

Data Lifecycle Technologies at Oak Ridge Leadership Computing Facility

National Science Data Fabric
All-Hands Meeting 2024

Presenter: Olga A. Kuchar, Ph.D.

ORNL is managed by UT-Battelle LLC for the US Department of Energy



U.S. DEPARTMENT OF
ENERGY

At the heart of scientific exploration is data

Creating resource ecosystems...

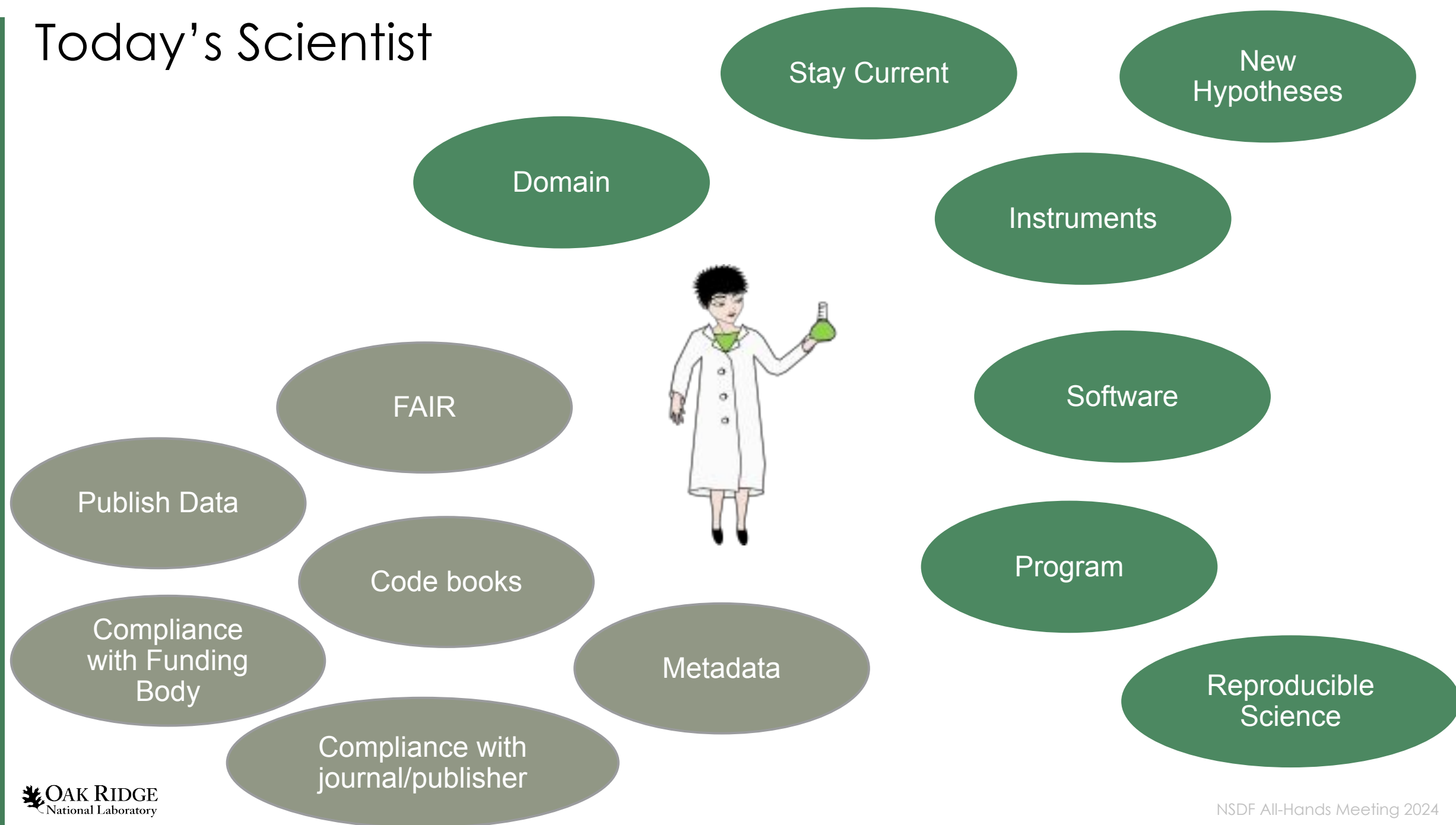


Maximize:

