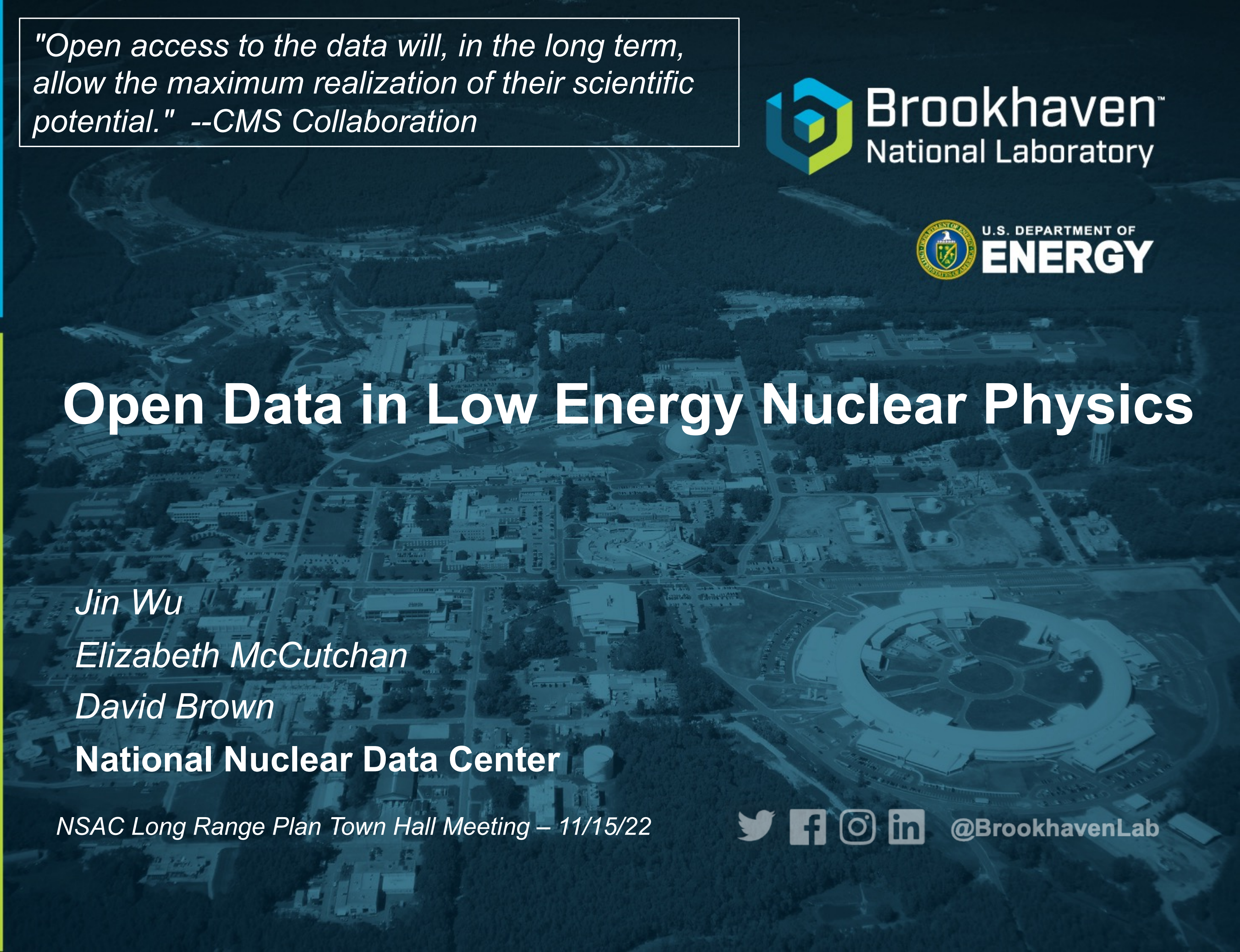


> *"Open access to the data will, in the long term, allow the maximum realization of their scientific potential." --CMS Collaboration*

