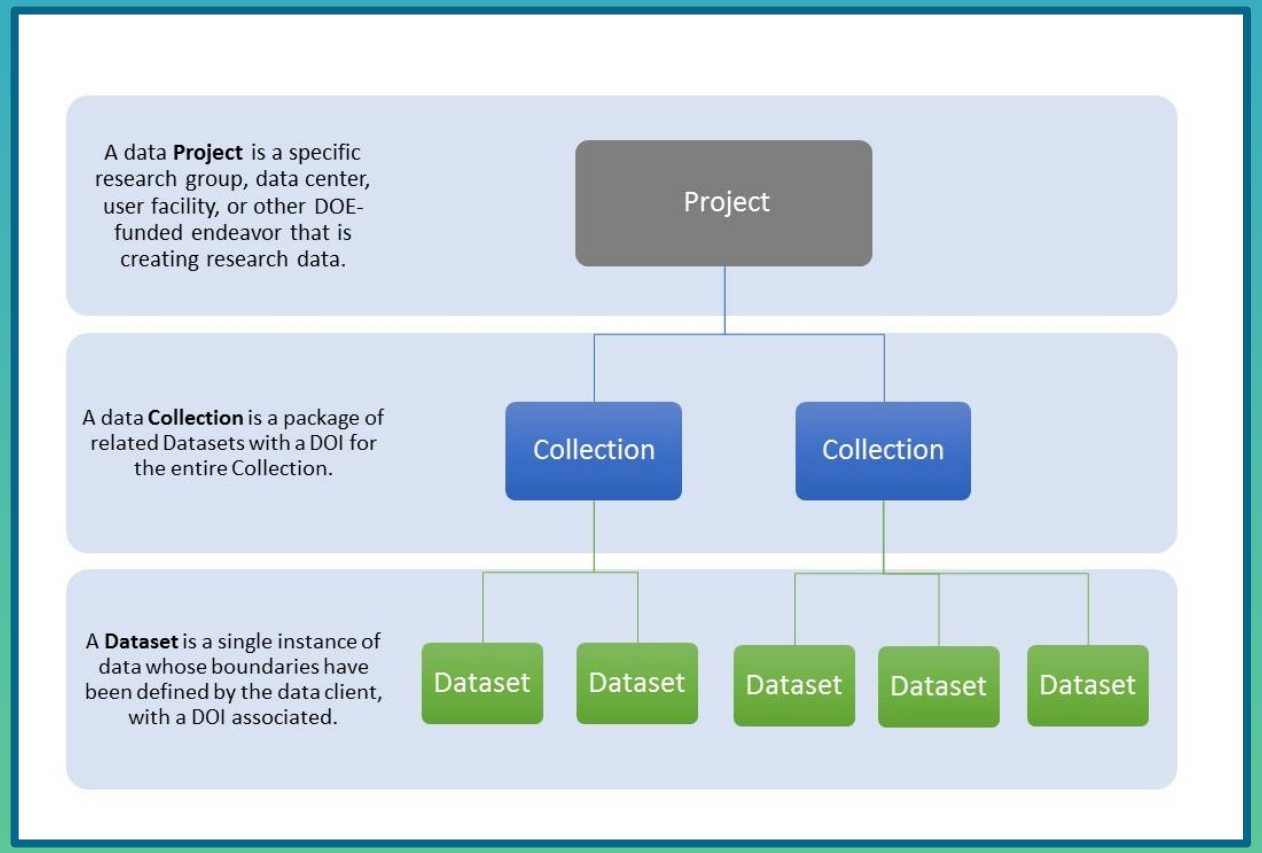
Data Discovery

Sara Studwell
Data Team Lead & DOE Data Explorer Product Manager
studwells@osti.gov

2019 DOE Data ID Service Workshop • September 17-18

DOE Data Explorer

DOE Data Explorer (DDE) is the search tool for finding DOE-funded, publicly available, scientific data submitted by data centers, repositories, and other organizations funded by the Department. The data themselves reside at national laboratories, data centers, user facilities, colleges and universities, or other websites.