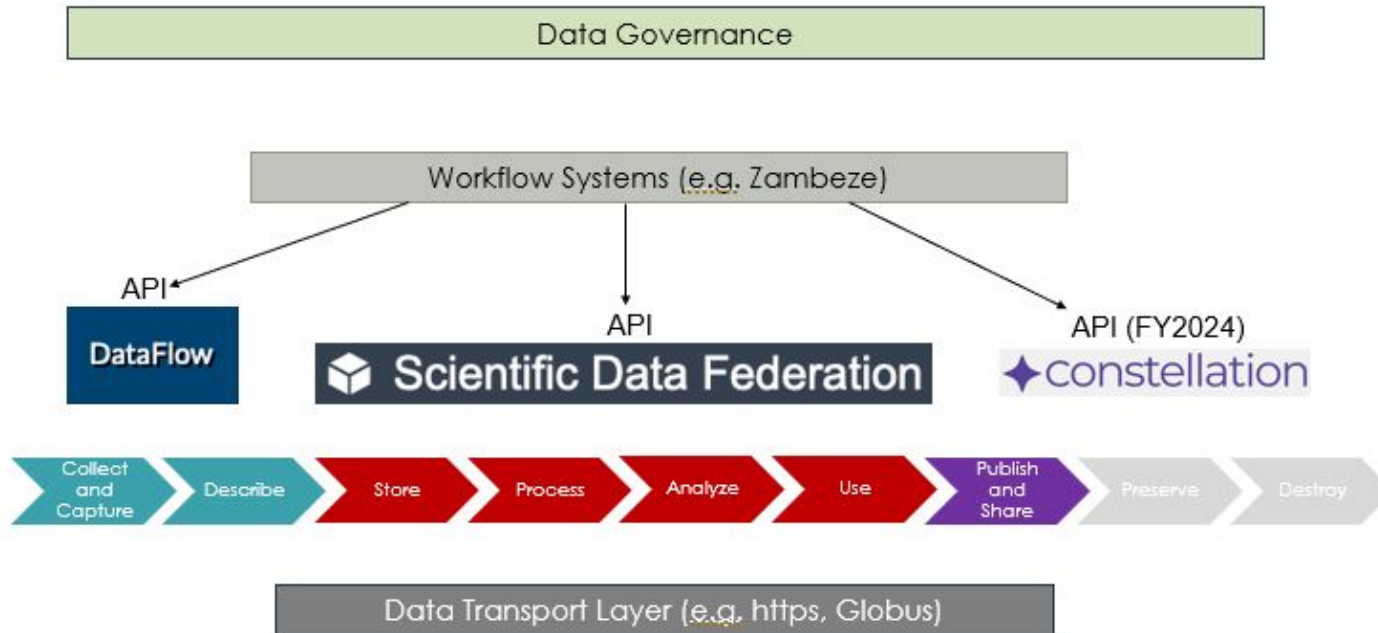
- Infrastructure, resources, and software
- Interface contracts
- Trust and interoperability
- Sharing responsibility and outcomes

While lowering the cognitive load...