Brookhaven™
National Laboratory

U.S. DEPARTMENT OF
ENERGY

Open Data in Low Energy Nuclear Physics

Jin Wu

Elizabeth McCutchan

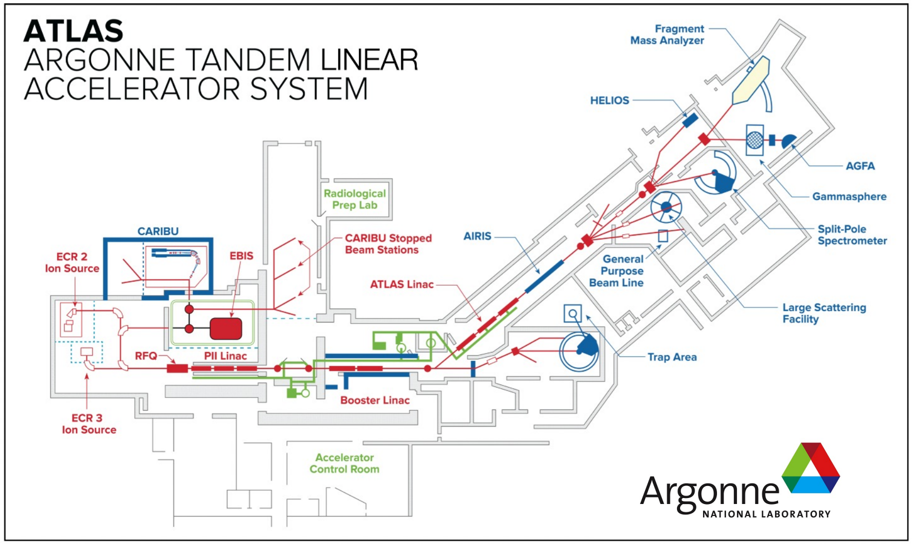
David Brown

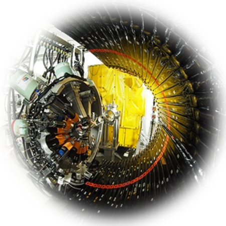
National Nuclear Data Center

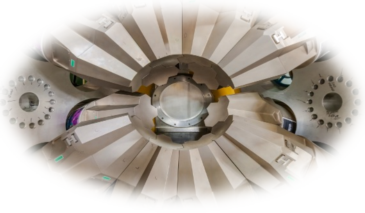
NSAC Long Range Plan Town Hall Meeting – 11/15/22

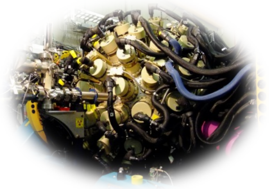
@BrookhavenLab

Where are we?





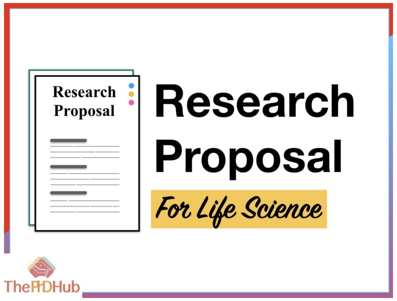




- Large amounts of experimental data will be produced attributing to the powerful accelerators and complex detection systems.

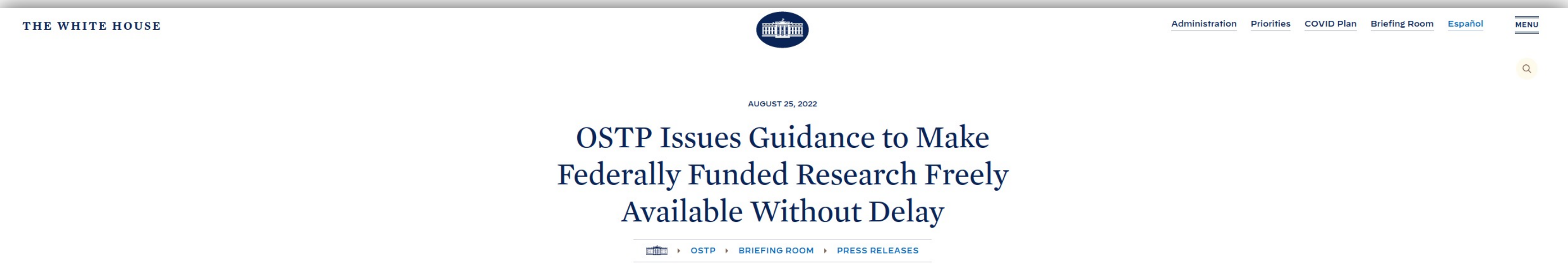
- This “self-curation” by individual research groups lacks uniformity and results in a situation where data discovery and reuse are often difficult or impossible.

Many scientists could do ...