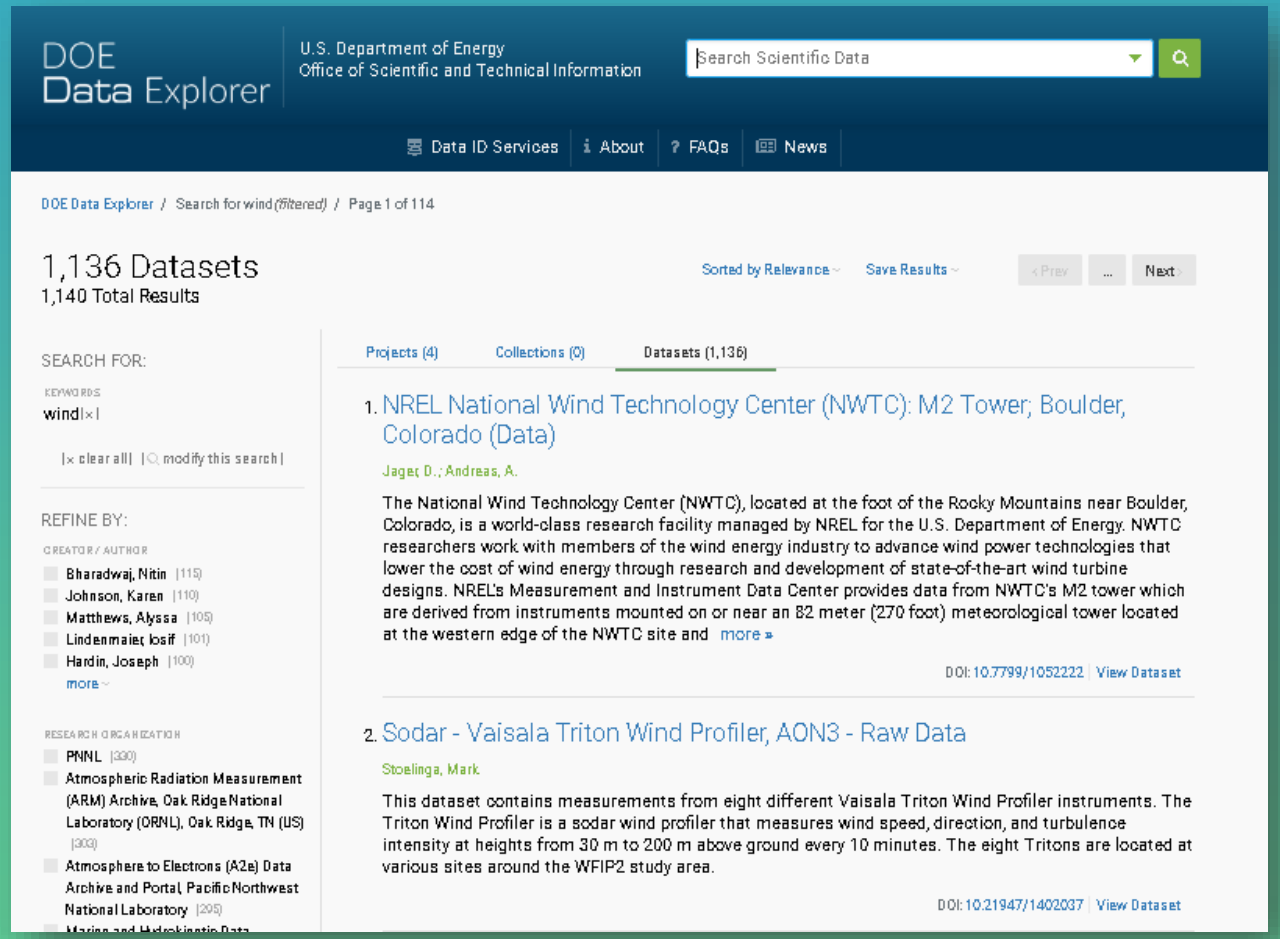


DDE includes data Project, data Collection, and individual Dataset records.

- A data Project is a scientific research group, data center, user facility, or other DOE-funded endeavor that is creating research data.
- A data Collection is a package of related datasets with a DOI for the entire collection.
- A dataset is a single instance of data whose boundaries have been defined by the data submitter, with a DOI associated.

