PNNL-34266

Data Archive and Portal (DAP) Platform for Solid Phase Processing Technologies

April 2023

Mohammad Fuad Nur Taufique
Matthew Macduff
Shuhao Bai
Devin McAllester
Osman Mamun
Jing Wang
Michael Kieburtz
Ram Devanathan

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, **makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from
the Office of Scientific and Technical
Information,
P.O. Box 62, Oak Ridge, TN 37831-0062**
www.osti.gov
**ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov**

**Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov**

PNNL-34266

Data Archive and Portal (DAP) Platform for Solid Phase Processing Technologies

April 2023

Mohammad Fuad Nur Taufique
Matthew Macduff
Shuhao Bai
Devin McAllester
Osman Mamun
Jing Wang
Michael Kieburtz
Ram Devanathan

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

The scale and speed of data generated by modern scientific experiments have constantly challenged the research community to store, curate, manage and optimally use it to drive scientific discoveries. In this work, we have developed a data archive and portal (DAP) platform including analytics capabilities to collect, curate, and manage data and metadata stream for solid phase processing (SPP) techniques. We successfully hosted around ~347K files of data related to processing parameters, microscopic images, and spectroscopic data related to solid phase processing. The DAP platform for SPP will establish an enduring capability to support machine learning and grow collaboration at the intersection of materials science and data science.

Summary

The Pacific Northwest National Laboratory generates a large amount of data every day from various research projects related to solid phase processing science (SPP) technologies such as Shear Assisted Processing and Extrusion (ShAPE), friction stir processing, friction stir welding, and cold spray. The data types include processing parameters, characterization data, and resulting mechanical testing data. To build a robust and reliable machine learning model, a high-quality dataset with proper metadata is required. However, the lack of a large set of high-quality data is a challenge in materials science. Therefore, it is advised to build a well-organized data management and curation infrastructure for materials data by incorporating the FAIR (Findable, Accessible, Interoperable, Reusable) data principles. Currently, there is no database that captures the data and metadata stream for SPP techniques, such as friction stir welding and ShAPE. To address this issue, the Pacific Northwest National Laboratory developed a data management and collaboration portal to collect, curate, and store the results of experiments and simulations, develop new scientific insights with data analytics tools, and integrate SPP research efforts. This work demonstrates proof of concept that can lead to a sustained data management framework with access control, offering researchers unprecedented access to experimental and modeling data from multiple SPP projects.

Acknowledgments

This research was supported by the Solid Phase Processing Science Initiative, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Contents

Figures

Tables

1.0 Introduction

Every day huge amount of data is generated from different research projects related to computational and experimental materials science under the umbrella of solid phase processing science initiatives (SPPSi) at the Pacific Northwest National Laboratory (PNNL). Solid phase processing (SPP) technologies available at PNNL are Shear Assisted Processing and Extrusion (ShAPE), friction stir processing, friction stir welding, and cold spray. The types of data vary from project to project, which include processing parameters from machine to machine, corresponding characterization such as, image data (SEM/TEM/OM), spectral data (XRD, Raman, Synchrotron), tabular data, numerical data, and resulting mechanical testing data, such as hardness, yield strength, Young's modulus etc. Each of the datum is generated from different machines at diverse facilities and associated with different categories of metadata. This huge stream of data including process parameters, characterization and mechanical tests has the potential to act as input parameters to build a decision-making tool by employing artificial intelligence (AI) and machine learning (ML)models. To build a robust, reliable and physics-informed machine learning model it is important to work on a high-quality dataset that includes all the data and metadata for a given set of the target property. To the best of our knowledge, most of the machine learning work related to materials science are challenged by the lack of a large set of high-quality data, and hence data are pre-processed manually to build the machine learning model from different literature or third-party data sources [1,2,3,4]. Sometimes published literature and third-party data sources have insufficient metadata and provenance, which are critical for supporting the FAIR (Findable, Accessible, Interoperable and Reusable) principles for reproducible science. To overcome the issues related to high-quality datasets, it is advised to build a well-organized data management and curation infrastructure for materials data by incorporating the FAIR data principles [5,6,7, 8].

