U.S. DEPARTMENT OF ENERGY | Office of Science

Public Reusable Research (PuRe) Data Resources

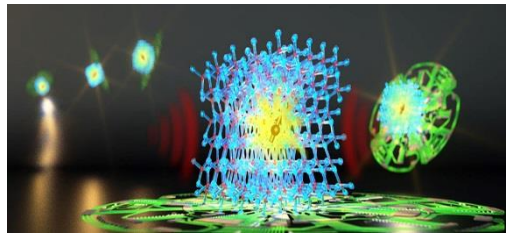
Biological and Environmental Research Advisory Committee Meeting

October 22, 2021

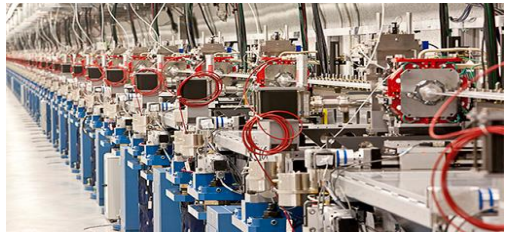
Michael Cooke
Senior Technical Advisor
Office of the Deputy Director for Science Programs
Department of Energy, Office of Science

DOE and Office of Science Mission

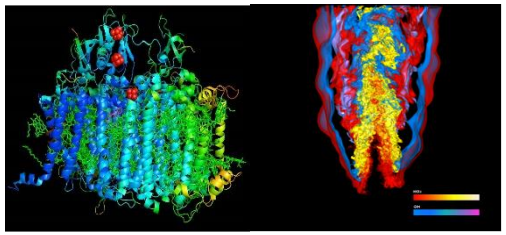
- **The Mission of the U.S. Department of Energy** is to ensure America's security and prosperity by addressing its energy, environmental and nuclear challenges through transformative science and technology solutions.
- **DOE Office of Science** is the lead federal agency supporting fundamental scientific research for energy and the largest supporter of basic research in the physical sciences in the United States, fulfilling a Mission to:
  - Deliver scientific discoveries and major scientific tools to transform our understanding of nature
  - Advance the energy, economic, and national security of the United States



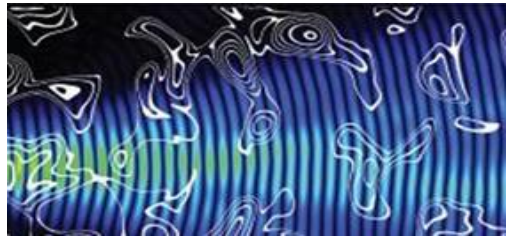
Advanced Scientific Computing Research



Basic Energy Sciences

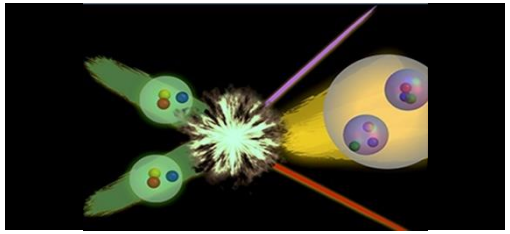


Biological and Environmental Research



Fusion Energy Sciences

High Energy Physics



Nuclear Physics

Office of Science at a Glance

- Fiscal Year 2021 Enacted Funding: $7.026B

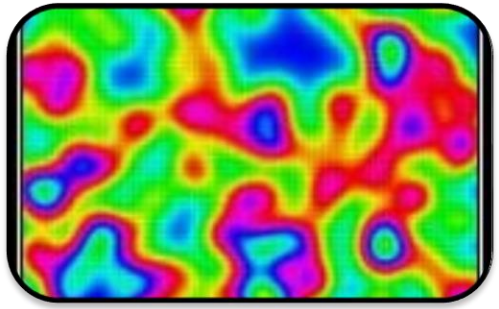
Largest Supporter of Physical Sciences in the U.S.

