OAK RIDGE National Laboratory

Towards a DOE Metadata Schema for Generalist Open Data Repositories

Meghan Berry, Information Science Specialist, ORNL

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

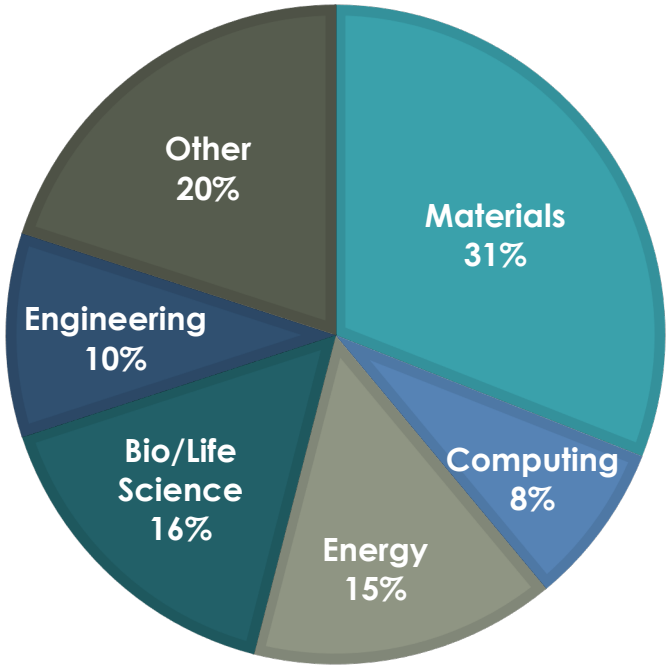
U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

Constellation

- Generalist open data repository hosted by OLCF
  - Available to all OLCF facility users and ORNL researchers
- Currently ~3PB published data
  - ~8PB in staging
  - Largest file ~17TB
- DOIs issued via OSTI Data ID Service
- Curation services to meet FAIR principles

