Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

NNDC Public Reusable (PuRe) Data

Alejandro Sonzogni

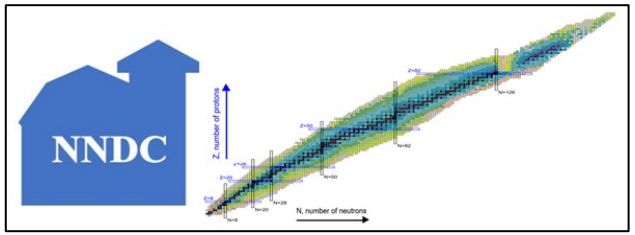
@BrookhavenLab

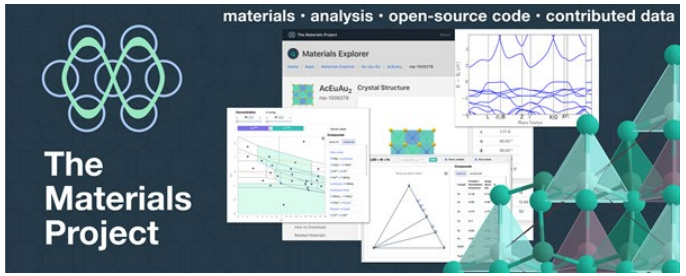
Public Reusable Research (PuRe) Data

Public Reusable Research (PuRe) Data is a designation for key data repositories, knowledge bases, analysis platforms, and other activities that strive to make data publicly available to advance scientific or technical knowledge. Spanning the range of the DOE Office of Science (SC) mission, these data resources include a data center for atmospheric data and model products, data repositories and knowledge bases for biological and environmental research, and a materials database for physical sciences. Each resource is an authoritative provider of data or capabilities in their respective subject area. Together, these high-quality public resources play a strategic role in advancing the SC mission while making data easier to find, access, and reuse across the broader scientific community.

Designation as a PuRe Data Resource does more than simply recognize the importance of these investments -- it carries the weight of SC stewardship. SC manages these resources under an oversight model with high standards for data management, resource operations, and scientific impact. The designated PuRe Data Resources go above and beyond the standard SC requirements for data management plans and act as community leaders in data stewardship.

https://science.osti.gov/Initiatives/PuRe-Data





ARM

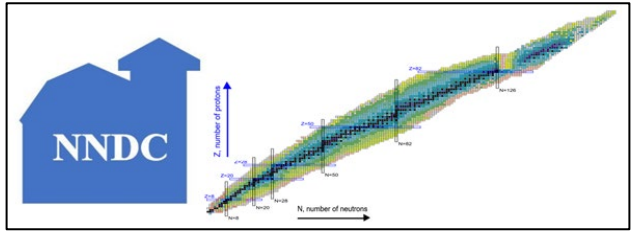
JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

PDG
particle data group

Brookhaven National Laboratory

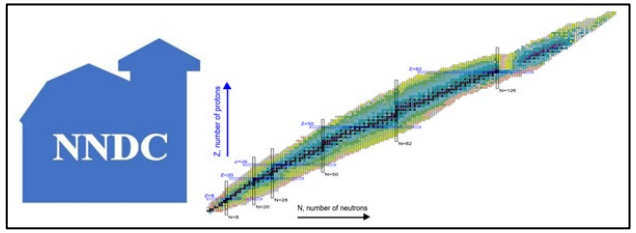
Public Reusable Research (PuRe) Data



An SC PuRe Data Resource has the following characteristics:

- An SC research program office directly supports and oversees an institution or institutions to sustain and manage the Resource.
- The Resource is considered to be an authoritative provider of data or capabilities in its subject area.
- The Resource is publicly available. The terms of access, such as licenses, are as open as possible.
- The Resource appropriately uses DOIs to facilitate discovery, reuse, and citation of data and software.
- The Resource provides sufficient information for a potential user to determine whether the resource is appropriate for their purposes.
- The supporting SC program office ensures there is regular peer review of the management, operations, and scientific impact of the data resource.
- There exists an SC- approved Data Management Plan (DMP) that holistically covers the Resource and that serves as a compact between the supporting SC program office and the managing institution or institutions, regarding the management of the Resource. At a minimum, the DMP must address:
  - A description of the data used and/or generated;
  - Data and/or software sharing agreements, licensing, and other terms of access and reuse;
  - Long- term preservation and stewardship of data and relevant software, including plans for handing data off to a preservation service;
  - Protection of data and code including security and integrity; and
  - Accountability of the data management plan.