Funding at >300 Institutions, including 17 DOE Labs

Over 24,000 Researchers Supported

Over 33,500 Users of 28 SC Scientific Facilities

~40% of Research to Universities



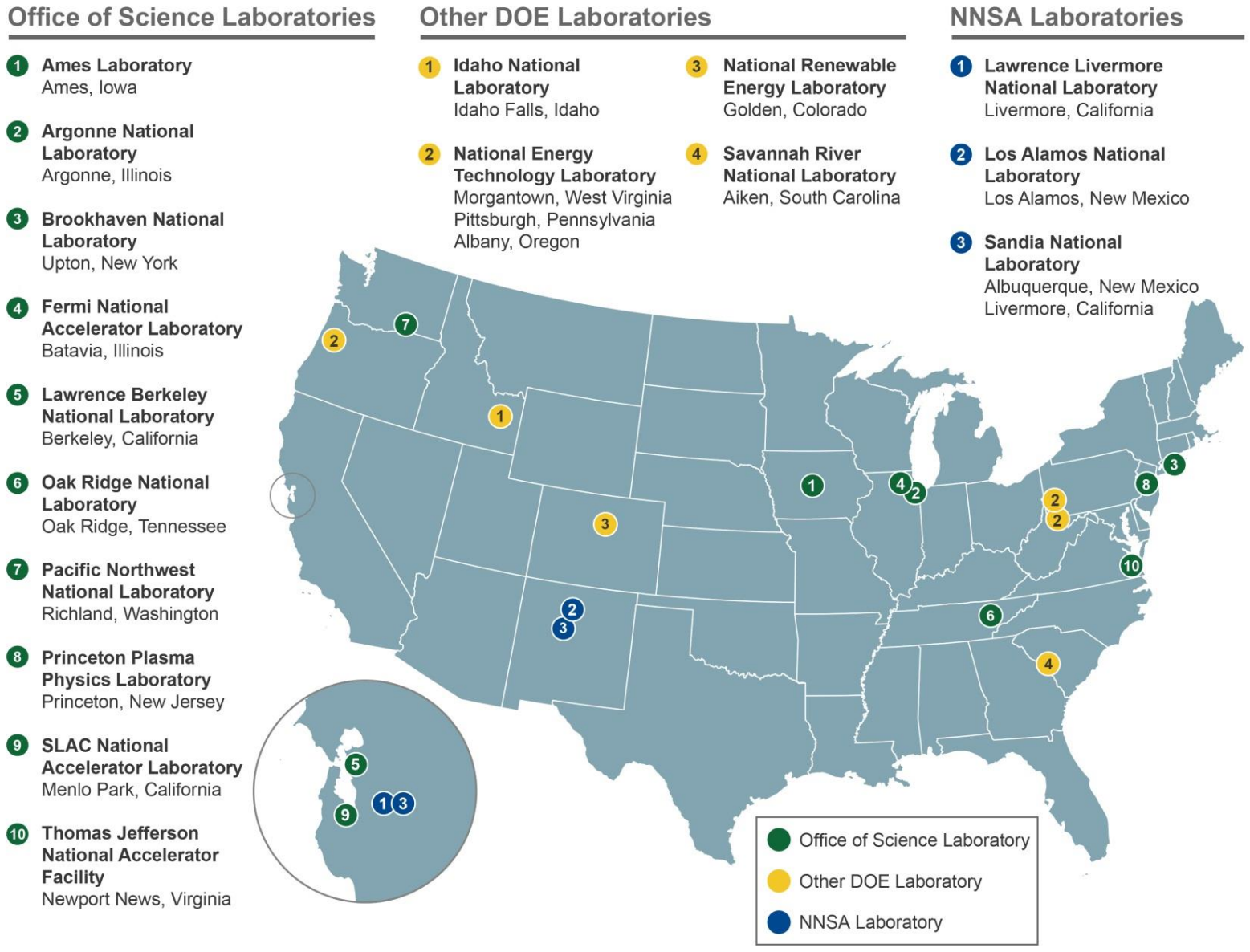
Research: 38.7%, $2.720B

Facility Operations: 37.0%, $2.602B

Projects/Other: 24.3%, $1.704B

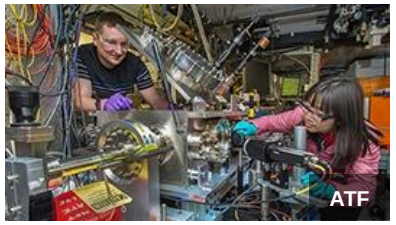
DOE National Laboratories

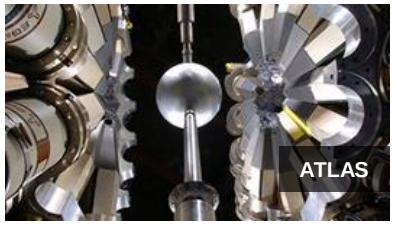
- The 17 DOE National Laboratories comprise a preeminent federal research system, providing the Nation with strategic scientific and technological capabilities
- SC stewards 10 DOE laboratories that provide essential support to the missions of the SC science programs