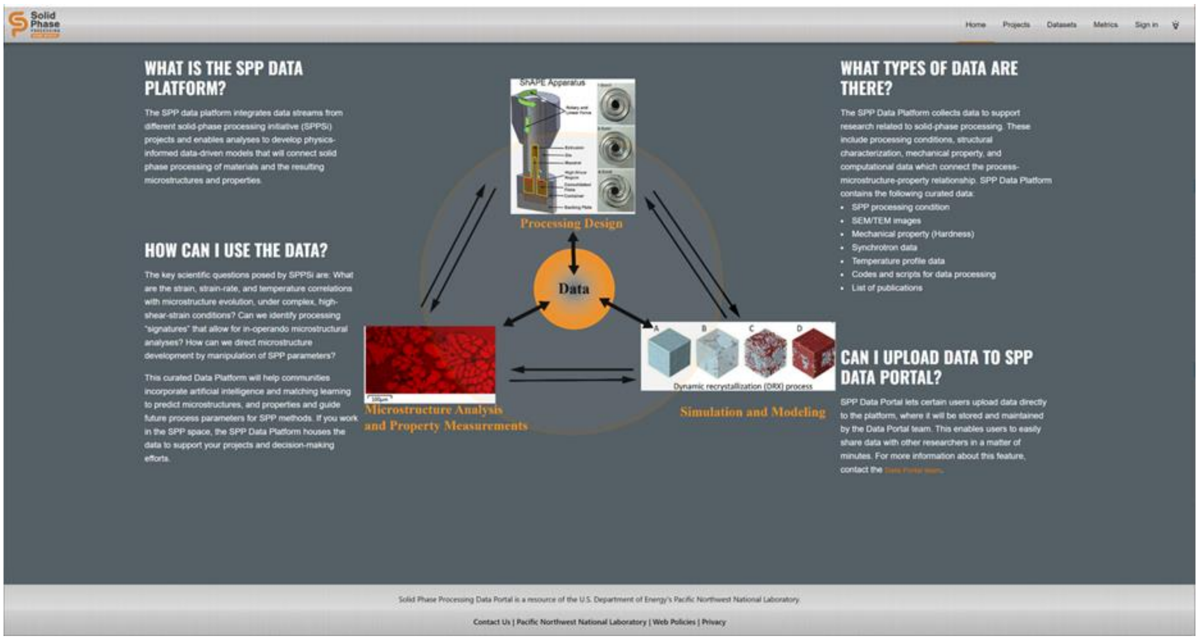
There are many efforts underway to manage and curate valuable scientific data related to materials research as presented in Table 1. However, to the best of our knowledge, there is no database that captures the data and metadata stream for SPP techniques, such as friction steer welding and shear-assisted extrusion and processing (ShAPE) [9].

Table 1. List of a few popular materials databases. A details list is provided by Lauri et. al. [4]

Name of the database	Link to the website	Focused area	Country of origin
The materials project	https://materialsproject.org	High throughput electronic structure calculations of single crystals, typically at 0 K.	USA
Materials Commons	http://www.prisms-center.org/#/mcommons/overview	Metallic alloy data for microstructural evolution and mechanical properties	USA
The Materials Data Facility	https://materialsdatafacility.org/	Metallic alloy data set	USA
Khazana	https://khazana.gatech.edu	Polymer data	USA

SUNCAT	https://suncat.stanford.edu/	Catalysis data	USA
MatNavi	https://mits.nims.go.jp/en	Polymer and metallic data	Japan
Citrine	https://citrine.io/	Materials and chemicals	USA
AFLOW	aflowlib.org	Computational data consists of 3,528,653 material compounds with over 733,959,824 calculated properties and growing	USA