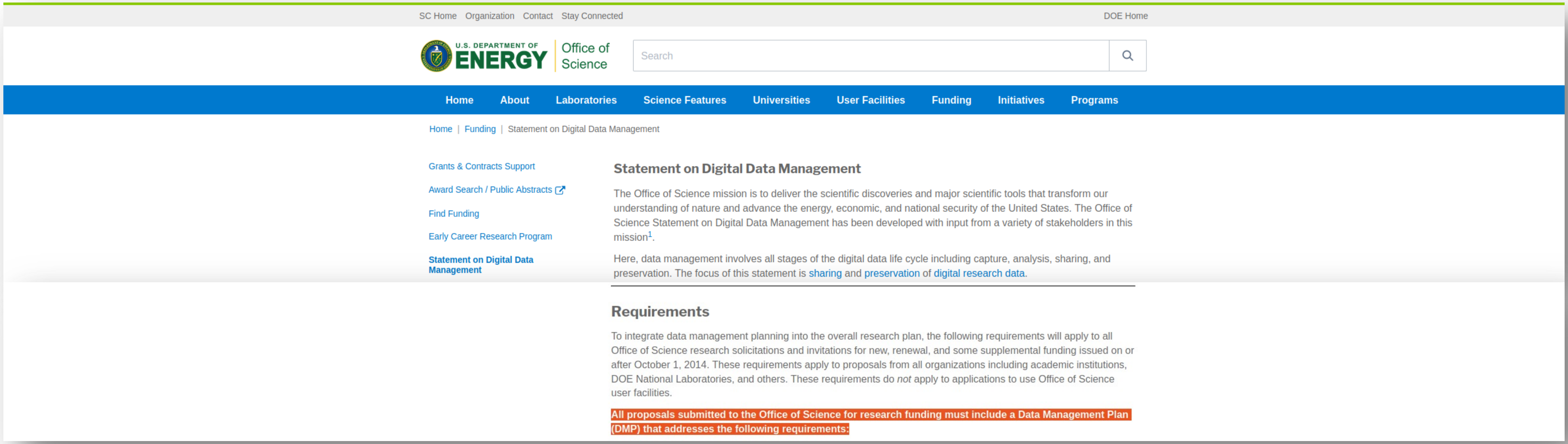
Much, Much, Much Later

Guidance from OSTP and DOE



AUGUST 25, 2022

OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay


SC Home
Organization
Contact
Stay Connected
DOE Home
U.S. DEPARTMENT OF ENERGY
Office of Science
Search
Home
About
Laboratories
Science Features
Universities
User Facilities
Funding
Initiatives
Programs
Home | Funding | Statement on Digital Data Management
Grants & Contracts Support
Award Search / Public Abstracts
Find Funding
Early Career Research Program
Statement on Digital Data Management

Statement on Digital Data Management

The Office of Science mission is to deliver the scientific discoveries and major scientific tools that transform our understanding of nature and advance the energy, economic, and national security of the United States. The Office of Science Statement on Digital Data Management has been developed with input from a variety of stakeholders in this mission[1].

Here, data management involves all stages of the digital data life cycle including capture, analysis, sharing, and preservation. The focus of this statement is sharing and preservation of digital research data.

Requirements

To integrate data management planning into the overall research plan, the following requirements will apply to all Office of Science research solicitations and invitations for new, renewal, and some supplemental funding issued on or after October 1, 2014. These requirements apply to proposals from all organizations including academic institutions, DOE National Laboratories, and others. These requirements do *not* apply to applications to use Office of Science user facilities.

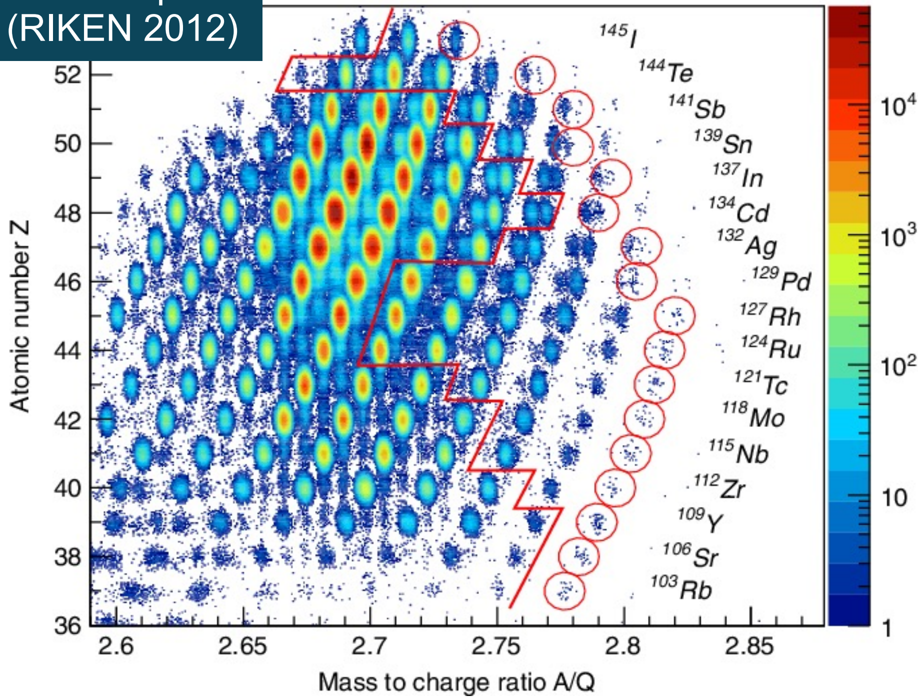
All proposals submitted to the Office of Science for research funding must include a Data Management Plan (DMP) that addresses the following requirements:

What we get from Open Data

- Validation of results
- Rescue / improve / repurpose existing data
- Byproducts – time and grants saving (e.g. fission, fragmentation, deep-inelastic, etc)
- Education
- Dataset citation
- Future experiment guidance
- Implemented ML/AL: Experiment optimization, and data analysis, data management.

110 Isotopes (RIKEN 2012)

G. Lorusso et al, PRL 114, 192501 (2015)

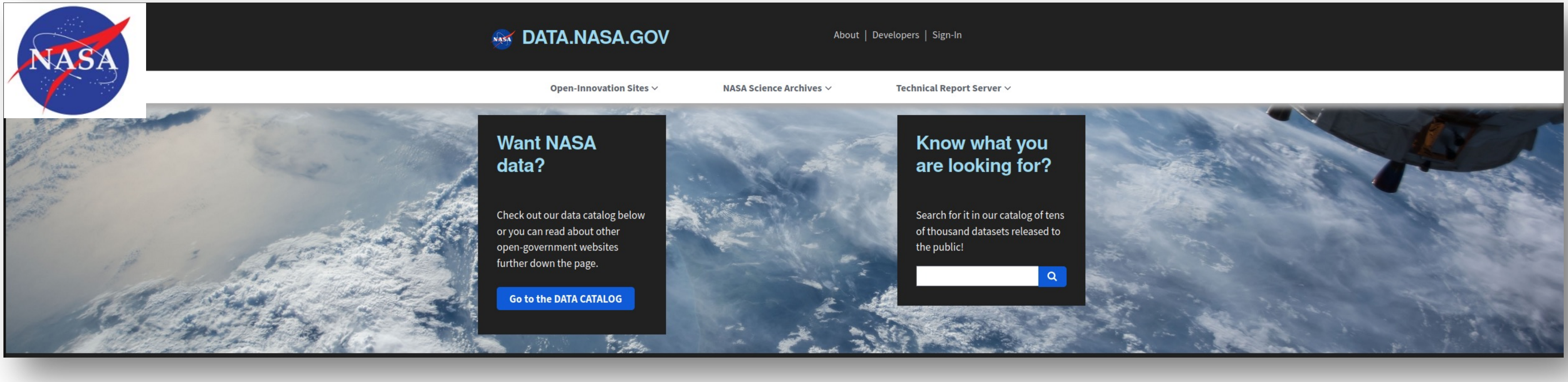


Measured: β-$T_{1/2}$, P_n, β-γ, Isomers, ICe

Published:
β-$T_{1/2}$ (110 Isotopes)
^{126}Pd, ^{128}Pd Isomers
^{126}Pd, β- γ
^{123}Ag, ^{125}Ag: β- γ
^{125}Pd, ^{127}Pd: isomers

Large amounts of byproducts are unanalyzed!

NASA Open Data



DATA.NASA.GOV

About | Developers | Sign-In

Open-Innovation Sites

NASA Science Archives

Technical Report Server

Want NASA data?

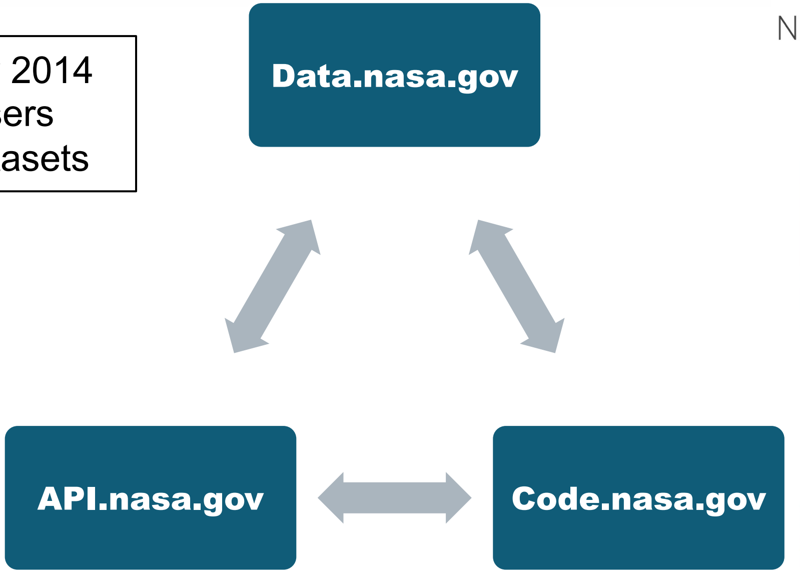
Check out our data catalog below or you can read about other open-government websites further down the page.