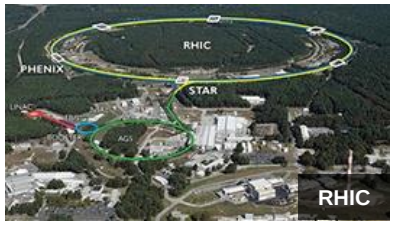


DOE SC Scientific User Facilities

FY 2021
28 scientific user facilities
33,500+ users

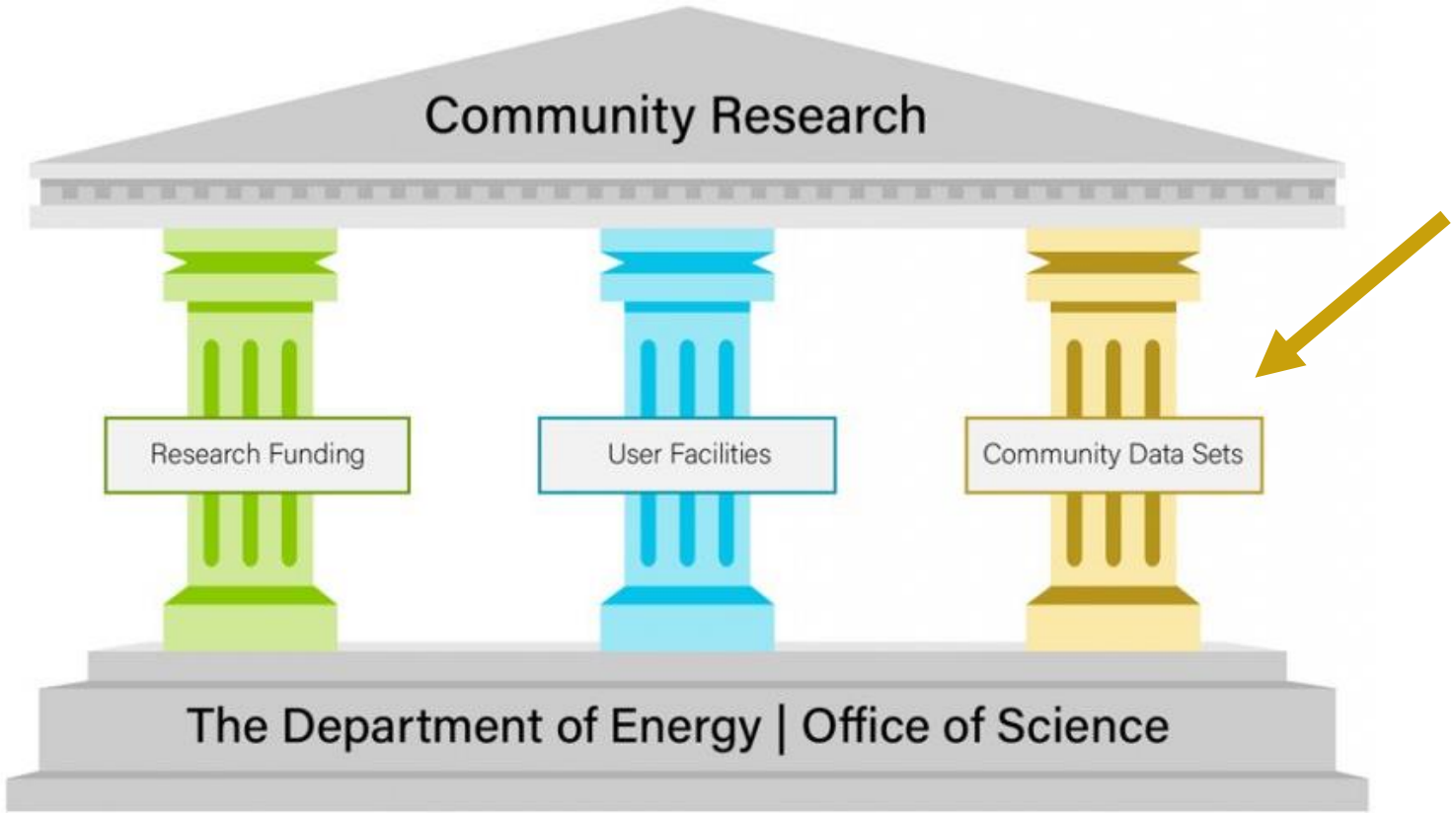






U.S. DEPARTMENT OF ENERGY
Office of Science

Data is the Third Pillar of the DOE SC Enterprise



https://www.energy.gov/science/office-science-pure-data-resources

Public Reusable Research (PuRe) Data Resources are:

- data repositories,
- knowledge bases,
- analysis platforms,
- and other activities

that aim to make data **publicly available** in order to advance scientific or technical knowledge.

PuRe Data Resource designations **highlight** and **improve stewardship** of SC-supported community data efforts with strategic impact on the SC mission.

Definition: A PuRe Data Resource is…

Data repositories

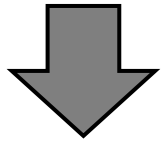
Knowledgebases

Analysis Platforms

Receives (sustained) funding from an Office of Science program