Currently, new materials knowledge generated resides almost entirely with individual researchers and only a small fraction of the data gathered is made available through publications and presentations. A large portion of this valuable data is locked away and sometimes not analyzed even by the data generator. The data generated by a project is currently invisible to other projects and is not even centrally archived resulting in investigators reinventing the wheel or going down previously encountered dead ends. To solve this issue, we developed a data management and collaboration portal with analytics capability included to collect, curate and store the results of experiments and simulations, develop new scientific insights with data analytics tools, and integrate SPP research efforts. It is important to mention that SPP data portal is available to the PNNL's employees only at this stage. A snapshot of the current data archive and platform (DAP) is presented in Figure 1. This work demonstrates proof of concept that can lead to a sustained data management framework with access control. It will offer researchers unprecedented access to experimental and modeling data from multiple SPP projects and can be adapted to include literature data as well.



WHAT IS THE SPP DATA PLATFORM?

The SPP data platform integrates data streams from different solid-phase processing initiative (SPPSi) projects and enables analyses to develop physics-informed data-driven models that will connect solid phase processing of materials and the resulting microstructures and properties.

HOW CAN I USE THE DATA?

The key scientific questions posed by SPPSi are: What are the strain, strain-rate, and temperature correlations with microstructure evolution, under complex, high-shear-strain conditions? Can we identify processing "signatures" that allow for in-operando microstructural analyses? How can we direct microstructure development by manipulation of SPP parameters?

This curated Data Platform will help communities incorporate artificial intelligence and matching learning to predict microstructures, and properties and guide future process parameters for SPP methods. If you work in the SPP space, the SPP Data Platform houses the data to support your projects and decision-making efforts.

WHAT TYPES OF DATA ARE THERE?

The SPP Data Platform collects data to support research related to solid-phase processing. These include processing conditions, structural characterization, mechanical property, and computational data which connect the process-microstructure-property relationship. SPP Data Platform contains the following curated data:

- SPP processing condition
- SEM/TEM images
- Mechanical property (Hardness)
- Synchrotron data
- Temperature profile data
- Codes and scripts for data processing
- List of publications

CAN I UPLOAD DATA TO SPP DATA PORTAL?

SPP Data Portal lets certain users upload data directly to the platform, where it will be stored and maintained by the Data Portal team. This enables users to easily share data with other researchers in a matter of minutes. For more information about this feature, contact the Data Portal team.

Solid Phase Processing Data Portal is a resource of the U.S. Department of Energy's Pacific Northwest National Laboratory.

Contact Us | Pacific Northwest National Laboratory | Web Policies | Privacy

Figure 1. The front page of DAP for the SPP projects.

2.0 SPP Data Management Workflow

The SPP data management workflow consists of three steps. These steps are:

1. Data collection
2. Data Curation
3. Data archive and portal platform

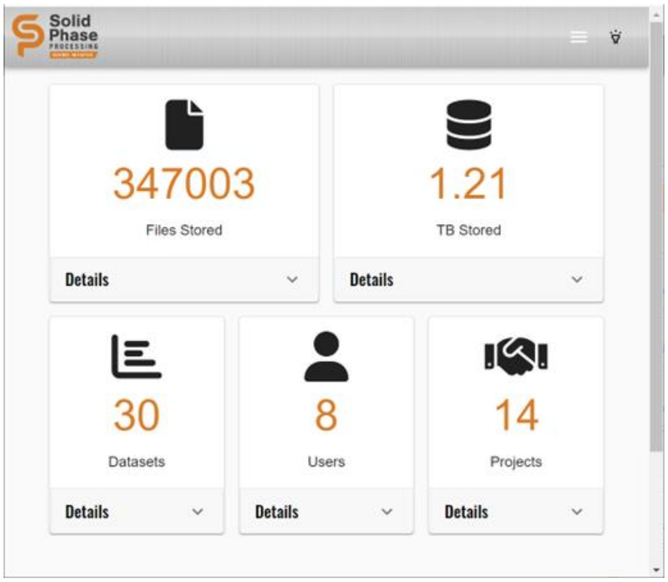
Details explanation of each step is given below.