Public Reusable Research (PuRe) Data



Timeline:

November 20, 2020, e-mail received to submit material by December 2, 2020.

December 11, 2020, additional information requested, which was provided on December 16, 2020.

Submitted our operations and plans to address issues such as

- Long-term data preservation
- DOI usage
- Adherence with FAIR principles, Findability, Accessibility, Interoperability, and Reusability.
- NNDC Diversity and Inclusion.
- ORCID Profiles
- Data and/or software sharing agreements, licensing, and other terms of access and reuse
- Better interaction with OSTI

Made a presentation to OSTI on February 12, 2021, and have been participating on DOE DATA ID monthly meetings.

NNDC Data Management Plan

- NNDC servers and cluster housed by RHIC-Atlas computing facility, benefitting from security, air-conditioning, clean power.
- Database, web and GitLab servers backed up weekly by ITD and Ramon Arcilla.

- As mentioned yesterday, for data preservation in case an asteroid hits Long Island, we have contracted Amazon Web Services to backup **Web, database& GitLab servers, 15 TB**.
- $18K / year.
- Considerable amount of paperwork!

On an unrelated note, the NNDC cluster now has 432 cores and much more RAM, thanks to an NNDC and NE investment. All made possible by Ramon Arcilla's work during the pandemic.

Brookhaven National Laboratory

DOI implementation & Permanent Landing Pages

- Data Object Identifiers, DOIs, are unique alphanumeric strings to identify content and provide a persistent link to its location on the internet (permanent landing page). Provided by OSTI through BNL's research library.
- Permanent landing page must provide:
  - Authors with affiliations and emails, as well as ORCID, a persistent digital identifier that you own and control, and that distinguishes you from every other researcher, to avoid name ambiguity.
  - Metadata.
  - Data in raw and graphical format.
  - API, or data in JSON format.
- It needs approval from BNL's cybersecurity chief and BNL's central library.
- Working pages should be available in December for ENSDF and XUNDL.
- A DOI will be assigned for every adopted, decay and reaction dataset. If dataset is updated, DOI stays the same. If dataset is re-evaluated, new DOI is issued, previous landing page stays.
- ENSDF & XUNDL will provide a learning experience to continue with EXFOR & ENDF/B.

Author's names in ENSDF files

- There is no consistency in the way authors names are given in ENSDF
- Some observed possibilities: 'A. Sonzogni', or 'A.A. Sonzogni', 'Alejandro Sonzogni', 'Sonzogni', plus their upper-case variations.
- We would like to give (i) affiliations for all authors, (ii) e-mail for corresponding author(s), (iii) contact e-mail, and (iv) optional ORCID.

```
144SM    ADOPTED LEVELS, GAMMAS                                    01NDS    200108
144SM  H TYP=FUL$AUT=A.A. Sonzogni$CIT=NDS 93, 599 (2001)$CUT=1-Dec-2000$
144SM  XA144EU EC DECAY
           .
           .
           .
           .
```