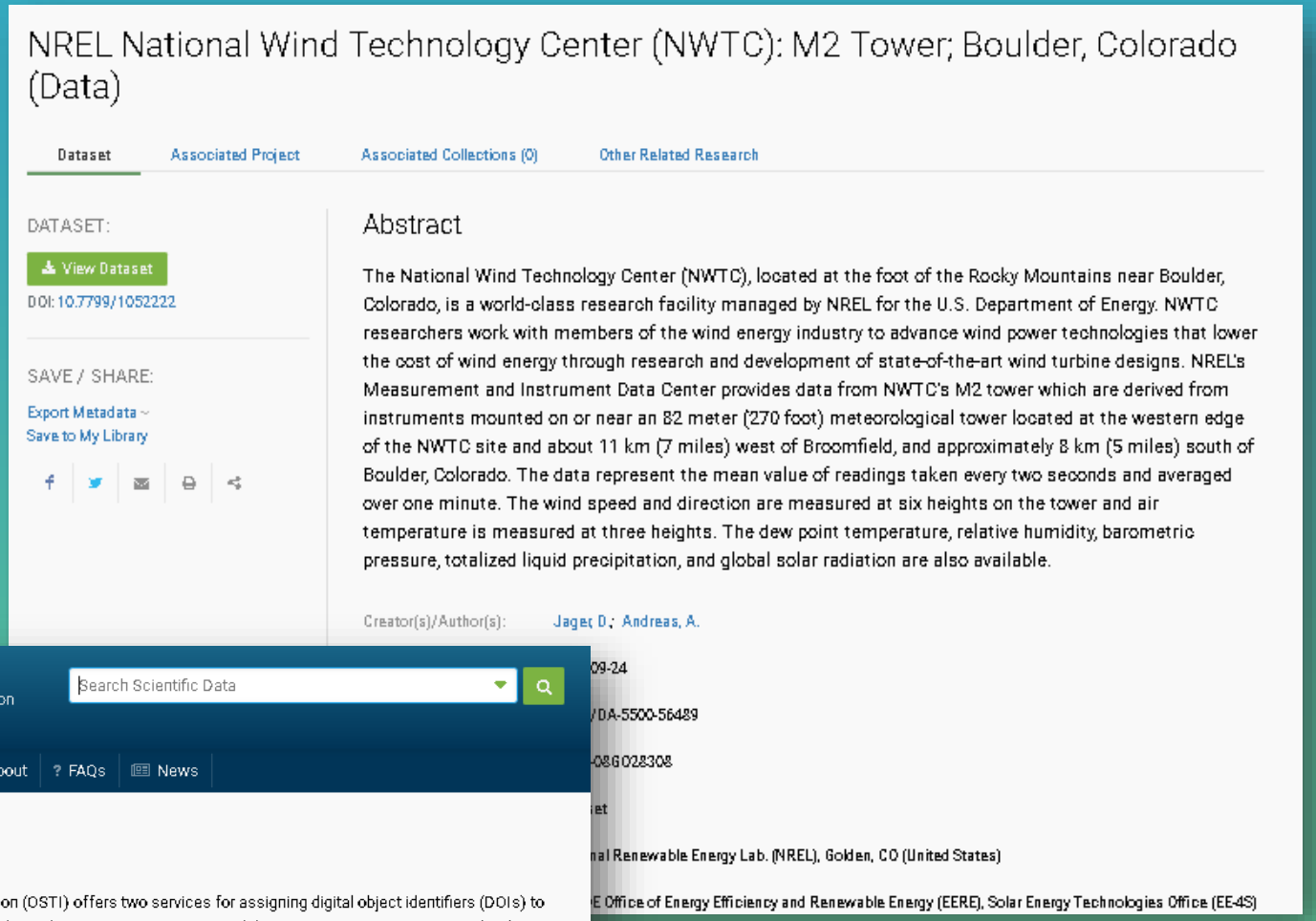
Search and Navigation

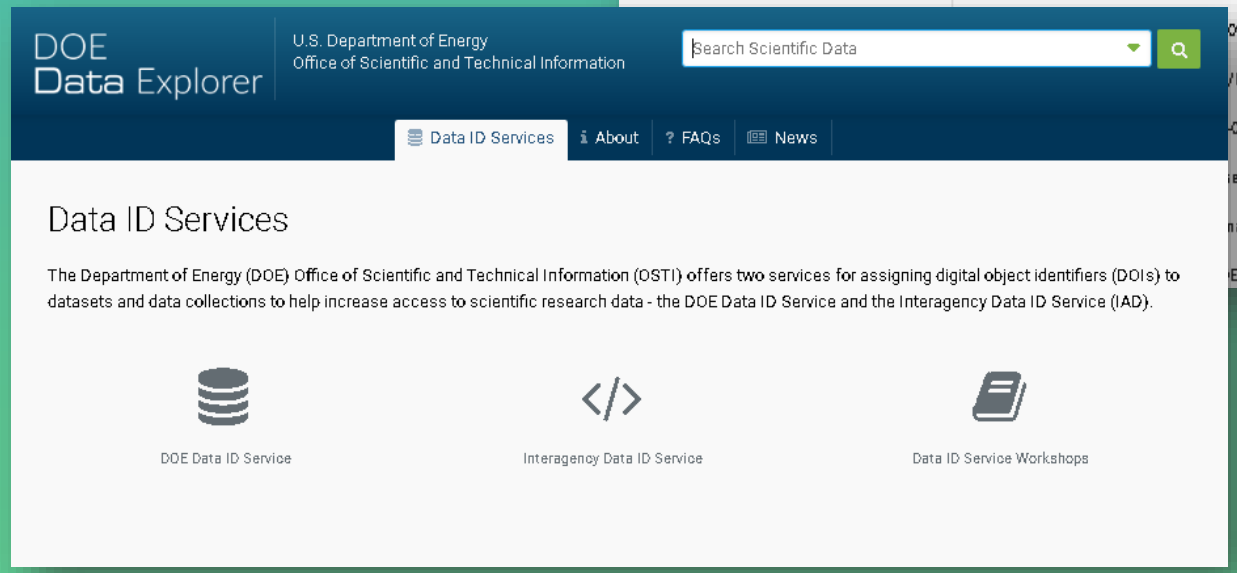
- Basic and advanced search options
- Toggle between Projects, Collections, and Datasets using tabs at the top of the results list