2.1 Data Collection

We collected data relevant to SPP by capturing the knowledge of researchers to identify the datasets that are available and the key parameters that need to be captured. We then collected literature data, experimental data and simulation results from SPP projects. Materials data are stored in a variety of formats, with varying levels of quality. We adapted established workflows to collect data from sources, digitize alloy data as needed, and store the data in an appropriate format. We developed a data template for different types of data that will capture the appropriate metadata with the data. The template initially includes the key parameters identified by domain experts in SPP and will be iteratively modified to keep up with the evolution of SPP. Data collection is the first and crucial step in data management and it will be an ongoing exercise for the life of the project.

2.2 Data Curation

In this step, we established data provenance, attached metadata and assessed quality. We have collected, curated and managed around ~347K data files corresponding to 30 individual datasets relevant to SPP as presented in Figure 2.



347003
Files Stored
Details

1.21
TB Stored
Details

30
Datasets
Details

8
Users
Details

14
Projects
Details

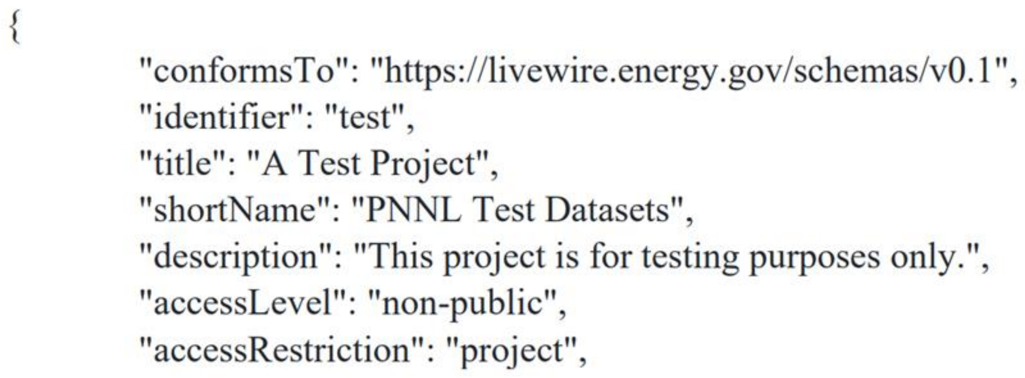
Figure 2. Data metrics with respect to filing storage and datasets for DAP.

2.3 Data Archive and Portal Platform

We developed a SPP Data Archive and Portal to support data sharing and knowledge discovery. This includes supporting the FAIR principles (Findable, Accessible, Interoperable, Reusable). This relates to a variety of features in the DAP that enable effective searching, publishing, access control, curating, visualizing, ordering, and download. These features are all driven by a metadata structure at the core of the DAP. The following section describes that structure and use.

Project Open Data: The core of the metadata structure is based on a standard known as DCAT-US Schema v1.1 or Project Open Data (https://resources.data.gov/resources/dcat-us/). This schema describes a structure using the JSON notation that includes 3 basic items in a hierarchy. Catalogs contain Datasets that contain Distributions. Each level has some key metadata fields that are commonly used such as a Title or Description or Contact. Because the format is just a JSON structure it is easily extended for the needs of a project or science domain. Wherever possible, the DAP makes use of the Project Open Data standards and extends it without breaking what is part of the standard. Using this approach provides for the ability to enforce compliance with a standard that ultimately enables greater sharing to metadata with other data repository and search portals.