Today's Scientist



Data Lifecycle Technologies Group



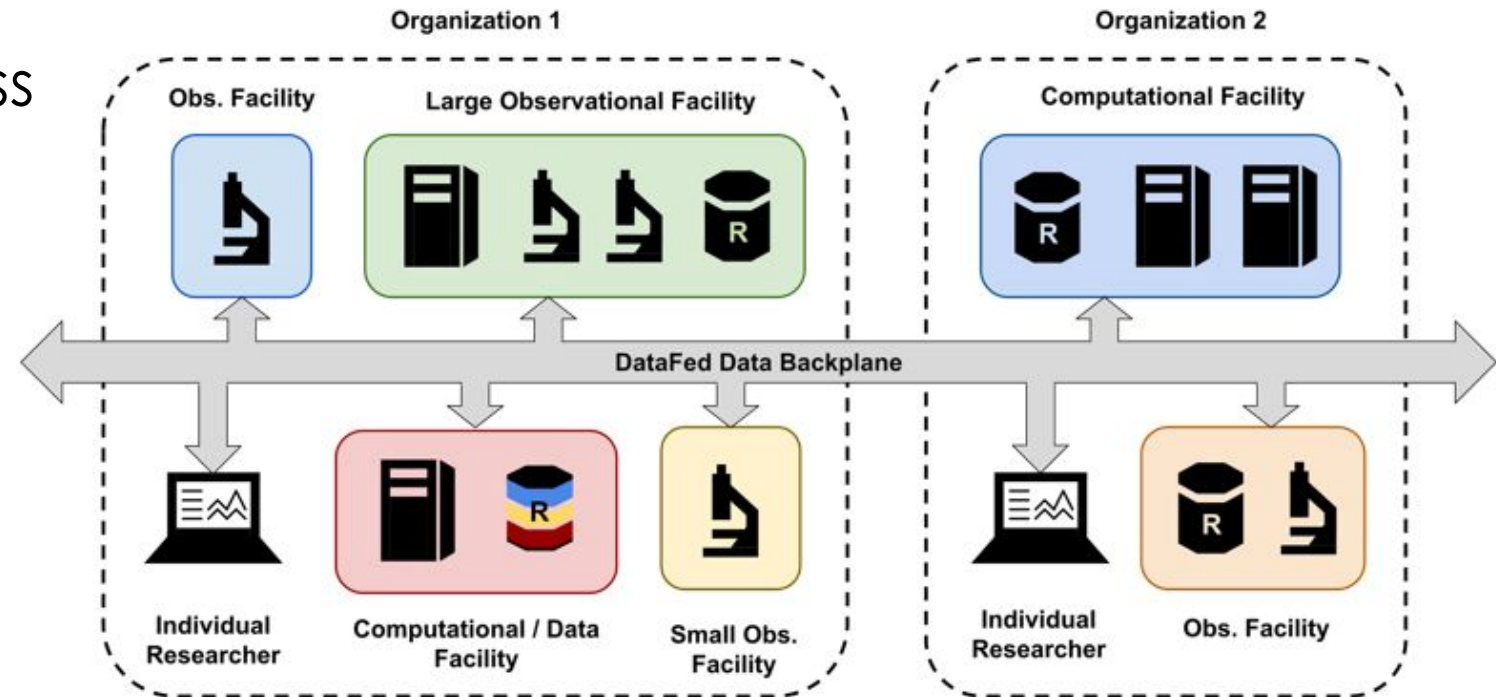
- Team of research and technical professionals
- Empower scientific communities and forge strong partnerships
- Holistic data management tools
- Data governance

DataFed – Federated Data Repository for Science

(<https://datafed.ornl.gov>)



- Federated data storage
- Simple, uniform data access from anywhere
- Federated ID and high-performance data transfer
- FAIR principled
- Captures domain-specific metadata and provenance
- Powerful search and discovery
- Expanding users via NW-BRaVE project, Lehigh University, Drexel University