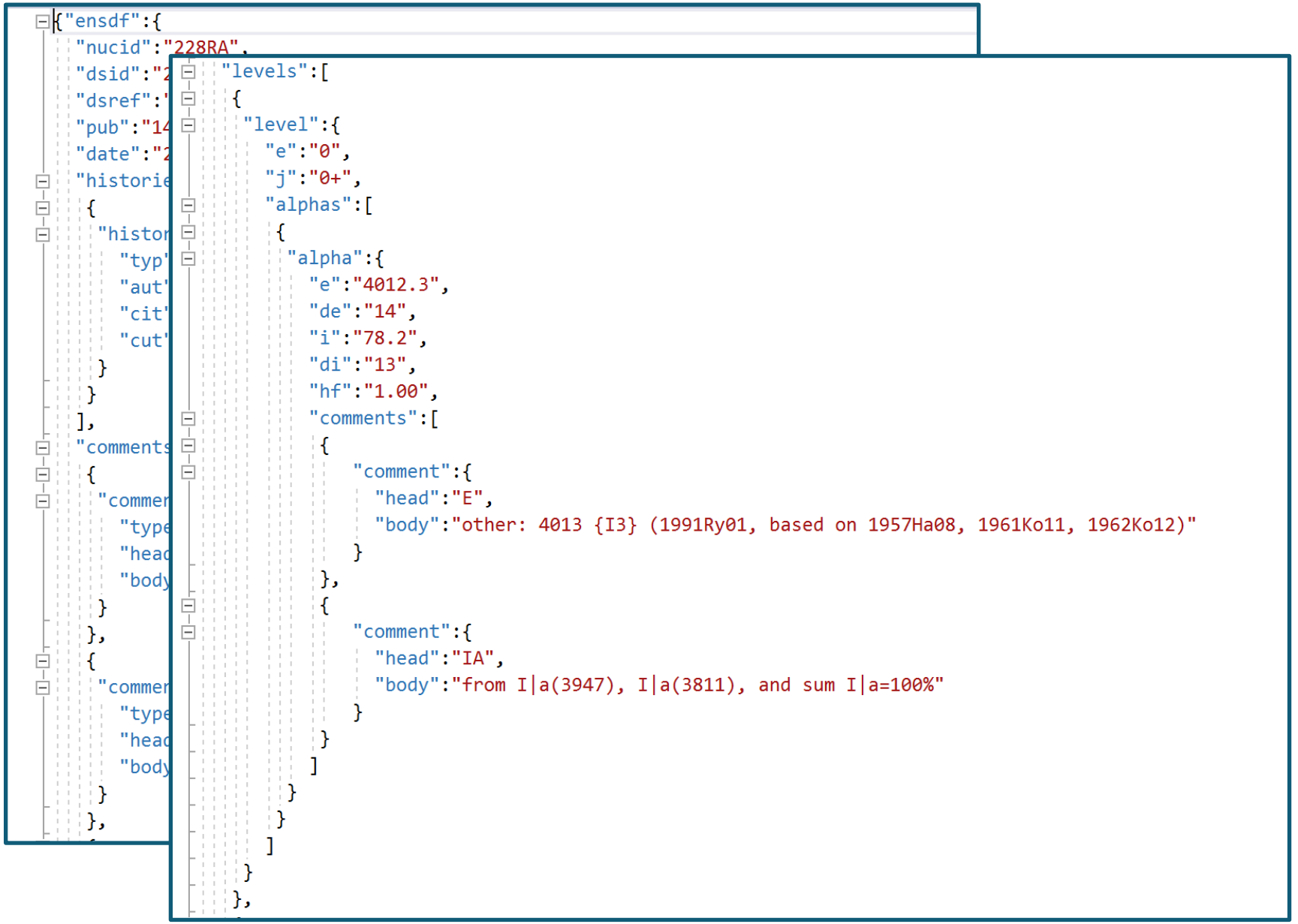
Suggest expanding the History record in the legacy ENSDF format and modify FMTCHK for the additional mandatory continuation records with their proper formats.

```
144SM    ADOPTED LEVELS, GAMMAS                                          01NDS    200108
144SM  H TYP=FUL$CORR_AUT=A.A. Sonzogni$E-MAIL=sonzogni@bnl.gov$
144SM 2H AFF=Brookhaven National Laboratory$ORCID=0000-0002-9438-6033$
144SM 3H AUT=J. Doe$AFF=Brookhaven National Laboratory$
144SM 4H ORCID=0000-0002-9438-6033$
144SM 5H CIT=Nuclear Data Sheets 93, 599 (2001)$CUT=1-Dec-2000$
144SM 6H CONTACT-EMAIL=ensdf@bnl.gov$
144SM  XA144EU EC DECAY
           .
           .
           .
           .
```

Brookhaven National Laboratory

JSON format for the Legacy ENSDF

- A JSON format for the legacy ENSDF format was developed.
- The implementation of an alternative XML format was a struggle due to '<' and '>' signs used profusely in ENSDF.
- For the naming convention, it follows the manual 100%.
- It frees us from providing an API to process ENSDF files.
- With help with Donnie Mason!



Brookhaven National Laboratory

Digitizing the NNDC library

As part of the PuRe designation, we have work with BNL's research library to:

- ❑ Identify reports, journal or books not available anywhere else.
- ❑ Submit reports' PDF files not available in OSTI – which requires access to NSR's MariaDB database.
- ❑ Inspect every single report, if available online then save the PDF and discard paper copy. If not, scan and then discard.
- ❑ Hired Catherine Dunn in August 2021 for this task with NNDC funds.
- ❑ It seems that only LANL has digitally preserved their reports.
- ❑ All material will be incorporated in NSR.
- ❑ Already data found in several reports are of interest, will present that next week in CSEWG.