DOE Office of Scientific and Technical Information (OSTI)
Artificial Intelligence and Machine Learning

DOE Data Days (D3)
June 2, 2022
Mary Beth West
DOE Invests ~$15B per Year in R&D

R&D Funding

- Advanced Research Projects Agency – Energy (ARPA-E)
- Office of Electricity Delivery & Energy Reliability
- Office of Energy Efficiency & Renewable Energy
- Office of Environmental Management
- Office of Fossil Energy
- Office of Legacy Management
- Office of Nuclear Energy
- Office of Science
- Office of Environment, Health, Safety & Security

NATIONAL LABS
- Ames
- Argonne
- Brookhaven
- Fermi
- Idaho
- Los Alamos
- Lawrence Berkeley
- Lawrence Livermore
- NETL
- NREL
- Oak Ridge
- Pacific Northwest
- Princeton
- SLAC
- Sandia
- Savannah River
- Thomas Jefferson

GRANTEES
TECHNOLOGY CENTERS
SITES

R&D Results
- Journal articles/accepted manuscripts
- Technical reports
- Conference papers
- Theses/dissertations
- Scientific and technical software
- Data objects
- Patents
- Workshop reports
- Videos

≈ 50,000 R&D results per year
**Main Drivers**

**OSTI’s Mission:** Advance science and sustain technological creativity by making R&D findings available and useful to Department of Energy researchers and the public.

The [Office of Scientific and Technical Information](https://www.osti.gov) (OSTI) collects, preserves, and disseminates DOE-funded R&D results, both through OSTI search tools and through other commonly-used search engines.

**OSTI.GOV**
- Primary search tool for all DOE-funded R&D results
  - Over 3.3 M metadata records; 800K full-text/resource available
  - Semantic search "keyword to concept" mapping; AI potential
AI / Machine Learning for R&D Results

R&D Results Data: Opportunities

• > 3.3M scientific and R&D results emanating from DOE-funded research and development (R&D) activities.

• Many types of research products represented, including journal articles, datasets, patents, software, technical reports.

• Data spans DOE research interests, research organizations, and sponsoring organizations.

• Metadata is searchable and available electronically, API / export use cases supported.

OSTI mission data is increasingly targeted as a valuable source of bulk metadata for analysis, machine learning exercises, and text mining.
AI / Machine Learning for R&D Results

R&D Results Data: Challenges

• R&D results are being generated more rapidly and in more varying forms.

• Access patterns are changing, conventional users / searches are no longer the only first-class use cases.

• New usage requirements for search results / formats are emerging.

• Existing data, structures, and formats don’t always lend well to analysis at scale.

• Association of related data is increasingly critical for research and reproducibility.

OSTI data and procedures need to support comprehensive, FAIR collections.
AI / Machine Learning for R&D Results

R&D Results Data: Addressing Challenges

• Improve ingest processes of R&D results to handle scale of direct submission (i.e., other alternatives for submission, such as Web Services and APIs for ingest).

• Data integrity for R&D results and their products.

• Mechanisms to support the reuse of data.

• Procedural improvements to augment tools that standardize reporting and analysis (i.e., additional metadata dimensions will allow for better topic / trend / funding analysis).
Data Opportunity: DOE OSTI Organization Authority

- DOE OSTI ingests and curates organization information provided about DOE-funded research outputs – including author affiliations, research organizations, and funding organizations.
- We developed an internal organization authority to standardize organization naming during metadata submission and within in the research output records in our search tools.
- Organization authority allows for enhancement of user submitted metadata. Providing consistent naming structures, enhanced organization labeling, and hierarchical mapping.

![Diagram of DOE OSTI Organization Authority](image-url)
Organization Disambiguation

Turn these … … into this

- Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States)
- Oak Ridge National Laboratory, Oak Ridge, TN (United States)
- ORNL
- Oak Ridge Nat. Lab
- Oak Ridge National Lab. (ORNL)
- Oak Ridge National Lab., Tenn.
- etc. …

Drivers

- Improves the findability of research outputs by enhancing our search tools through normalization of metadata.
- Enables organization and sector-level analysis of research outputs.
- Supports enhancements to organization data that enable meaningful analytics (e.g., geolocation, congressional district).
- Makes data more interoperable by supporting organization data with persistent identifiers.
- Supports additional metadata for other research objects (e.g., award DOIs, ORCID iDs, etc.).
Organization Disambiguation

How it works ...

- Organization data is aggregated from multiple authoritative sources, including ROR, GRID, OSTI authorities, and others.
- Authority application developed to maintain / query authoritative references and extended metadata.
- Disambiguation tools (indexing, reconciliation APIs) leverage NLP, similarity, and authority info to resolve variations and assign authoritative entity data to variant organization names.
- Normalized labels and enhanced organization metadata are made available to downstream processes (e.g. submission and search tools).
Organization Authority and Reconciliation API

Organization Name

Aliases

Organization Type

Geographic Information

Organization PIDs

{ "id": "2139",
  "name": "Oak Ridge National Laboratory",
  "lat": "-84.3170013",
  "lng": "35.9329987",
  "city": "Oak Ridge",
  "state": "Tennessee",
  "state_code": "TN",
  "country_code": "US",
  "org_type": "facility",
  "org_acronyms": "ORNL",
  "parent_id": "{2052}",
  "ext_ids": {
    "fundref": "100006228",
    "isni": "0000 0004 0446 2659",
    "ror": "https://ror.org/01qz5mb56",
    "grid": "grid.135519.a",
    "geonames_city_id": "4646571",
    "wikidata": "Q714439",
    "orgref": "38147"
  }
}
Reconciliation API

- The API adheres to the draft reconciliation API spec for query behavior and result formats.
- Query results return confidence scores and match IDs, which can be used to fetch extended organization metadata (e.g., PID, sector, MSI status).
- Supports additional query fields (e.g., city, state, country) to improve match resolution and multiple queries per request to enable bulk operations.
- The API integrates with data cleaning and transformation tools like OpenRefine.
- Standard JSON is used for both input and output.

Application Status: Currently in development and testing; anticipate deploying late FY22.
Comments/Questions

Mary Beth West
westm@osti.gov

www.osti.gov
www.osti.gov/ai/

AI@osti.gov
comments@osti.gov

@osti.gov