Considered an authoritative provider of data or capabilities in its subject area

Publicly available & uses DOIs to facilitate discovery, reuse, and citation of data and software



Recognized for the strategic impact on the Office of Science mission

Full definition: https://science.osti.gov/Initiatives/PuRe-Data/Policies-and-Procedures/Definition

Benefits to the Office of Science

Better Communication

Talk about high value public, reusable data

Articulate the benefits of these investments

#SCPuReData! Track highlights and link these to other investment outcomes

Better Stewardship

Sustained support

Recognize the long-term strategic value of these assets

Long-term preservation of the scientific record and the investment

Better Science

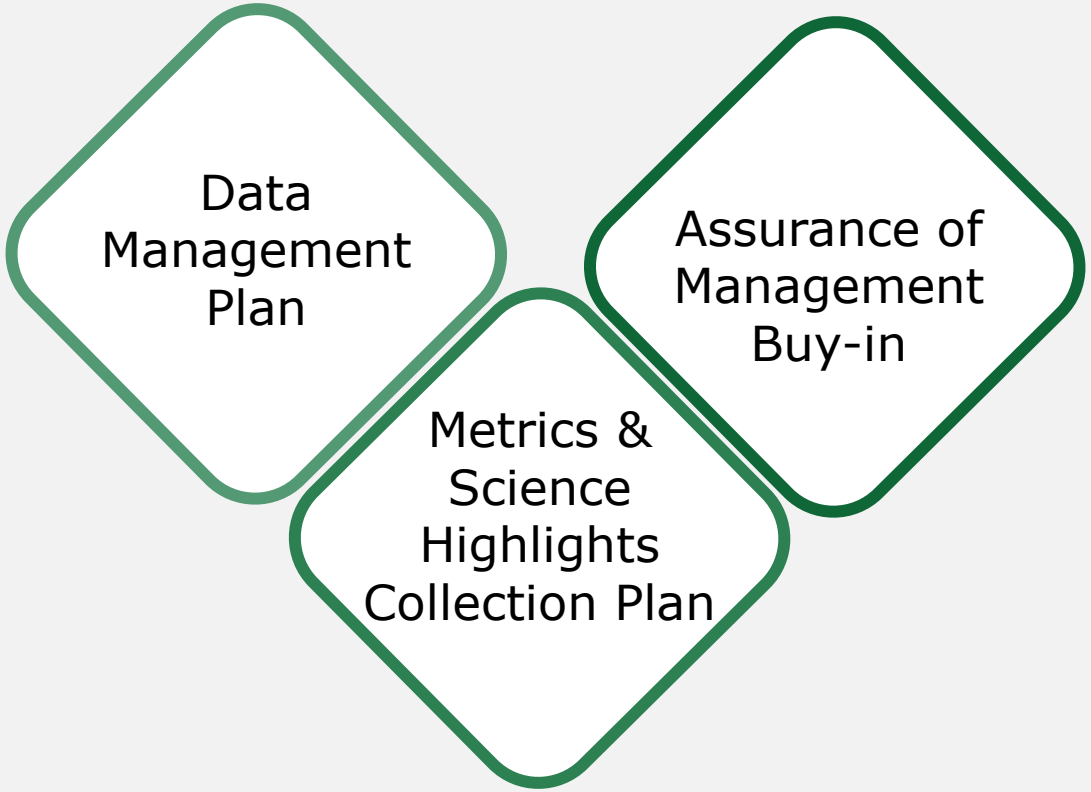
Hold data resources to high standards in data management, operations, and scientific impact