DataFed – Web, API, Data Lineage/Provenance

Collection

Record with
Raw Data

Metadata

Location of raw
data

The screenshot shows the DataFed web interface. On the left, a sidebar displays a tree view of collections: 'Personal Data', 'Root Collection', 'Experiment1', and 'MolecularDynamics'. Under 'Experiment1', there are four records: 'First Experiment', 'Fourth experiment', 'Second Experiment', and 'Third Experiment'. The 'First Experiment' record is selected, and its details are shown in the main panel. The details include a 'General' tab with tags, schema, data location, data source, data size, data extension, relationships, and owner. The 'Metadata' tab is also visible. The interface includes a top navigation bar with 'My Data', 'Catalog', and 'Provenance' tabs. At the bottom, there are buttons for 'Provenance', 'Annotate', 'Upload', 'Download', 'Edit', 'Permissions', 'Delete', and 'Schemas'. A notice at the bottom states: 'DataFed is for open data, fundamental research and/or open-source applications and does not meet requirements for sensitive, proprietary or other controlled unclassified information. DataFed servers will be off-line for regular maintenance every Sunday night from 11:45 pm until 12:15 am EST Monday.'

The screenshot shows a terminal window with two panels. The left panel displays a list of datasets with their IDs, names, and descriptions. The right panel shows a Python script that uses the DataFed API to interact with the datasets. The script includes the following code:

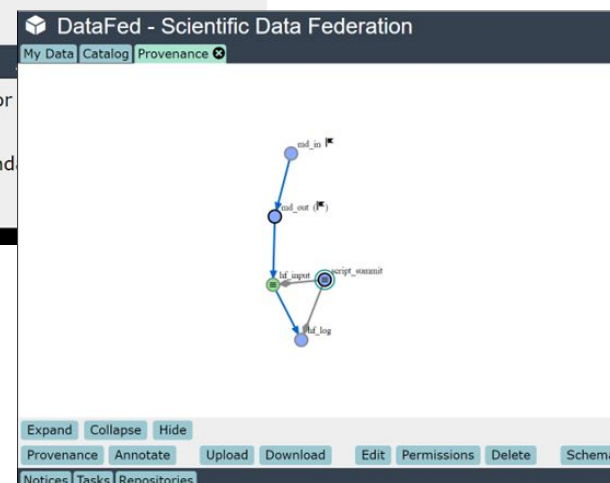
```
>ls
1. c/53987481 (demo) "Demo"
2. c/51962668 (share) "Share"
3. c/53986889 (test) "Test"
>ls demo
1. d/67781402 (cal-titan-018119) "Calibration Titan"
2. d/67781826 (config-titan-123) "Config Titan 123"
3. d/67776565 (data-a) "Dataset A"
4. d/67776684 (data-b) "Dataset B"
5. d/67776778 (data-b-1) "Dataset B-1"
6. d/67777979 (data-e) "Dataset E"
7. d/67779849 (data-f) "Dataset F"
8. d/67776773 (data-c) "Dataset-C"
9. d/67776892 (data-d) "Dataset-D"
10. d/67780272 (exp-1) "Experiment 1"
>dv 3.
DataID d/67776565 (data-a)
Title Dataset A
Desc An interesting scientific dataset
Topic n/a
Keywords data science, machine learning
Owner u/stansberrydv
Locked No
Size 0
Repo repo/core
Created 2019-04-03 16:28:00
Updated 2019-04-03 16:42:30
Meta (available)
Refs Child of d/67780272 (exp-1)
Source of d/67776773 (data-c)
>^C
[3ub0or-condo-login01 ~]$ exit
logout
Connection to or-condo-login01.ornl.gov closed.
(base) persimmon:~ 3ub$
```

```
Python 3.5.5 | packaged by conda-forge | (default, Jul 23 201
#, 23:45:11)
Type 'copyright', 'credits' or 'license' for more information
Python 7.8.0 -- An enhanced Interactive Python. Type '?' for
help.
In [1]: from datafed.CommandLib import API
In [2]: df = API()
In [3]: df.setContext('p/mat001')
In [4]: df.projectView('p/mat001')
Out[4]:
{proj: {
  id: "p/mat001"
  title: "MAT001 - BELINE Mining Pilot"
  desc: "Collecting, standardizing, and then analyzing Band
Excitation Imaging data from Scanning Probe Microscopes for s
cientific and operational insights"
  owner: "u/somnaths"
  repo: "cades-cnms"
  data_limit: 50955800320
  data_size: 58674891797
```

