex 3. National Centers for Environmental Information** (NCEI) is the official archive for data collected by NOAA scientists, observing systems, and research initiatives.

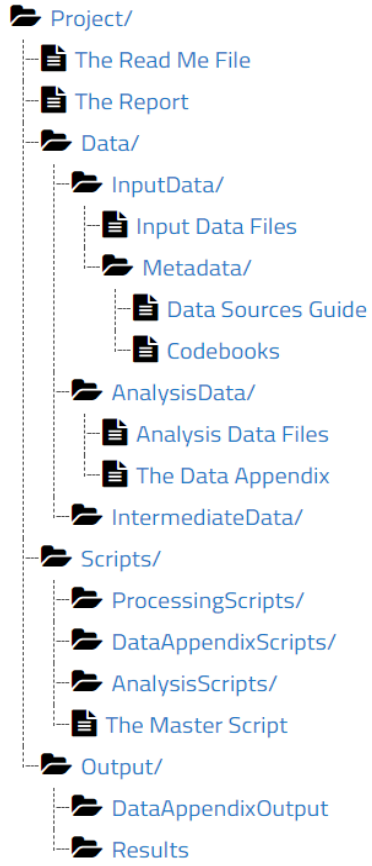

Fig. 1 - TIER Protocol 4.0 (2021)

What are the five criteria for Yes/No response for NIH Open Domain-Specific Data Sharing Repositories?

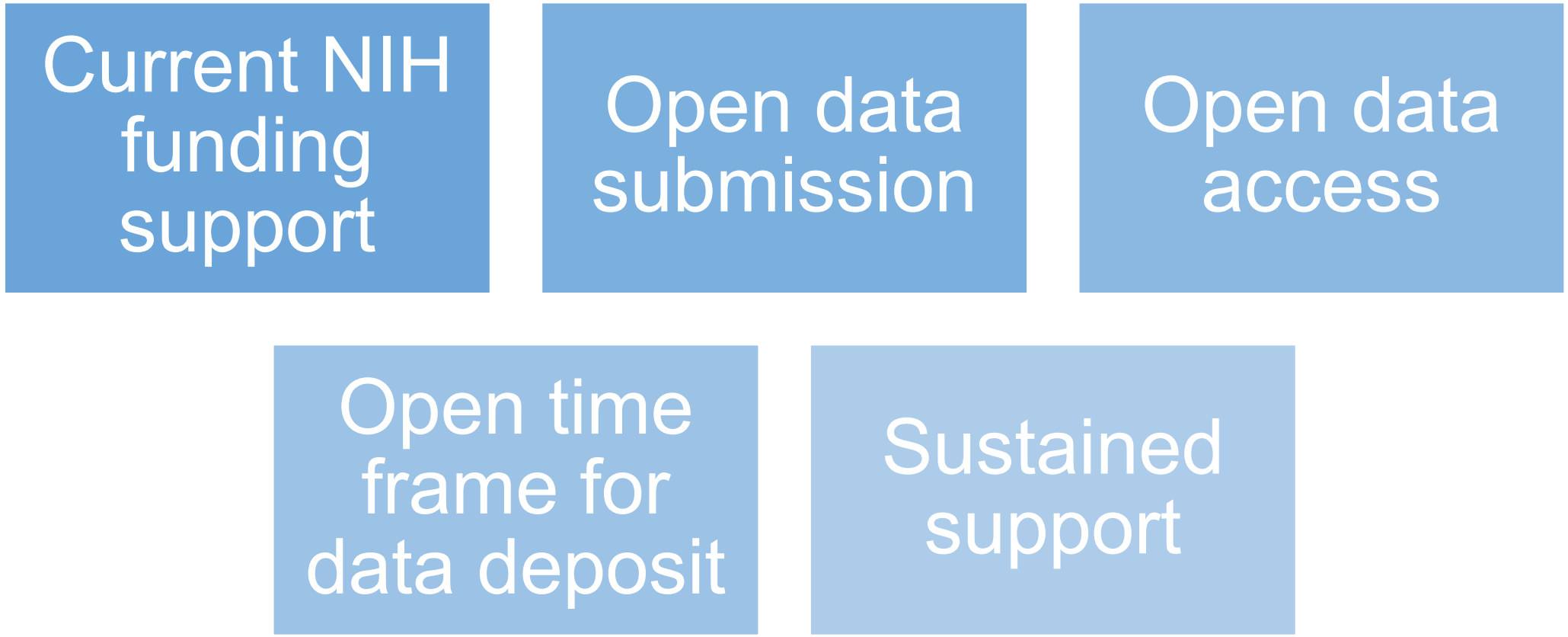

Fig. 2 – NIH Trans-NIH Biomedical Informatics Coordinating Committee (BMIC) five criteria for NIH data sharing repositories (NIH BMIC, 2018)