Go to the DATA CATALOG

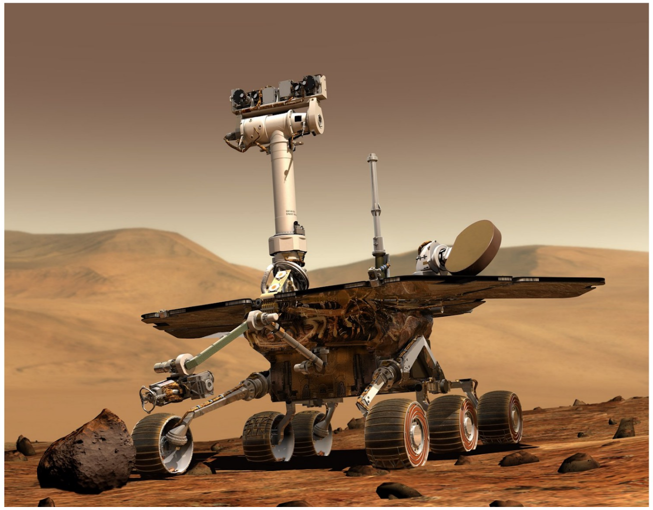
Know what you are looking for?

Search for it in our catalog of tens of thousand datasets released to the public!

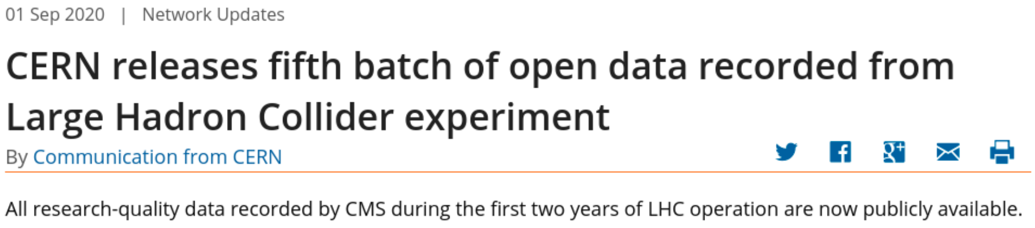
- Started in July 2014
- 20K unique users
- 40K active datasets



Texting robots on Mars using Python, Flask, NASA APIs and Twilio MMS



CERN Open Data

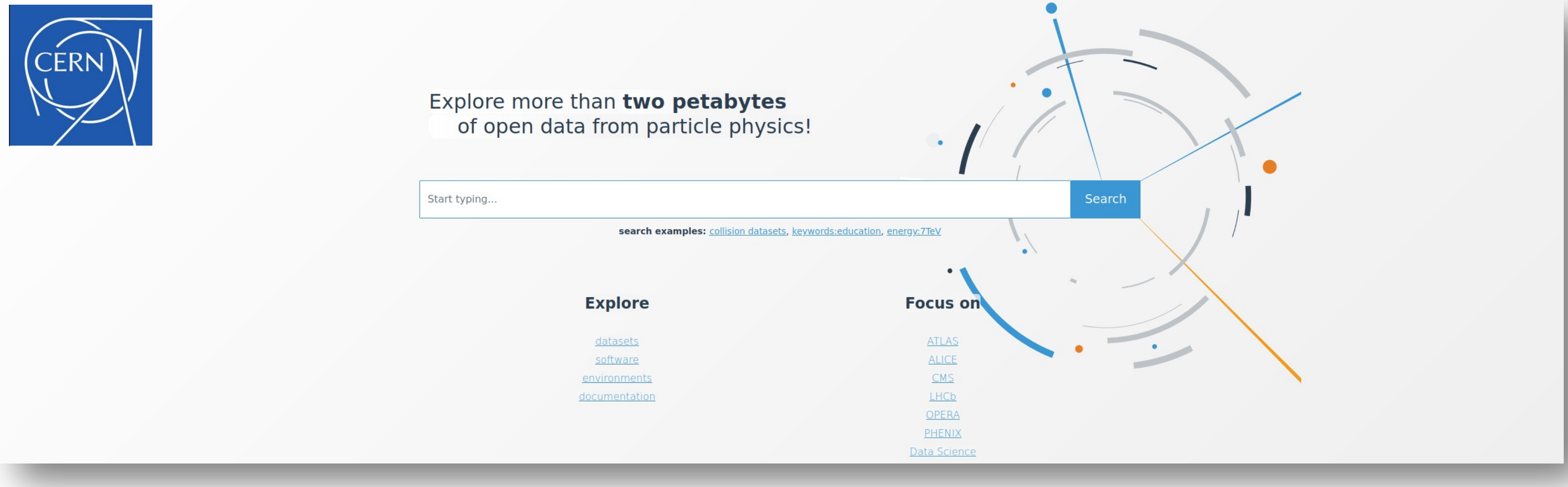


01 Sep 2020 | Network Updates

CERN releases fifth batch of open data recorded from Large Hadron Collider experiment