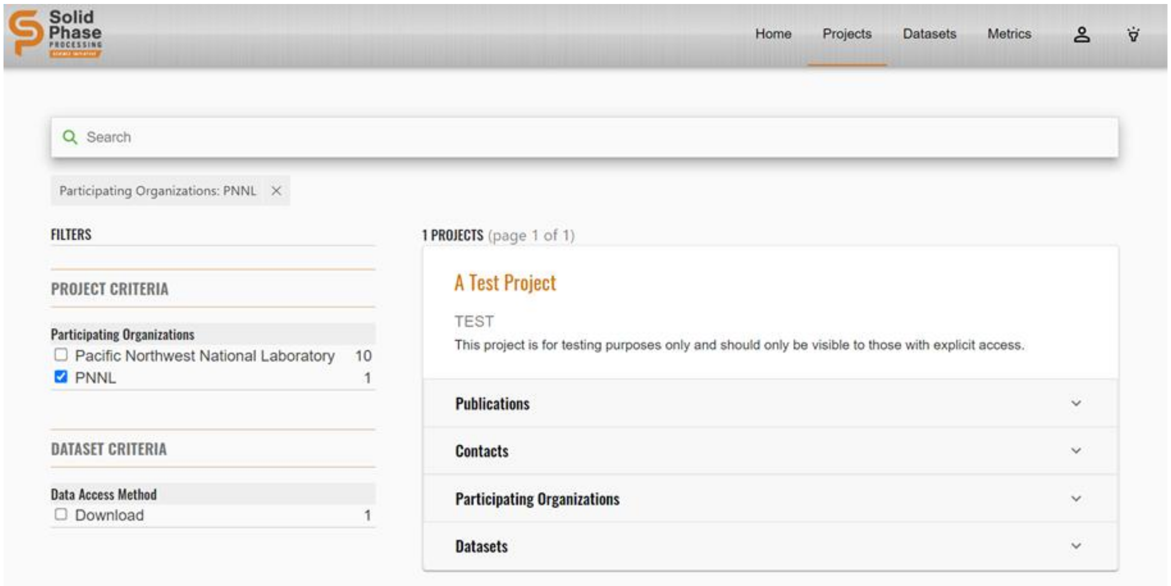
Catalogs (aka Projects): The top of the metadata structure is a Catalog. In the DAP, these are displayed as “Projects” on the website but the structure is still that of a Catalog. An example of some of this structure is shown in Figure 3.



```
{
        "conformsTo": "https://livewire.energy.gov/schemas/v0.1",
        "identifier": "test",
        "title": "A Test Project",
        "shortName": "PNNL Test Datasets",
        "description": "This project is for testing purposes only.",
        "accessLevel": "non-public",
        "accessRestriction": "project",
```

Figure 3. An example of the catalogue structure.

The metadata includes items that would appear on a web page to enable a user to understand more about the project, but also provide directions to the DAP system on how to manage and connect the data. In the example, the data is restricted to only members of the project (which is managed through a separate user management process). The Project level metadata can be indexed for searching as well as include facets to enable discovery on the Project web page. In the example below presented in Figure 4, projects from “PNNL” are displayed. Additional arbitrary extensions to the project metadata can be displayed as part of the project as in the ‘Publications’ section below.



Participating Organizations: PNNL ×

FILTERS

PROJECT CRITERIA

Participating Organizations

☐ Pacific Northwest National Laboratory	10
☑ PNNL	1

DATASET CRITERIA

Data Access Method

☐ Download	1

1 PROJECTS (page 1 of 1)

A Test Project

TEST

This project is for testing purposes only and should only be visible to those with explicit access.