NIH Trans-NIH Biomedical Informatics Coordinating Committee (BMIC) Data Sharing Resources

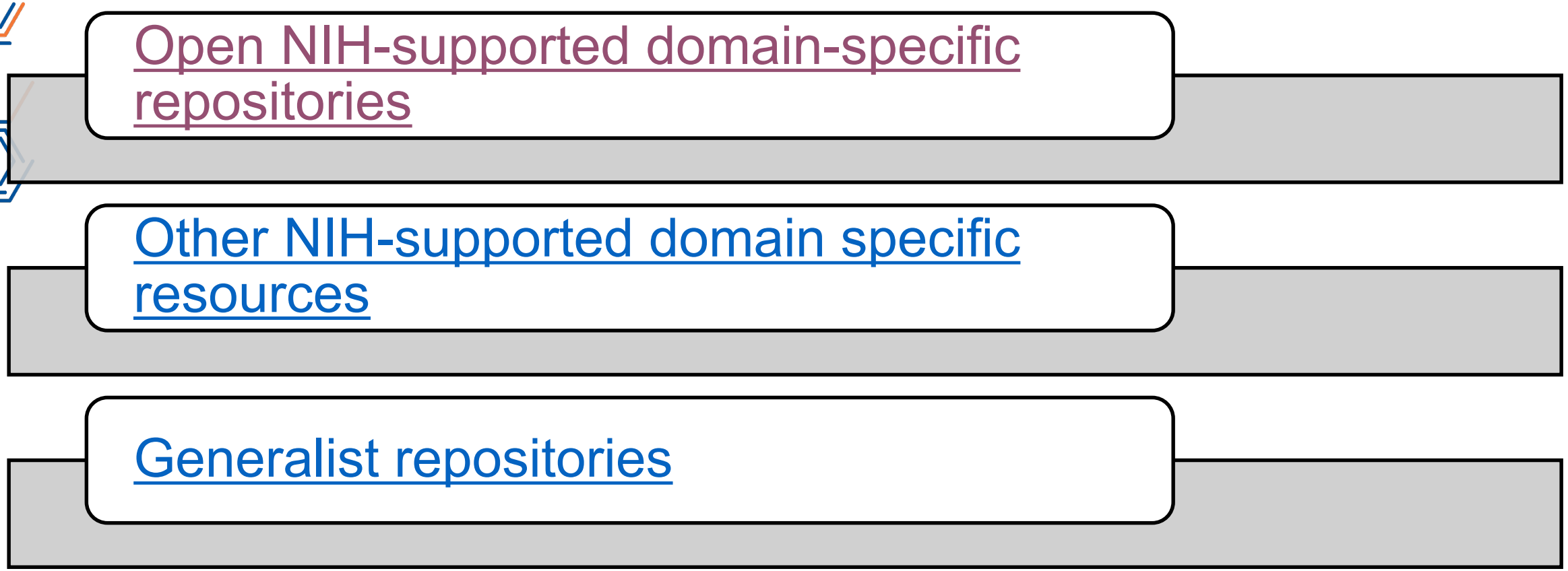

Fig. 3 – NIH BMIC Data Sharing Resources (NIH BMIC, 2020)

How to search for discipline-specific data repositories?

Institute, Center, Funder recommended (i.e. ABTA, NIH BMIC, USGS)

NIH-supported domain-specific repositories

re3data; FAIRsharing

Fig. 4 – Examples of searching for discipline-specific data repositories

How to search for discipline-specific data repositories?

Consult

- Identify a data repository and develop a DMP requests
  - ❑UF College of Medicine Asst. Professor – Div. of Infectious Diseases and Global Medicine (NIH)
  - ❑UF College of Medicine Asst. Professor – Div. of Pediatric Hematology Oncology (ABTA)

Resources

- ❑Open NIH-supported domain-specific →ICO: NIAID: Repository Name →TB Portals
- ❑ABTA Approved Repositories → NCBI BioProject

Fig. 5 – UF College of Medicine Asst. Prof. find a data repository use cases

When to select a general data repository?

Discipline-specific repository not available or feasible

General data repository can serve as backup

Archive code, analyses, or supplemental data

Fig. 6 – Some reasons to select a general data repository (See: Generalist Repository Comparison Chart [Stall et .al, 2020])

What are two general data repositories approved by UF IRM FPS for open data?

UF IRM FPS

- University of Florida Integrated Risk Management (IRM) Fast Path Solutions (FPS)
- Pre-assessed software and computing environments

Figshare

- Approved on 2021-06-07
- 2nd General data repository
- 100 GB free per Scientific Data manuscript
- Additional fees for larger datasets

Zenodo

- Approved 2020-04-29
- 1st General data repository
- 50 GB per dataset
- DOI
- Version control
- OAI-PMH feature
- Plan S Principles

Fig. 7 – General repositories approved by UF IRM FPS

Have can an ERN solution compliment a data repository?