By Communication from CERN

All research-quality data recorded by CMS during the first two years of LHC operation are now publicly available.



Explore more than **two petabytes** of open data from particle physics!

Start typing... | Search

search examples: collision datasets, keywords:education, energy:7TeV

Explore

- datasets
- software
- environments
- documentation

Focus on

- ATLAS
- ALICE
- CMS
- LHCb
- OPERA
- PHENIX
- Data Science

- >2 petabytes of data available to the public LHC, ATLAS, CMS, ALICE, etc.
- Preprocessed data
- Example source code

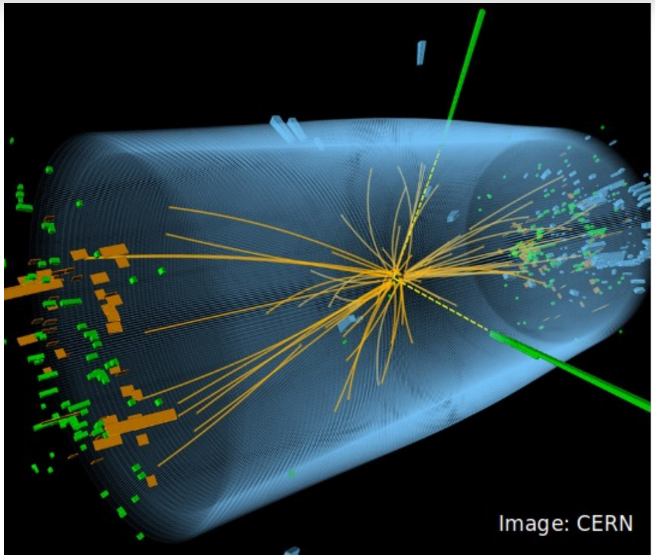
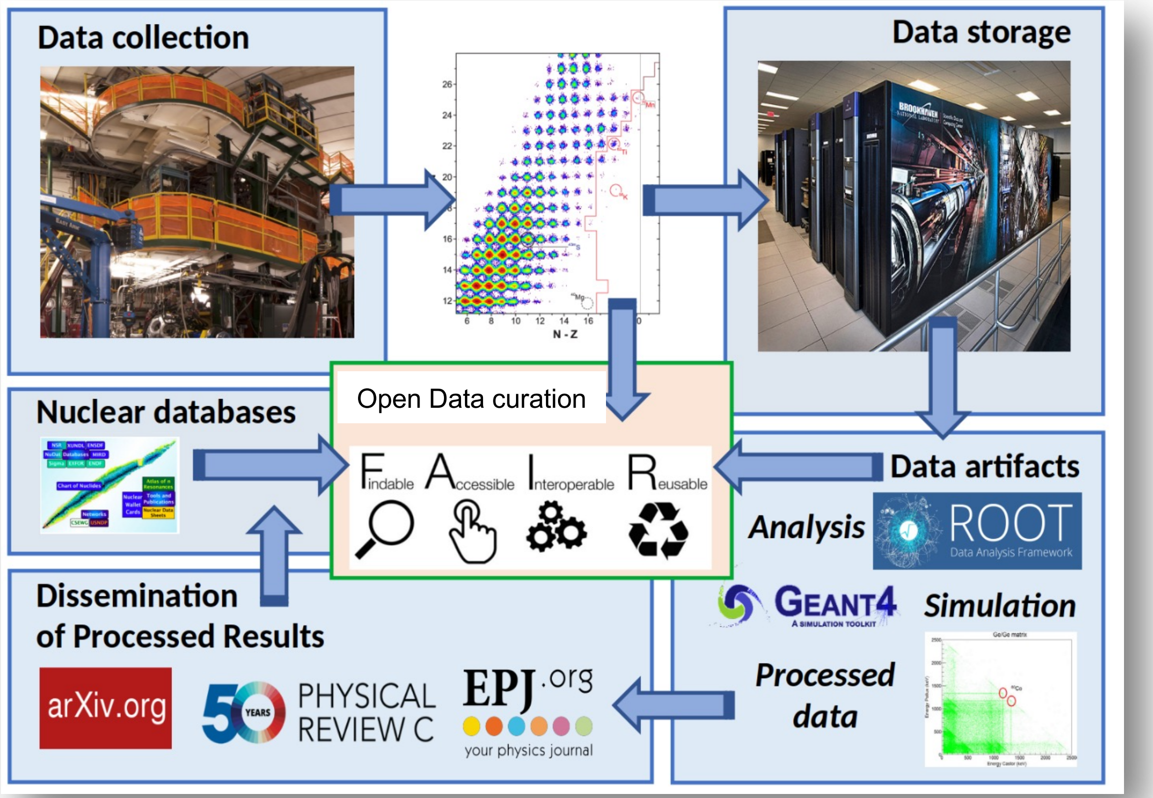