- Refine options change based on results product type



Additional Features

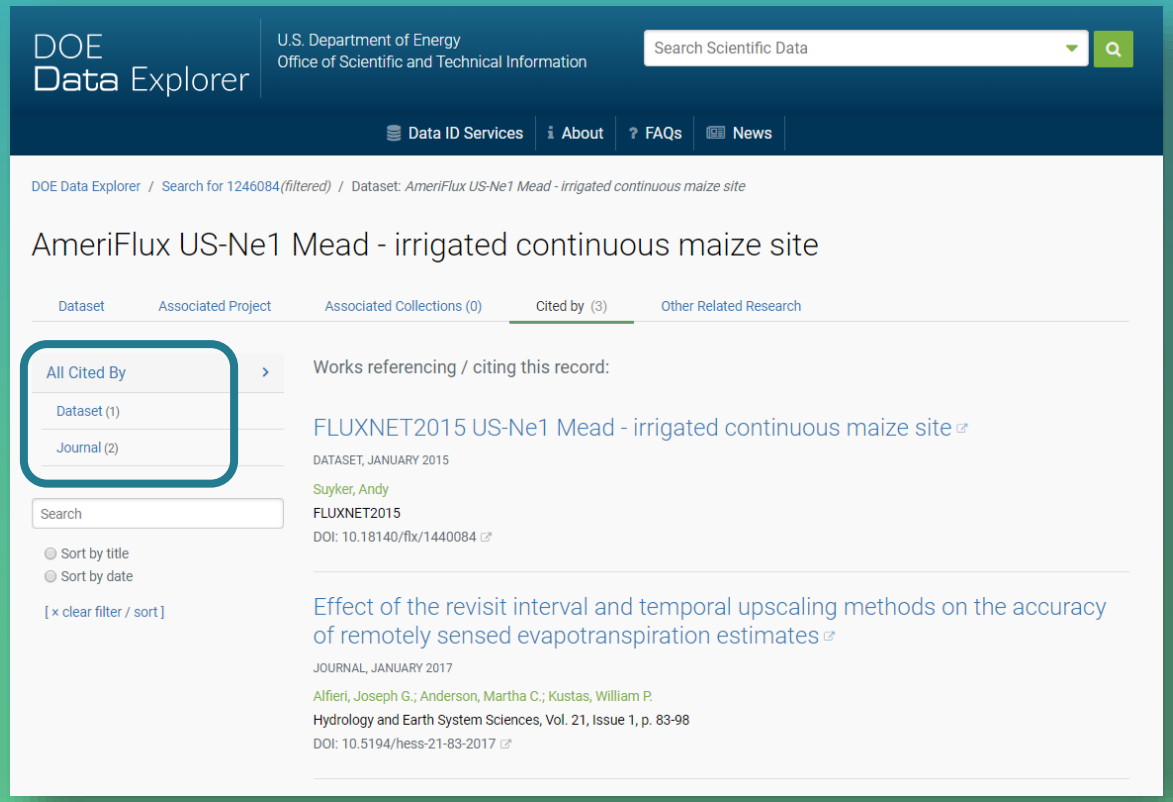
- Save/share options
- Account functionality
- News and information pages about Data ID Services
- Other Related Research