Command-line & Python interfaces

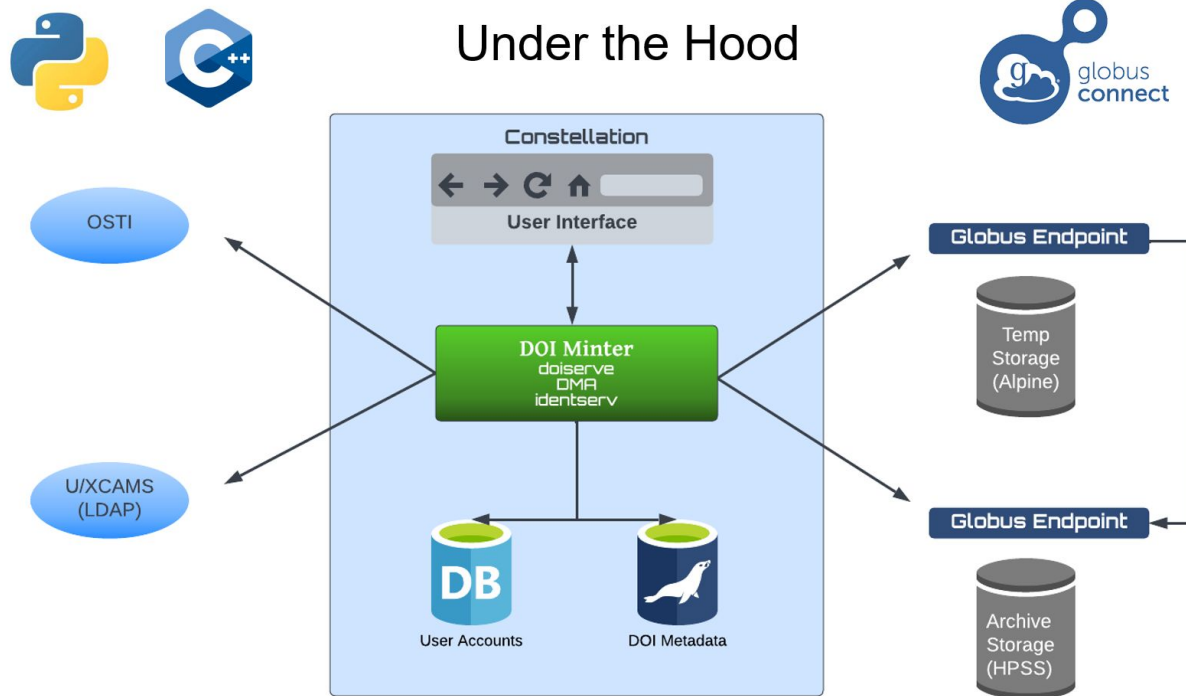
Interactive & non-interactive scripting
(e.g. - HPC environments)

Provenance



Constellation – Generalist Open Data Repository

(<https://doi.ccs.ornl.gov>)



- ~7 PB of unstructured data
 - Largest dataset: 4 PB
 - Largest file: ~17 TB
- Curated data helps science
 - Challenges of large-scale data curation
 - Data for/with AI
- Working with privacy, legal, cyber security, and Institutional Review Board (IRB) to keep the scientists, lab, and DOE protected