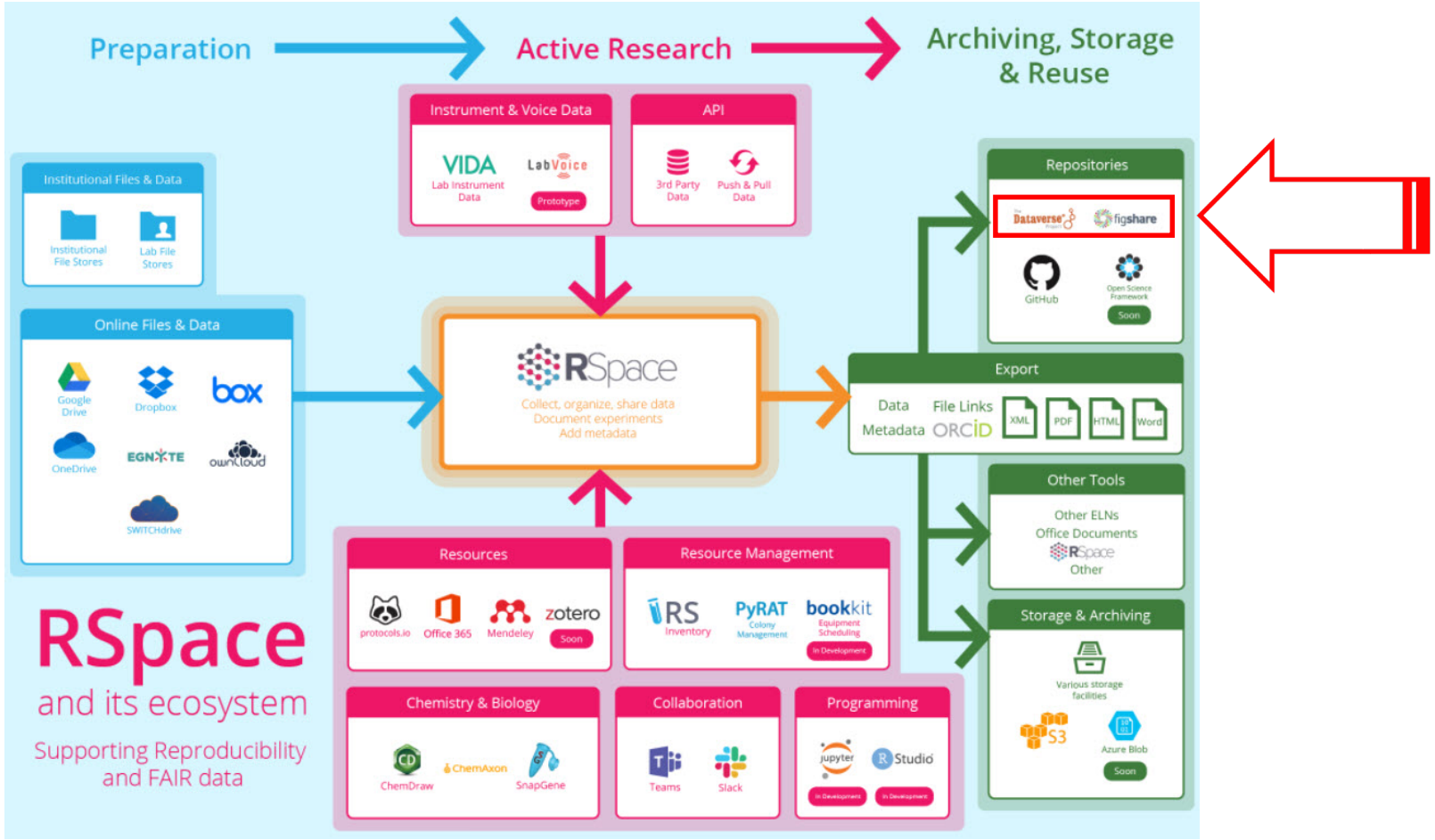

Fig. 8 – RSpace electronic research notebook (ERN) ecosystem - https://www.researchspace.com/#case-studies

What is a trusted digital repository?

Criteria for a United State Geological Survey (USGS) Trusted Digital Repository (TDR)

A USGS TDR must:

1. Accept responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of users;
2. Be an organizational system that supports not only the long-term viability of the repository but also the digital information for which it has responsibility;
3. Demonstrate fiscal responsibility and sustainability;
4. Be designed in accordance with commonly accepted system conventions and standards to ensure the ongoing management, access, and security of materials deposited within it; and
5. Establish methodologies for system evaluation that meet community expectations of trustworthiness.