Develop a key workforce around data

Work toward making data more Findable, Accessible, Interoperable, and Reusable

Designation Process: A Collaborative Effort

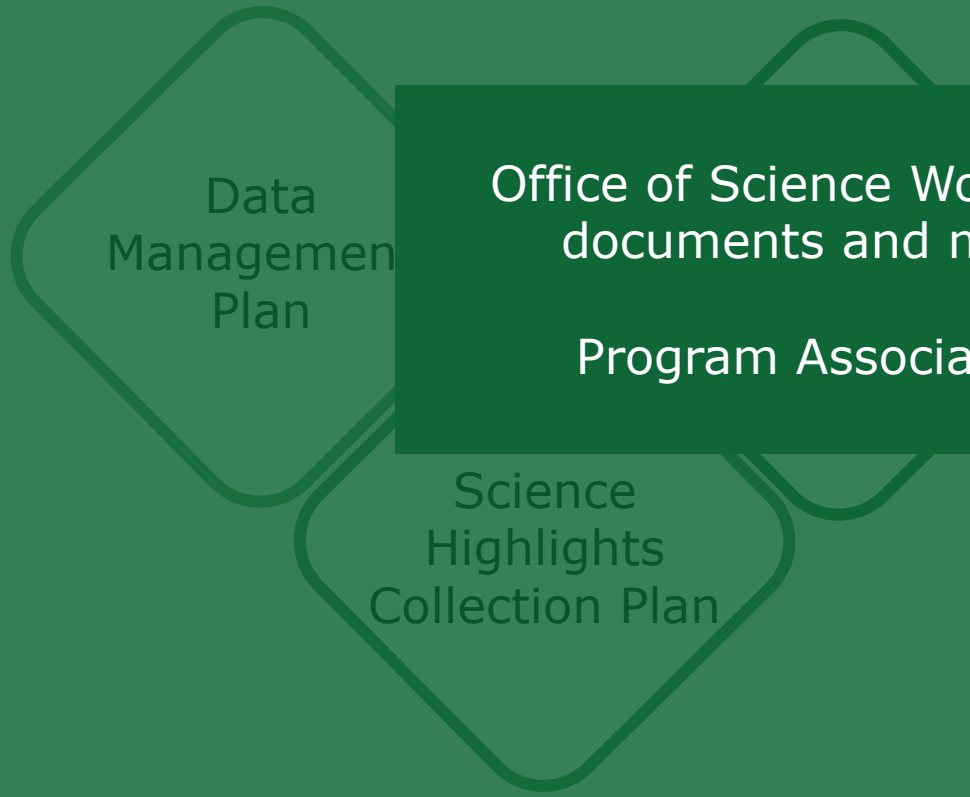
Candidates Resource Provides:



Program Office Provides:

Designation Process: A Collaborative Effort

Candidates Resource Provides:



Program Office Provides:

Office of Science Working Group on Digital Data reviews documents and makes an official recommendation

Program Associate Director approves designation

FAIR Principles

Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorization procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

Reusable

- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

The FAIR Guiding Principles for scientific data management and stewardship: https://doi.org/10.1038/sdata.2016.18

Desirable Characteristics of PuRe Data Resources

For designation as a PuRe Data Resource, the resource meets or strives towards the following characteristics:

- Plan for long term sustainability
- Provide clear user guidance
- Assign and use DOIs to facilitate discovery, reuse, and citation of data and software
- Provide metadata to accompany data
- Provide mechanisms for curation and quality assurance
- Offers public access to data and resources
- Track data provenance

Based on OSTP RFC, *Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research*, published January 17, 2020, available here.