CONTENTS BY DOMAIN



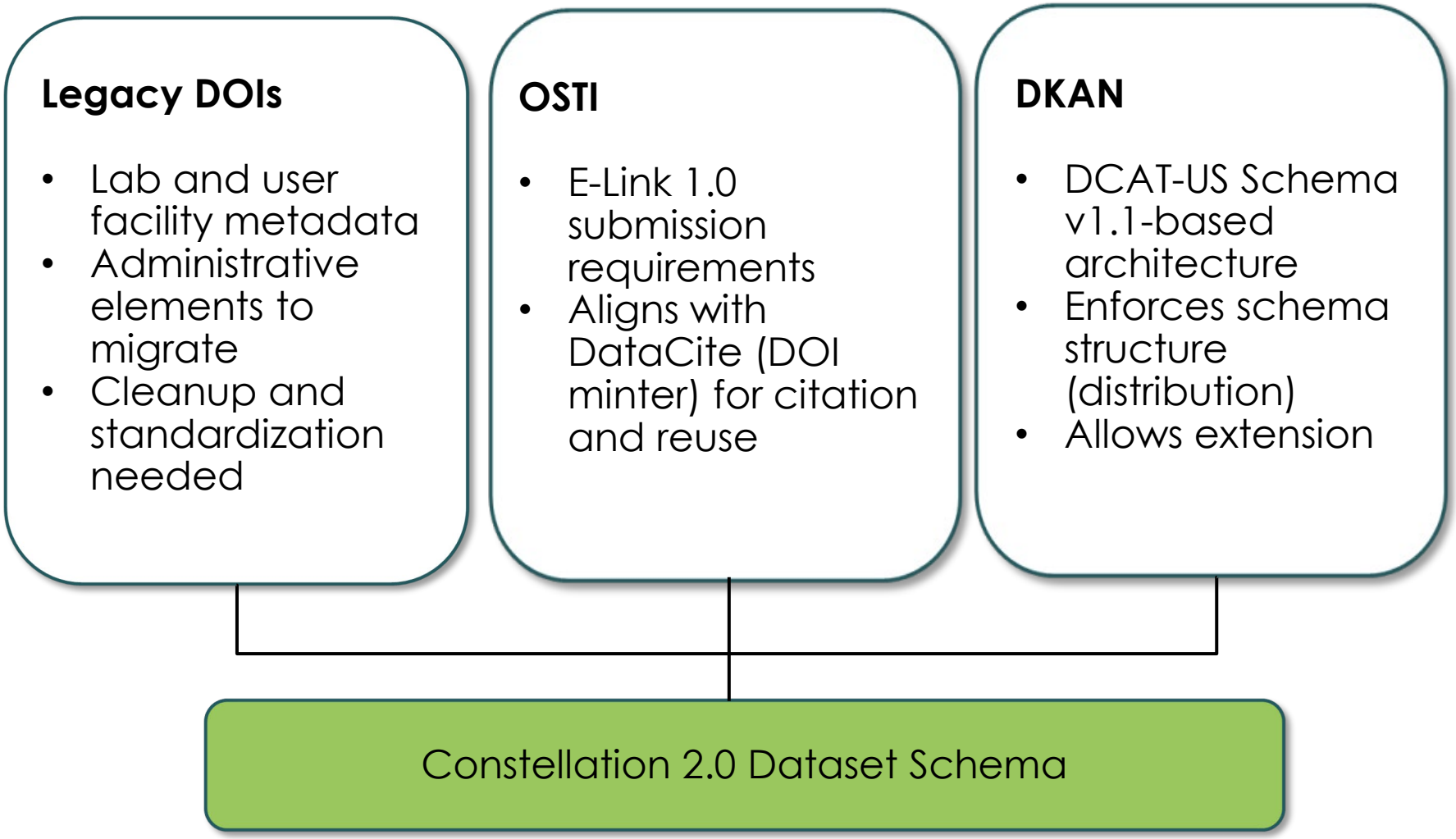
Repository Migration

- FY24 development and migration from homegrown platform to DKAN (Drupal 10)
- Phased record migration
  - Previously published content with existing schema
  - Deposits accepted throughout development
  - Three schema iterations (so far)
- Iterative metadata cleanup for validation and ingest

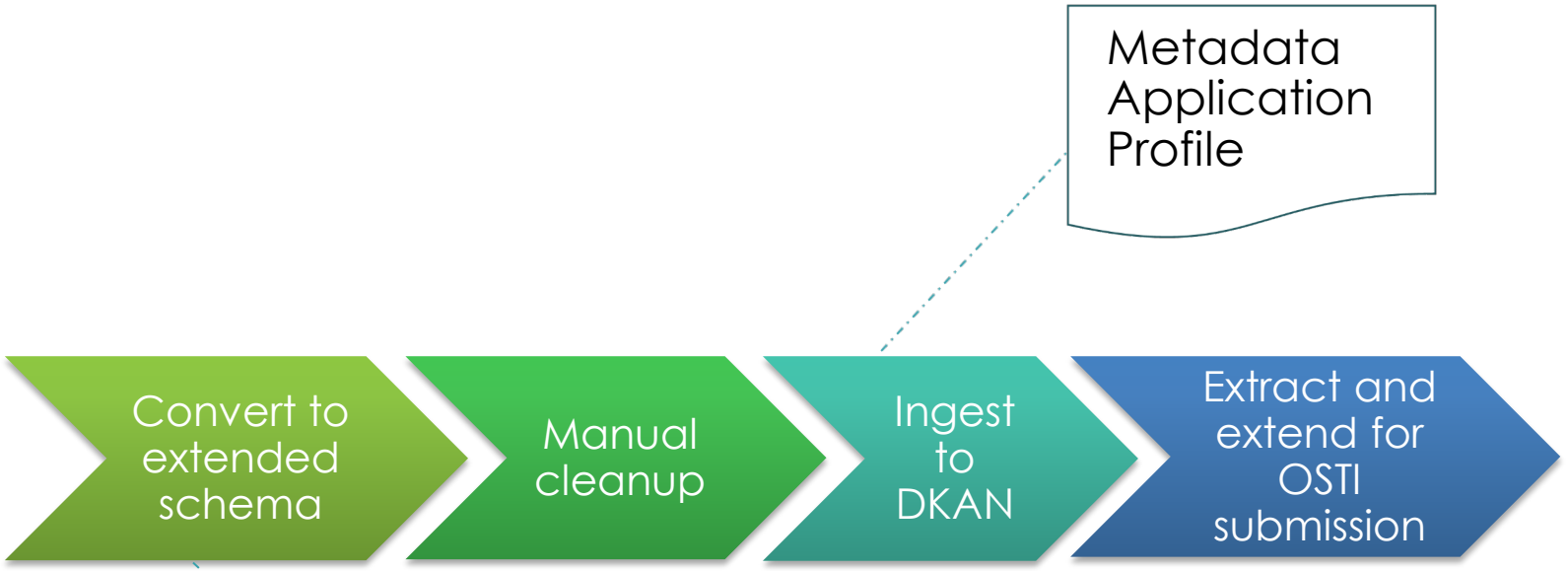
Why DKAN?