Image: CERN

Modern database technology: curating open data



Credit: Adam Hayes

Acceptance of Proposal

During Experiment

Acceptance of publication

The end of embargo period

Future Perspective

1. Developing & maintaining Open Data website

2. Organizing Community workshops

3. Developing Data/Metadata, Persistent Identifiers system

4. Recruiting future contributors

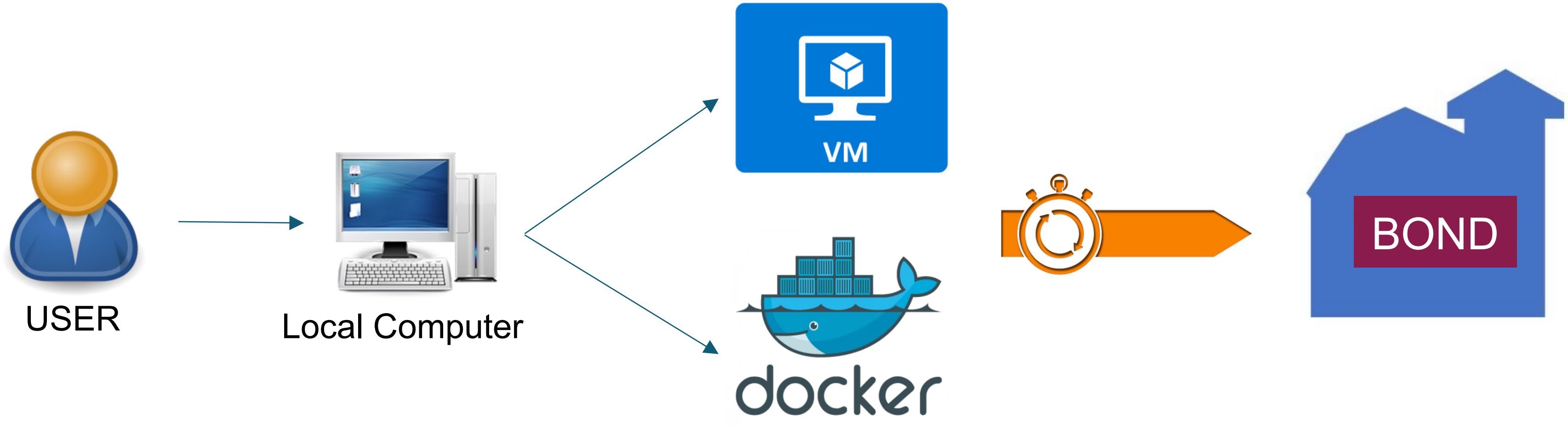
Jin Wu
jwu2@bnl.gov

Elizabeth McCutchan
mccutchan@bnl.gov

David Brown
dbrown@bnl.gov

Thank you!