DOE Data Curation Working Group



2023



30+ from across the national laboratories



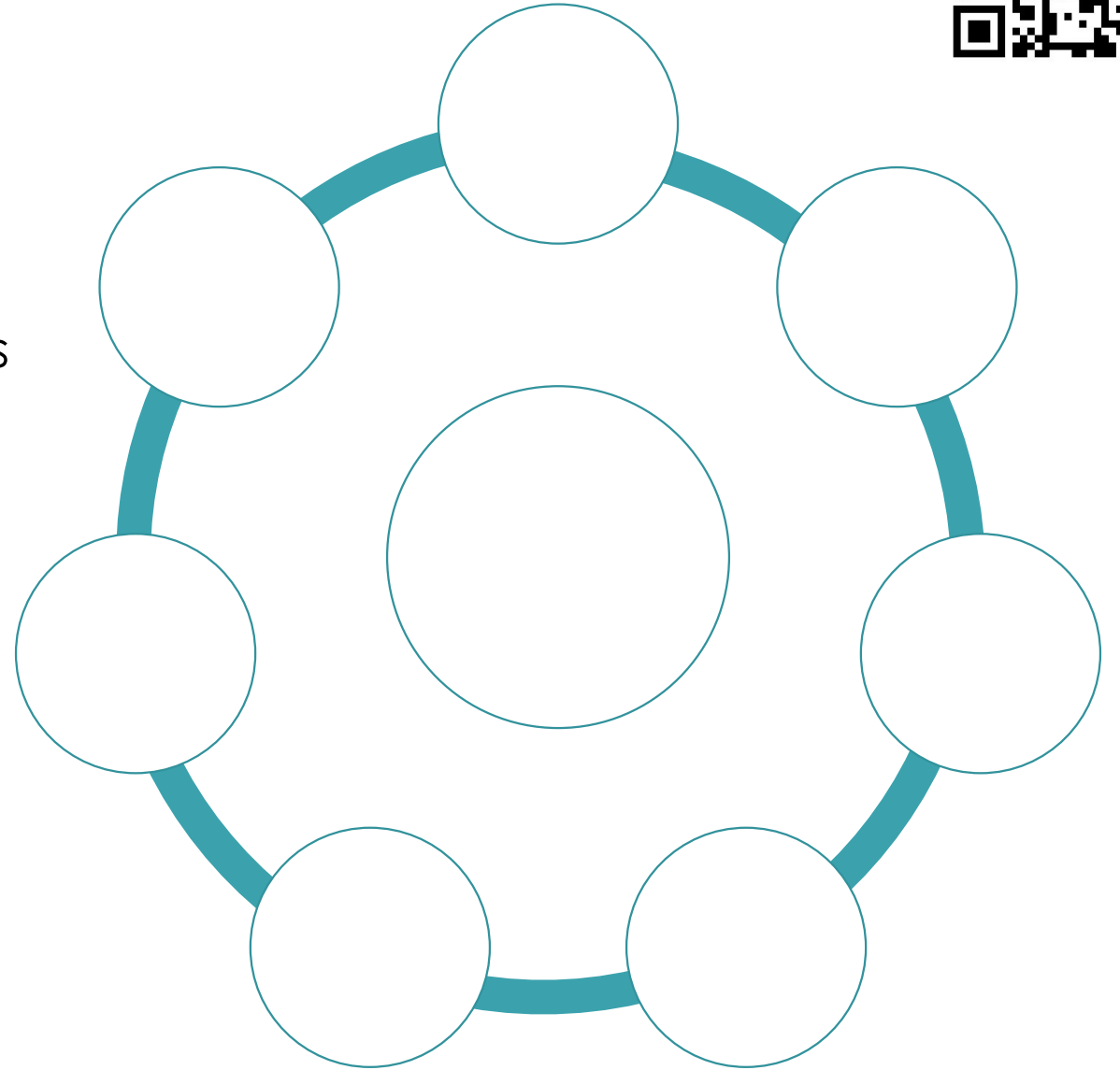
Larger group meets monthly; objective groups meet on an as-needed basis to advance work on their deliverables



Microsoft Teams



Membership is open to any interested individual across the DOE National Laboratory Complex



What does success look like for us?



This Photo by Unknown Author is licensed under CC BY-SA-NC

- Collaborations and partnerships that evolve current data management practices
- Our tools being used by the scientific community and integrated into data ecosystems
- Researching/prototyping/deploying data solutions to meet the next generation of scientific challenges

Thank you!



Olga Anna Kuchar, Ph.D.
kucharoa@ornl.gov



Acknowledgement:

This work uses resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.