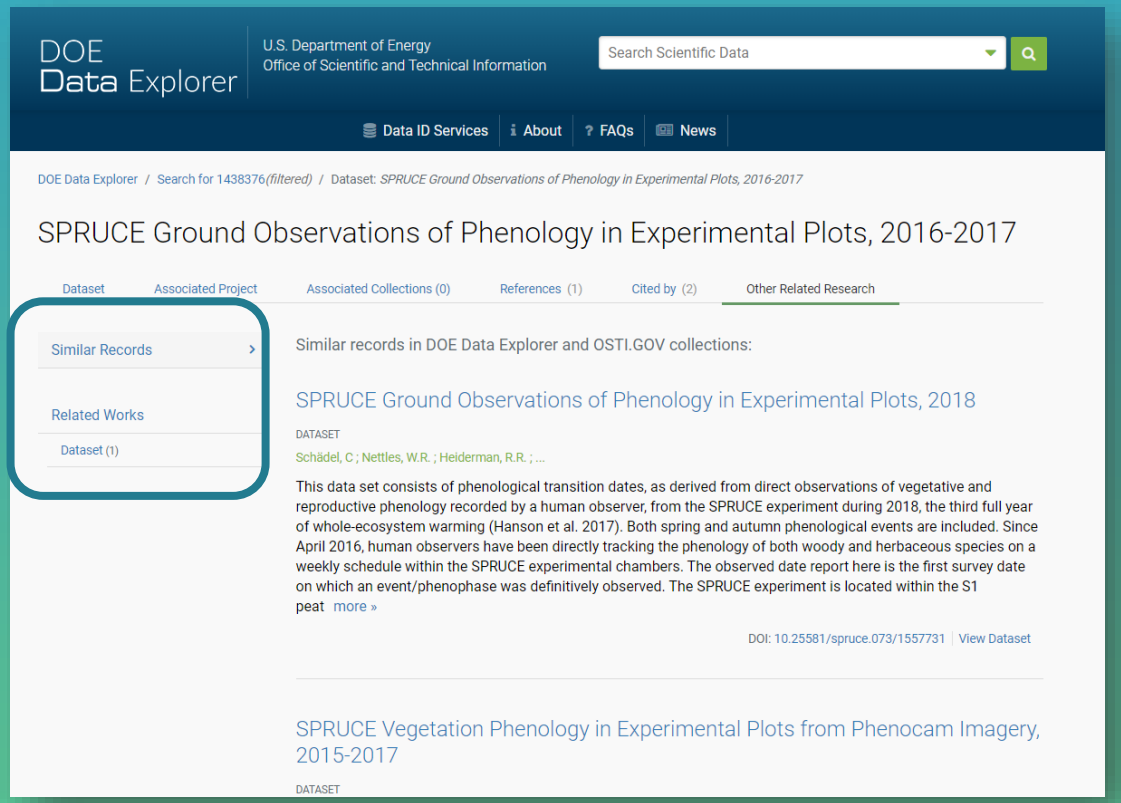




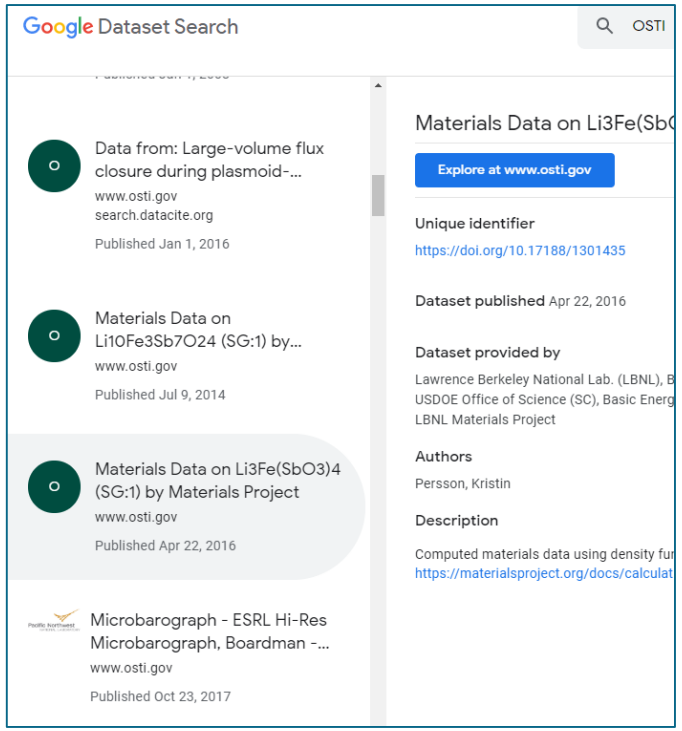
@OSTIGOV • #DOEDataID2019

Other Related Research





Google Dataset Search Beta



- Much like Google and Google Scholar, Google Dataset Search allows users to search and access datasets hosted in thousands of repositories.
- Currently indexes repositories that use the schema.org structured data guidelines.
- OSTI implemented Google's structured data guidelines (largely schema.org) for datasets, using both JSON-LD and microdata representations.
- Structured data changes were implemented well in advance of Google Dataset search release, ensuring OSTI STI datasets could be made available immediately.
- Structured data was tested and validated using Google structured data tools.
- Source and provenance best practices were applied, ensuring that sitemaps and canonical references for dataset STI were in place.

DOE DATA ID Service

The FAIR Guiding Principles

To be Findable:

- (meta)data are assigned a globally unique and persistent identifier
- data are described with rich (meta)data
- (meta)data clearly and explicitly include the identifier of the data it describes
- (meta)data are registered or indexed in a searchable resource

To be Accessible:

- (meta)data are retrievable by their identifier using a standardized communications protocol
  - the protocol is open, free, and universally implementable
  - the protocol allows for an authentication and authorization procedure, where necessary
- (meta)data are accessible, even when the data are no longer available

To be Interoperable:

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

To be Reusable:

- meta(data) are richly described with a plurality of accurate and relevant attributes
  - (meta)data are robust, accurate, and descriptive, adequately describing the associated data.
  - (meta)data are released with a clear and accessible data usage license
  - (meta)data are associated with detailed provenance
  - (meta)data meet domain-relevant community standards