PuRe Data Resources at a Glance

PuRe Data Resources

https://science.osti.gov/Initiatives/PuRe-Data/Resources-at-a-Glance

- Initially designated resources:
  - Atmospheric Radiation Measurement Data Center
  - Joint Genome Institute
  - Materials Project
  - National Nuclear Data Center
  - Particle Data Group
  - Systems Biology Knowledgebase (KBase)

JGI JOINT GENOME INSTITUTE

NNDC

PDG particle data group

The Materials Project

ARM

KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase

The Materials Project

The Materials Project is a recognized leader of **data generation, storage, and distribution in the domain of materials science.**

- Computational materials science database and analysis platform
- Provides scientists information of known and predicted materials and inspires the design of novel materials
- Allows researchers to data-mine scientific trends in materials properties

Sponsored by the office of Basic Energy Sciences (BES)

https://materialsproject.org/

» Accelerates innovation in materials research

» Predicts materials properties before they are synthesized in a laboratory

150,000 registered users, approximately 10,000 distinct daily users, and data on >130,000 inorganic compounds

Primary single publication associated with the Materials Project has been cited >3,000 times

Atmospheric Radiation Measurement Data Center

The Atmospheric Radiation Measurement (ARM) Data Center (ADC) is the **leading provider of ground-based atmospheric measurements with the goal of advancing atmospheric and climate research.**

- Atmospheric measurements from station-based locations around the country
- All data are available free of charge through Data Discovery
- ADC's two HPC Clusters allow scientists to access and conduct research using archived ARM data

Sponsored by the office of Biological and Environmental Research (BER)

https://www.arm.gov/data

» **Promotes the advancement of atmospheric, Earth, and climate science**

More than 11,000 data products totaling over 2.3 petabytes of data that dates back to 1992

Joint Genome Institute

The Joint Genome Institute (JGI) is the **global leader in generating genome sequences of plants, fungi, microbes, and metagenomes.**

- Provides a variety of computational biology tools and data sets through their website
- Accelerates basic research in the development of cellulosic ethanol and other biofuels and bioproducts, as well as microbial behavior that impact environmental systems

Sponsored by the office of Biological and Environmental Research (BER)

https://jgi.doe.gov/data-and-tools/

» **Advances the scientific challenges related to energy production and environmental systems**

10,154 Data users that downloaded at least one dataset from JGI

10 PB JGI data repository size as of December 2020

Systems Biology Knowledgebase (Kbase)

The Systems Biology Knowledgebase is a **data platform that aims to enable researchers to predict and ultimately design biological function.**

- Knowledge creation and discovery environment designed for biologists and bioinformaticians
- Provides access to public data and more than 300 analysis tools
- Enables users to analyze, share, and collaborate using data and tools designed to help build increasingly realistic models for biological function

KBase
PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Sponsored by the office of Biological and Environmental Research (BER)

https://www.kbase.us/

» Facilitates collaboration and accelerates the pace of scientific discovery

More than 15,000 user accounts

290 TB data and more than 100 new publication citations

Particle Data Group

The Particle Data Group (PDG) is an **international collaboration that provides an authoritative and comprehensive summary of particle physics** as well as related areas of cosmology and astrophysics.