Publications

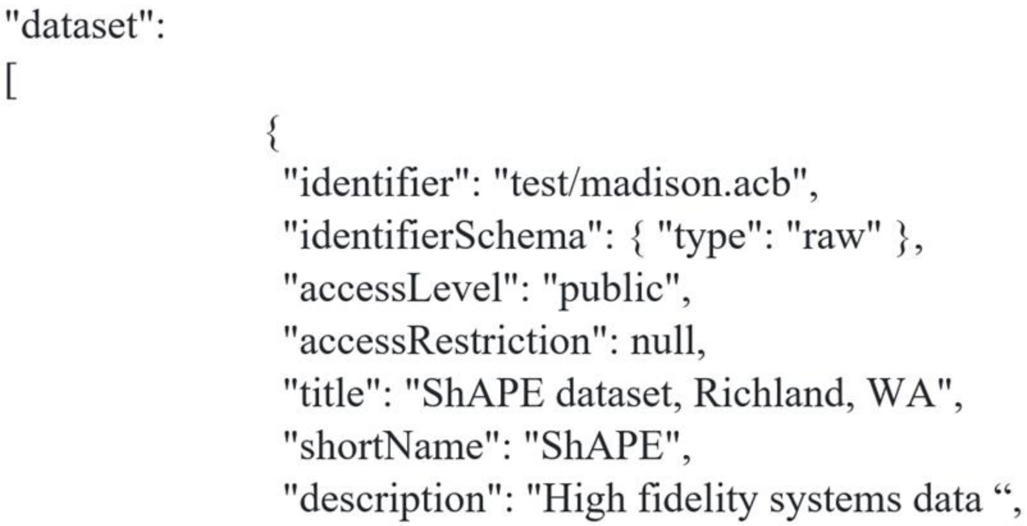
Contacts

Participating Organizations

Datasets

Figure 4. A project catalog for the DAP.

Datasets: The key object in the DAP repository is the Dataset. It is where a particular collection of data is fully described and ultimately enables access to the data. Future users of the data will likely only have the information described in the dataset to understand the data. Having sufficient fields of metadata enable useful searching. The basic JSON structure for a Dataset is simple and similar to a project. An example of a portion of a metadata record is shown in Figure 5.

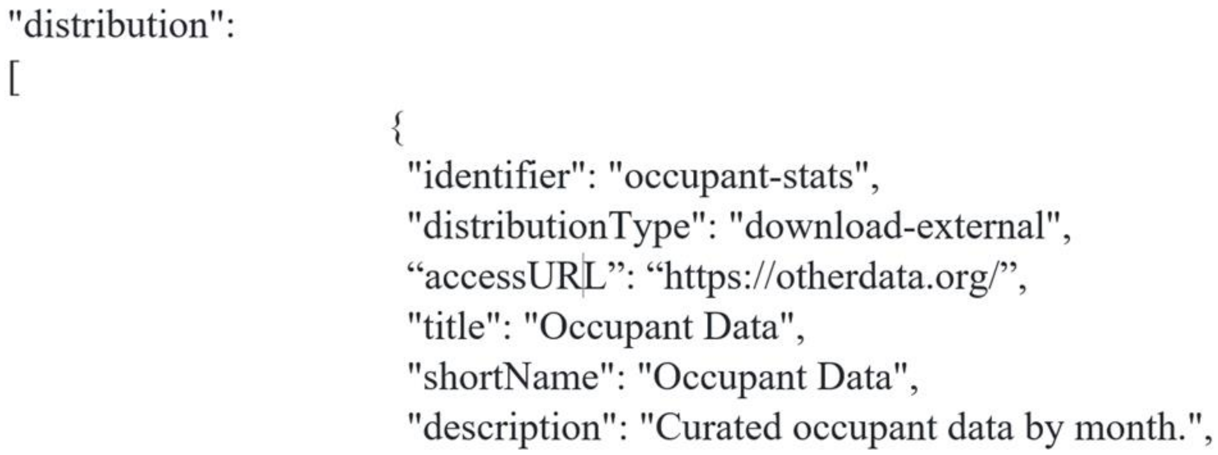


```
"dataset":
[
                {
                 "identifier": "test/madison.acb",
                 "identifierSchema": { "type": "raw" },
                 "accessLevel": "public",
                 "accessRestriction": null,
                 "title": "ShAPE dataset, Richland, WA",
                 "shortName": "ShAPE",
                 "description": "High fidelity systems data “,
```

Figure 5. An example of the metadata record.

