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

DOE Data Days (D3)
October 5 - 7, 2020

Josh Nelson and Mary Beth West

U.S. DEPARTMENT OF ENERGY | Office of Science | Office of Scientific and Technical Information

DOE Invests ~$12B per Year in R&D

U.S. DEPARTMENT OF ENERGY

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

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

Congress Requires DOE to Provide Public Access to our R&D Results

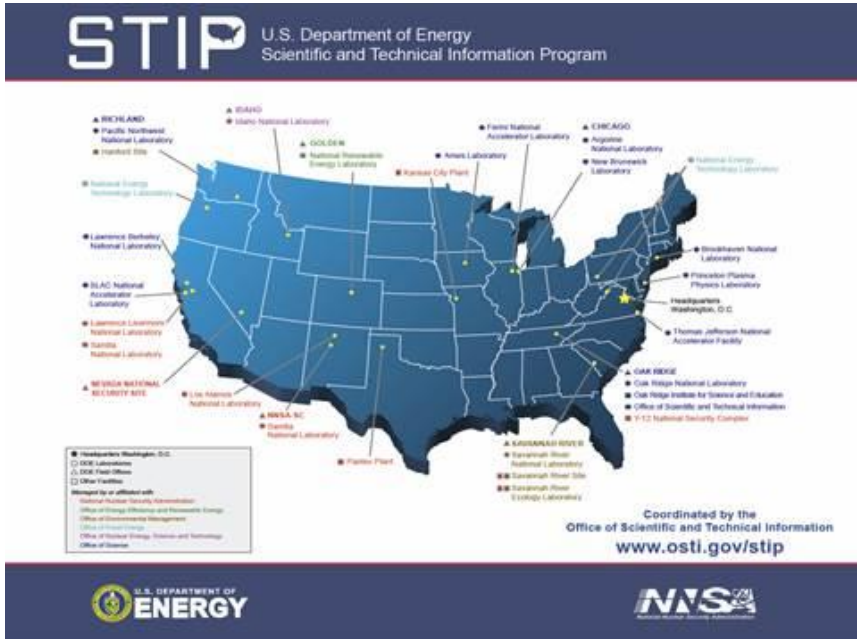
- Enabling legislation for AEC/ERDA/DOE: *"The Secretary shall . . . maintain a central source of information . . . in furtherance of the research, development, and demonstration mission . . . The information shall be made available to the public."* (resulted in establishment of OSTI in 1947)
- Energy Policy Act of 2005, P.L. 109-58, Section 982: *"The Secretary, through the Office of Scientific and Technical Information, shall maintain within the Department publicly available collections of scientific and technical information resulting from research, development, demonstration, and commercial applications activities supported by the Department*."

The Office of Scientific and Technical Information (OSTI) collects, preserves, and disseminates DOE-funded R&D results, both through OSTI search tools and through other commonly-used search engines.

OSTI's Three Core Functions:
Collect, Preserve, and Disseminate DOE R&D Results

Collect

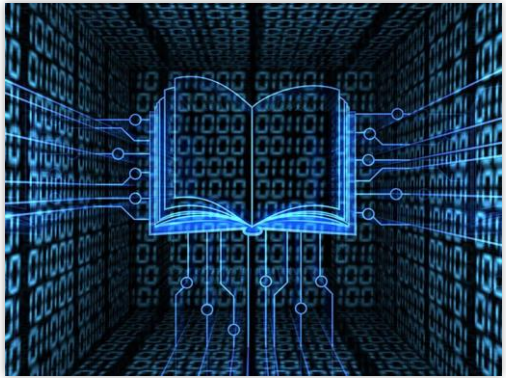
FY19 R&D Result Records = 54,017
(publications, datasets, software, patents, etc.)



- Collection is all electronic, through simplified submission systems
  - APIs to labs' publication management systems
  - Embedded submission links in grantees' reporting tools
- OSTI develops and maintains DOE corporate ingest system (E-Link) to collect the results of DOE-funded research from DOE labs and grantees.
- OSTI coordinates the Scientific and Technical Information Program (STIP), which is a Department-wide collaboration, with points of contact at every DOE office, laboratory, and facility.