Source: USGS Trusted Digital Repositories (TDR) - https://tinyurl.com/ygouqh8p

How to upload and share data using the Zenodo general data repository sandbox?

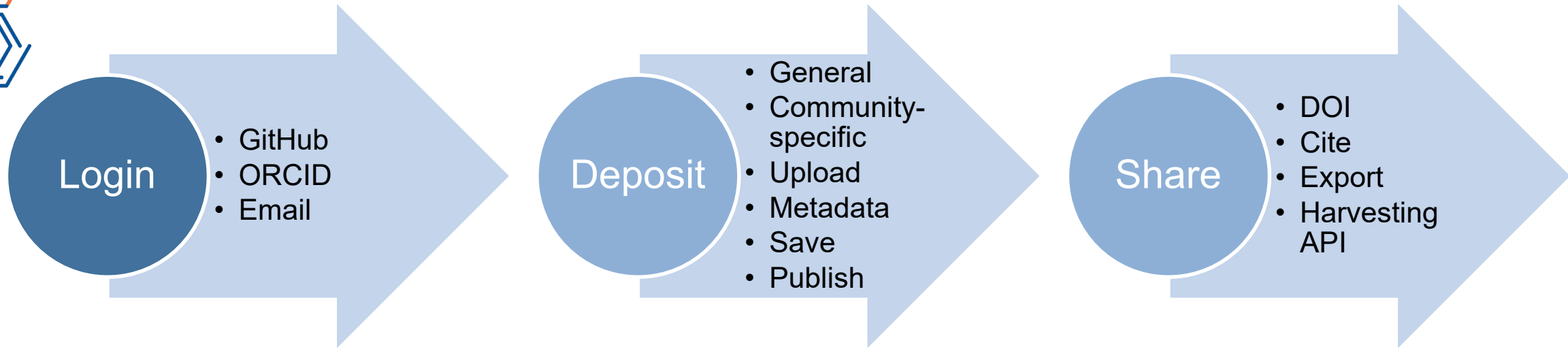

Fig. 9 – Demonstrate deposit and share via Zenodo sandbox - https://sandbox.zenodo.org/

UF George A. Smathers Libraries
UNIVERSITY of FLORIDA

References

- DCMI. (2021). Dublin Core Metadata Initiative. Metadata Basics. https://www.dublincore.org/resources/metadata-basics/.
- FAIRsharing. (2021). https://fairsharing.org/.
- Knowledge Network for Biocomplexity (KNB). (nd). About the KNB. https://knb.ecoinformatics.org/about.
- Library of Congress. (2017). Sustainability of Digital Formats: Planning for Library of Congress Collections. Formats, Evaluation Factors, and Relationships. https://tinyurl.com/yfryn259.
- National Institutes of Health (NIH).(2018). Trans-NIH Biomedical Informatics Coordinating Committee (BMIC) Data Sharing Resources. The five criteria for Yes/No response used in the query to repositories. https://www.nlm.nih.gov/NIHbmic/query_criteria.html.
- National Institutes of Health (NIH).(2020). Trans-NIH Biomedical Informatics Coordinating Committee (BMIC) Data Sharing Resources. https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html.
- re3data. (2021). Registry of Research Data Repositories. https://www.re3data.org/.
- Scientific Data. (2021). Data Repository Guidance. https://www.nature.com/sdata/policies/repositories.
- TIER. (2021). TIER Protocol 4.0. https://www.projecttier.org/tier-protocol/protocol-4-0/.
- UF IRM. (2021). Fast Path Solutions. https://irm.ufl.edu/fast-path-solutions/.
- USAID. (n.d.). Welcome to the Development Data Library User Guide. For Data Submitters. Preparing Data for Submission. https://data.usaid.gov/stories/s/Preparing-Data-for-Submission/2aex-zbcs.
- USAID. (n.d.). Frequently Asked Questions (FAQ) about Open Data. https://data.usaid.gov/stories/s/7nq9-vptc#1.-what-is-open-data.
- USDA. (n.d.). Find a Data Repository. https://www.nal.usda.gov/main/data/find-data-repository.
- Zenodo sandbox. (2021). Zendo sandbox. https://sandbox.zenodo.org/.

Thank you

Who can help you with selecting a data repository at the University of Florida?

1. Your funder
2. Your department
3. Subject and departmental librarians – Subject/Area Specialists - https://uflib.ufl.edu/specialists/
4. UF Academic Research Consulting & Services - https://arcs.uflib.ufl.edu/

Contact information

UF Libraries Data Management Librarian

plato.smith@ufl.edu

UF George A. Smathers Libraries
UNIVERSITY of FLORIDA