Simple searching on keywords and descriptions is in the core DAP platform. But significantly, any addition to the dataset metadata structure can be added to the search and facets that people use to discover data. This means that domain specific metadata about a dataset can be reliably added to the search facets just like it was for a project. Many domains have a standard vocabulary that can be inserted into the metadata. In the case of SPP, this could include the type of material used or the instrument used. These metadata are then available to the user, included in facetted searches and displayed with the dataset in a consistent way for ease of use.

Distribution: For data access, a dataset can have one or more ‘distributions’. This could be a simple list of links to various files to download. However, as with the other parts of the schema this is also extensible. An example in the Project Open Data standard is to access the data through an API. In that case metadata fields would point to documentation and even software to enable interaction with the data. Figure 6 provides an example to access the data on another site.

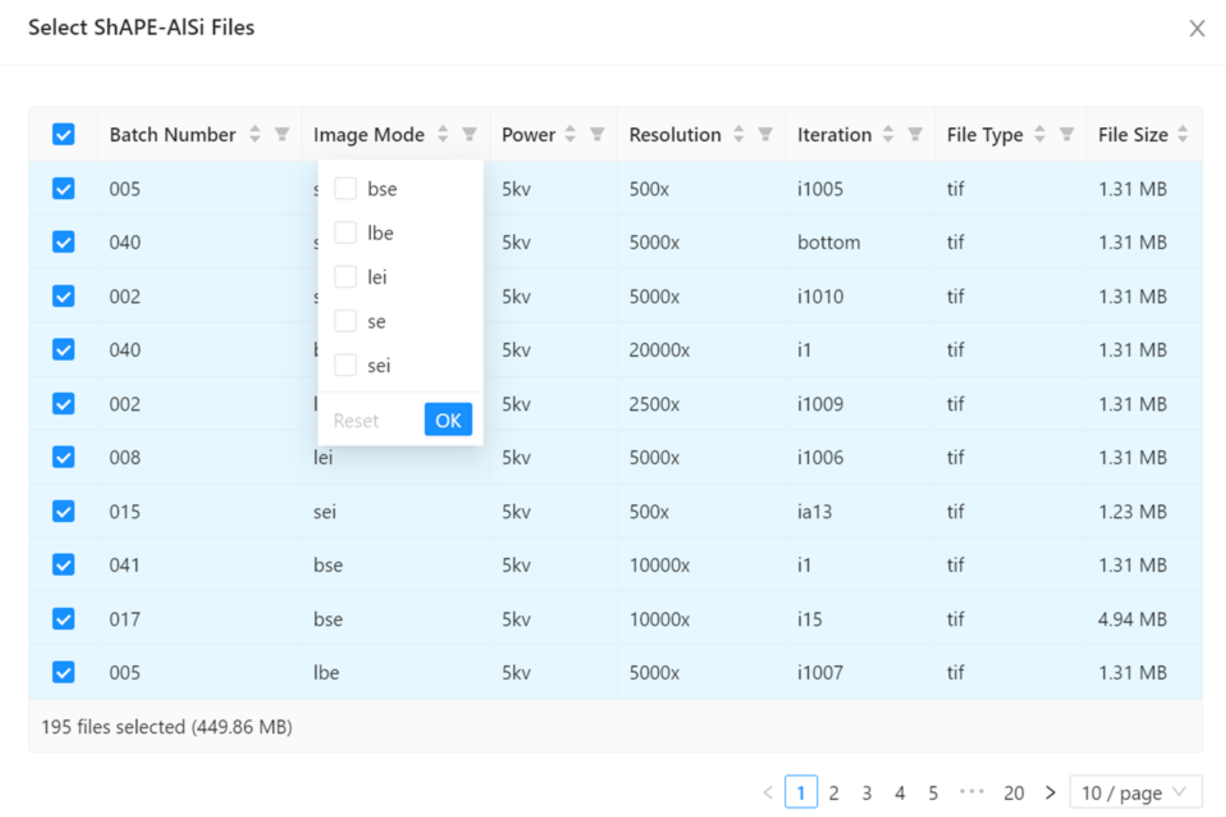


```
"distribution":
[
                        {
                         "identifier": "occupant-stats",
                         "distributionType": "download-external",
                         “accessURL”: “https://otherdata.org/”,
                         "title": "Occupant Data",
                         "shortName": "Occupant Data",
                         "description": "Curated occupant data by month.",
```

Figure 6. An example of data distribution.