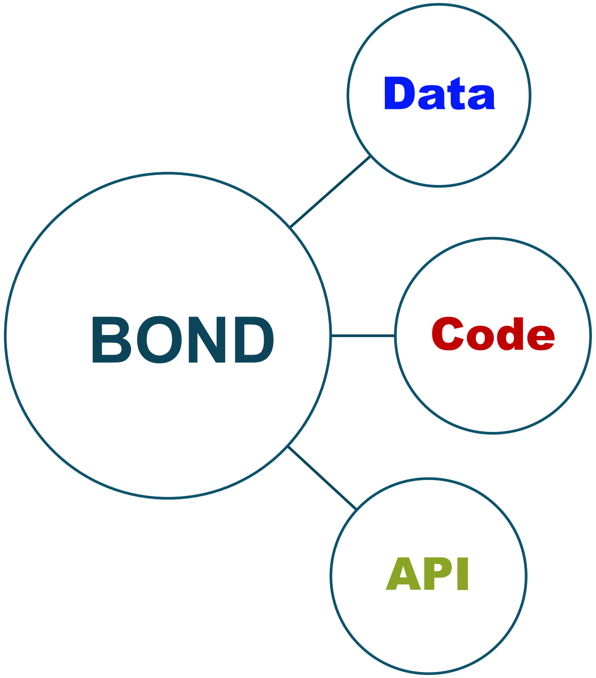
Analyzing data remotely

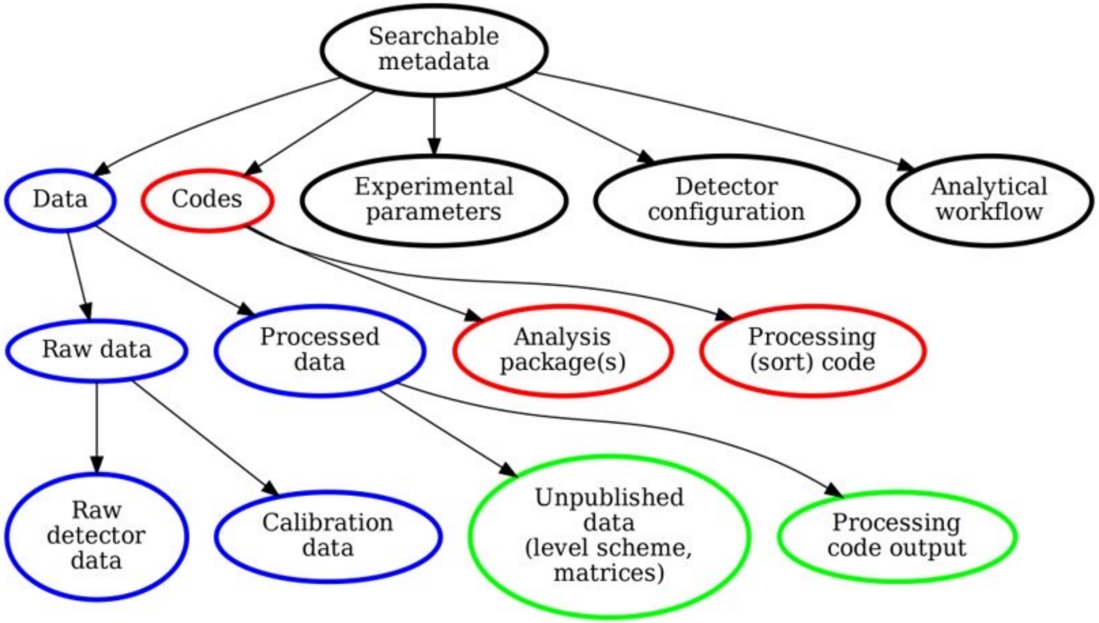


The goal is to remove a need for the installation of the experiment software and to minimize the number of platforms (compiler-OS combinations) on which experiment software needs to be supported and tested.

- Virtual Machine (VM) represents a complete, portable and easy-to-configure user environment for developing and running nuclear data analyses locally, independently of Operating System platforms (Linux, Windows, macOS).
- Docker is one kind of container image, to access the open data directly in your computer locally.

Data Structure





- Using JSON (JavaScript Object Notation) and object-oriented databasing
- To achieve: 1) uniformity, (2) expandability, and (3) support for heterogeneous data/metadata.

Proposed Data Release Policy (part)

- Level 1 data provides more information on published results in publications, such as extra figures and tables

- Level 2 data includes simplified data formats for outreach and analysis training, such as a ROOT file with some simple information.

- Level 3 data comprises calibrated/sorted files together with analysis-level experiment-specific software, allowing to perform complete full scientific analyses

- Level 4 data covers basic raw data (including all the calibration data) with accompanying the associated documentation (log/E-log, proposal, etc) and the full analysis code, as well as all the results analyzed and published.

Level 1-2: After major publication(s), 2-3 years
Level 3: 3-5 years
Level 4: > 5 years (Max. 10 years)

Each stage needs to be approved by the board of collaborators.

QUIET PLEASE
BOARD
MEETING
in progress
It's settled. We put out a press release saying we can't remember what we decided.
Nicholson
25 April '09

The "FAIR" data principle

1. **Findable** with a globally-unique identifier and rich metadata
2. **Accessible** through a free and standard protocol
3. **Interoperable** with as much standardization as possible
4. **Reusable** with accurate and rich provenance metadata

Findable	Accessible	Interoperable	Expansionable and Reusable
• Metadata • Persistent Identifiers (e.g. DOI) • Data Management Plan	• Web portal • APIs • Data embargo	• Workflow starting from acceptance of proposal	• Policy • Expansionable • Infrastructure