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data 3* (2016). DOI: 10.1038/sdata.2016.18

DOE DATA ID Service

FAIR Guiding Principles at OSTI

Findable	Accessible	Interoperable	Reusable
• OSTI assigns a Digital Object Identifier (DOI) to every dataset metadata submission. • OSTI offers 13 required metadata fields. • DOI is included in the metadata record found in OSTI.GOV and DDE. • All metadata records are registered with DataCite. Records are searchable in OSTI.GOV, DDE, DataCite, Google Dataset Search, DX.DOI.ORG.	• Publicly available metadata is found in OSTI.GOV and DDE, as well as their respective APIs. • OSTI provides a REST API for metadata discovery. • Metadata records are always available at OSTI.GOV or DDE. If the URL is broken or the data is no longer available, the URL should indicate the location/availability of the data.	• OSTI offers dataset metadata submission via E-Link API (XML) or web submission interface. • OSTI dataset metadata follow the DataCite schema required for obtaining a DOI, and necessary DOE-relevant metadata. • OSTI allows for linkages to be created between research through the use of the Related Identifiers metadata field.	• OSTI offers 13 required metadata fields. • OSTI encourages data metadata submitters to provide accessibility information. • OSTI encourages data clients to provide provenance information on dataset landing pages • OSTI encourages metadata submitters to adhere to respective domain-relevant community standards.

Data Metrics

- DataCite
  - https://stats.datacite.org/
  - Registrations by members, clients, prefixes
    - By year
    - DOI metadata findable, registered
  - Resolutions by month
    - Total attempted, successful, failed, top 10 successes/failures
- OSTI
  - Submitted Datasets
    - By site, date range
  - Submitted Collections
    - By site, date range
- Repository
  - Can be accomplished with Google Analytics or analog counts
  - Pageviews, downloads, referrals, etc.

Questions?

ddecomments@osti.gov

studwells@osti.gov