- Provides world averages of particle properties and related quantities
- Resource for measurements of known particle properties
- Provides summaries of searches for new particles
- Reviews of key topics in particle physics and cosmology

Sponsored by the office High Energy Physics (HEP)

https://pdg.lbl.gov/

» **Summarizes and disseminates authoritative information for particle physics and cosmology**

Information about 45,000 measurements from 12,000 publications

More than 6,000 citations per edition, more than 84,000 citations for all editions

National Nuclear Data Center

The National Nuclear Data Center (NNDC) is a **worldwide resource for nuclear data.**

- Lead unit of the U.S. Nuclear Data Program (USNDP)
- Specializes in nuclear structure and low-energy nuclear reactions along with nuclear databases and information technology
- Compilation, evaluation, and dissemination of nuclear properties

Sponsored by the office Nuclear Physics (NP)

https://www.nndc.bnl.gov/

» **Provides current, accurate, and authoritative data for workers in pure and applied nuclear science and engineering**

Consists of 6 widely used databases

OSTI Resources (https://www.osti.gov/)

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

Collection

DOE invest $12B annually with national labs and grantees; R&D results emanate from DOE funding.

Research & Development Results

- Journal articles/accepted manuscripts
- Technical reports
- Conference papers
- Theses/dissertations
- Scientific and technical software
- Datasets
- Patents
- Workshop reports
- Videos

~50k R&D "products" submitted annually to OSTI

DOE Office Of Scientific & Technical Information

Enabling Open Science – Persistent Identifiers

Persistent Identifier Services – the key to interlinking research objects

- Currently assigning digital object identifiers (DOIs) to data objects, software, technical reports, conference posters and presentations
- ORCID IDs – persistent identifiers for researchers; lead US Government ORCID Consortium
- Pilot projects assigning and using persistent identifiers for awards and organizations

Data ID Services

US Government ORCID Consortium

Interagency Data ID Service

Award DOI Service

Discovery

Primary search tool for all DOE-funded R&D results

OSTI.GOV — Over 3.25M R&D records
919K full-text/resource available

DEPARTMENT OF ENERGY DOE PAGES Public Access Gateway for Energy & Science — 124K full-text accepted manuscripts and articles fulfilling DOE Public Access Policy

DOE Data Explorer — 151K data records linking to data repositories

DOE CODE — 3.1K software projects
Provided to Code.gov

DOE Patents — 42.5K patents from 1940s to present
Harvested from USPTO

DOE ScienceCinema — 3.1K videos from DOE and CERN
Speech indexed by IBM Watson

Science.gov — Searches 98+% of all federal science databases

WorldWide Science.org — Searches scientific databases from 70+ countries
Uses Microsoft Translator to search in 10 languages

DOE SC Data Management Overview

SC data management principles		
Enable discovery	Share, preserve, validate	Cost management

Data Management Plan (DMP) requirements			
Share, preserve, validate	Make data associated with publications accessible	Availability of data management resources	Privacy, security, confidentiality

- DMPs are reviewed as part of the overall SC research proposal merit review process
  - Additional requirements and review criteria for the DMP may be identified in a solicitation
  - Proposals may include requested funding to implement a DMP, which will be considered during merit review

Complete information available at: https://science.osti.gov/Funding-Opportunities/Digital-Data-Management

Updates to Digital Data Management Guidance

- Office of Science is updating the **Suggested Elements of a Data Management Plan** (DMP) and adding **Guidance for Reviewers of Data Management Plans**
  - Current guidance will remain in effect for all solicitations issued through December 31, 2021
  - Updated guidance will be effective for all solicitations issued after **January 1, 2022**
  - There are **no changes to formal DMP requirements** that are part of solicitations

Suggested Elements of a DMP

- Suggested Elements offer guidance to researchers about what to include in a DMP
- Provide a framework for planning a DMP that satisfies requirements
- Tool to aid in aligning with best practices in data management

Guidance for DMP Reviewers

- Reviewers are asked if the DMP is suitable and supports validation of the proposed research
- Reviewer guidance connects suggested elements to DMP requirements
- Encourages constructive feedback to continue improving future DMPs

PAMS Updates for Reviewers

- PAMS emails to reviewers will include a link to the Guidance for Reviewers of DMPs
- Reviewers will need to certify once a year that they have read the Guidance for Reviewers of DMPs

Complete information available at: https://science.osti.gov/Funding-Opportunities/Digital-Data-Management

U.S. DEPARTMENT OF
ENERGY | Office of Science