- Scalability
- Flexible content taxonomy
- Advanced user permissions
- Drupal customization
- Security

Schema Requirements



Current Metadata Workflows



Crosswalk

Legacy DB	DCAT-US	DataCite	E-Link 1.0	E-Link 2.0

DOE Metadata Landscape

- Complex repository ecosystem- varied data, varied metadata
- Expected increase in open data publishing following OSTP memo
- Expansion of PID services led by OSTI
- Key objective – citation and attribution
  - Handling of authors varies by schema
  - Service to our users and part of tracking provenance

From 'Desirable Characteristics of Data Repositories for Federally Funded Research' [1]

Metadata	The repository ensures datasets are accompanied by metadata to enable discovery, reuse, and citation of datasets, using schema that are appropriate to, and ideally widely used across, the communities that the repository serves.

From OSTP Public Access Memo [2]

a) Collect and make publicly available appropriate metadata[15] associated with scholarly publications and data resulting from federally funded research, to the extent possible at the time of deposit in a public access repository. Such metadata should include at minimum:
   i) all author and co-author names, affiliations, and sources of funding, referencing digital persistent identifiers,[16] as appropriate;
   ii) the date of publication; and,
   iii) a unique digital persistent identifier for the research output;

Next Steps for Constellation

Alignment with DCAT-US 3.0

- Collaborate with DKAN developers

PID implementation

- Incorporate into schema
- Reconcile legacy records (ORCIDs and RORs)

Scaling beyond curator intervention

- Automation and user guidance

Prepare for E-link 2.0

- Authorities and validation

Pathways to a Generalist Repository Schema

- Agreement on core elements for sharing DOE-funded research
  - DataCite required fields?
  - Understand needs for capturing provenance and funding
- Recommendations around PID requirements and implementation
- Interoperability
- Work underway in DOE Data Curation Working Group:
  - Establishing DOE-wide ontologies
  - Shared understanding of DOE data publishing landscape
  - Schema discussions

OAK RIDGE National Laboratory

Thank you! Questions?

References

1. The National Science and Technology Council. (2022). *Desirable Characteristics of Data Repositories for Federally Funded Research.* https://doi.org/10.5479/10088/113528

2. Nelson, A. (2022). Ensuring Free, Immediate, and Equitable Access to Federally Funded Research. Office of Science and Technology Policy. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

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

This work was carried out at Oak Ridge National Laboratory, managed by UT-Battelle, LLC for the U.S. Department of Energy under contract DE-AC05-00OR22725

U.S. DEPARTMENT OF ENERGY