OSTI's Three Core Functions:
Collect, Preserve, and Disseminate DOE R&D Results

Preserve



From 1940s – Present

- ≈ 800K electronic publications
- ≈ 80K metadata records and DOIs for datasets
- ≈ 40K patents
- ≈ 3,500 software projects
- ≈ 3M electronic records
- Redundant storage infrastructure - combination of on-premise, cloud, and offsite
- 1M physical documents
- Digitization of legacy collection

OSTI's Three Core Functions:
Collect, Preserve, and Disseminate DOE R&D Results

Disseminate

OSTI.GOV

Primary search tool for all DOE-funded R&D results
- Over 3.1 M metadata records; 800K full-text/resource available
- Semantic search "keyword to concept" mapping; AI potential

DEPARTMENT OF ENERGY DOE PAGES Public Access Gateway for Energy & Science
- 83K accepted manuscripts
- Fulfills public access mandate

DOE Patents
- ≈41,000 patents from 1940s to present
- Harvested from USPTO

DOE Data Explorer
- Metadata and links for 76K datasets

DOE ScienceCinema
- ≈3,700 videos from DOE and CERN
- Speech indexing by IBM Watson

DOE CODE
- ≈2,100 software projects
- Feeds Code.gov

OSTI enables indexing by Google, Google Scholar, Google Dataset Search, Bing, etc., resulting in 50% of OSTI's web traffic being referred by Google.

OSTI's Approaches to Artificial Intelligence and Machine Learning for R&D Results

AI / Machine Learning for R&D Results

R&D Result Data: Opportunities

- > 3.1M 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 Result 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

Data Challenges: Establishing a comprehensive corpus of R&D results

- Improvements to 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).
- Digitization of physical R&D results and their assets (e.g., technical reports) and registering DOIs (via Crossref), when applicable.
- Identification of emerging product types and mechanisms for ingest (e.g., datasets, software) and related support services (e.g., Data ID Service for minting DOIs).
- Identification of reliable resources to supplement ingested data and missed resources (i.e., USPTO, Scholix, etc.).
- Development of strategic data partnerships (i.e., Crossref, CHORUS, WoS, ORCID, etc.).

AI / Machine Learning for R&D Results

Data Challenges: Ensuring data integrity for R&D results and their products

- Development of standards to support the structural homogeneity of metadata, supporting submission requirements (i.e., metadata schema for Web Services, consistency among forms and submission resources).
- Mitigation of data quality issues (e.g., submission errors, OCR errors).
- Disambiguation and consistent use of entity labels (i.e., in-house resources for organizational entity hierarchy curation).
- Development of data authorities, vocabularies, and thesauri.

AI / Machine Learning for R&D Results

Data Challenges: Providing mechanisms to support the reuse of data

- Development of discovery and research tools ensuring products and their metadata are findable.
- Ensuring interoperability and machine readability by following relevant schemas for structured data (e.g., XML Sitemaps, JSON-LD, etc.).
- Working with 3rd Parties to ensure that data is findable (e.g., Google Scholar, Google Datasets, Microsoft Academic, DATA.GOV, Code.gov, etc.).
- Development of ingest tools to support the registering of persistent identifiers (e.g., Data ID Service, DOE CODE, Award DOI Service, etc.).

AI / Machine Learning for R&D Results

Data Challenges: Providing mechanisms to support the reuse of data

- Expanding or enhancing the data to support interoperability.
- Improvements to referencing and interrelating data.
- Developing data import and curation workflows.
- Extracting and improving extracted metadata from full-text (e.g., figures and tables).
- Implementing dataset version support.
- Improving API support (e.g., RAML and OpenAPI) for communication and for generating documentation and client code.
- Supporting updated API output formats for improved interoperability (e.g., CSL-JSON).

AI / Machine Learning for R&D Results

Data Challenges: Procedural improvements

- Improvement of integration with the STIP Community (e.g., laboratories and Financial Awarding offices) to modernize and automate R&D results collection.
- Identification of Departmental expertise and resources (e.g., OTT reporting standards, common AI initiatives) to leverage best practices.
- Leveraging Data Management Plans (DMPs) to improve visibility and impact of DOE research, document standards for preservation and sharing.
- Augmenting tools to standardize reporting and analysis (i.e., additional metadata dimensions will allow for better topic / trend / funding analysis).
- Expanding use of persistent identifier services to improve data reuse.
- Identification and support of interaction use cases (e.g. REST APIs).

AI / Machine Learning for R&D Results

Data Challenges: Future development and resources

- Ongoing development of tools to capture high quality, detailed metadata for the improved findability, accessibility, interoperability, and reuse of digital assets (FAIR).
- Developing publicly available Natural Language Process (NLP) and Machine Learning (ML) models trained on the DOE-funded research corpus.
- Expanding the development of R&D results-specific technical vocabularies, thesauri, authorities and other lexical datasets.
- Improving machine readability and data quality / structure for the scalable mining and analysis of data.

THANK YOU!

Josh Nelson
nelsonjc@osti.gov

Mary Beth West
westm@osti.gov

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

AI@osti.gov
comments@osti.gov

@osti.gov