For the DAP a special type of distribution was added to enable support for hundreds or even millions of files. In these cases, it is not reasonable to list the files or for the user to browse a list. Instead, they will want to filter the data based on their needs and enable an automated download process. This capability is a key feature of the DAP in providing a method and structure for storing, searching, and downloading large numbers of data files. While the distribution metadata is often very simple in just pointing to a file, the download extension to the metadata includes the ability to describe all the facets of the data which enables precise filtering for the users to focus on just the data needed.

File Naming: The core of the metadata is focused on discovering which dataset is of interest. When there are many files involved a file naming convention is essential for all parts of the system and especially for the end user. By describing and enforcing a data file naming convention, it can be described precisely in the metadata which then enables granular searching for particular data files. As in this example in Figure 7 from a SPP dataset the file name is unpacked into facets that can be sorted, filtered and searched prior to data download.



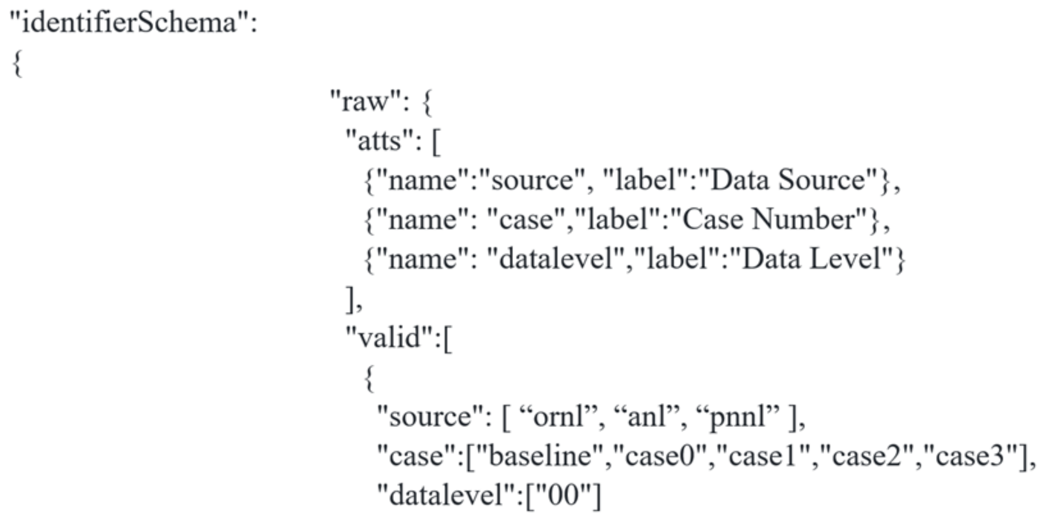
Select ShAPE-AlSi Files

☑	Batch Number	Image Mode	Power	Resolution	Iteration	File Type	File Size
☑	005	s	5kv	500x	i1005	tif	1.31 MB
☑	040	s	5kv	5000x	bottom	tif	1.31 MB
☑	002	s	5kv	5000x	i1010	tif	1.31 MB
☑	040	b	5kv	20000x	i1	tif	1.31 MB
☑	002	l	5kv	2500x	i1009	tif	1.31 MB
☑	008	lei	5kv	5000x	i1006	tif	1.31 MB
☑	015	sei	5kv	500x	ia13	tif	1.23 MB
☑	041	bse	5kv	10000x	i1	tif	1.31 MB
☑	017	bse	5kv	10000x	i15	tif	4.94 MB
☑	005	lbe	5kv	5000x	i1007	tif	1.31 MB

195 files selected (449.86 MB)

Figure 7. Customizable data searching based on the need of end user.

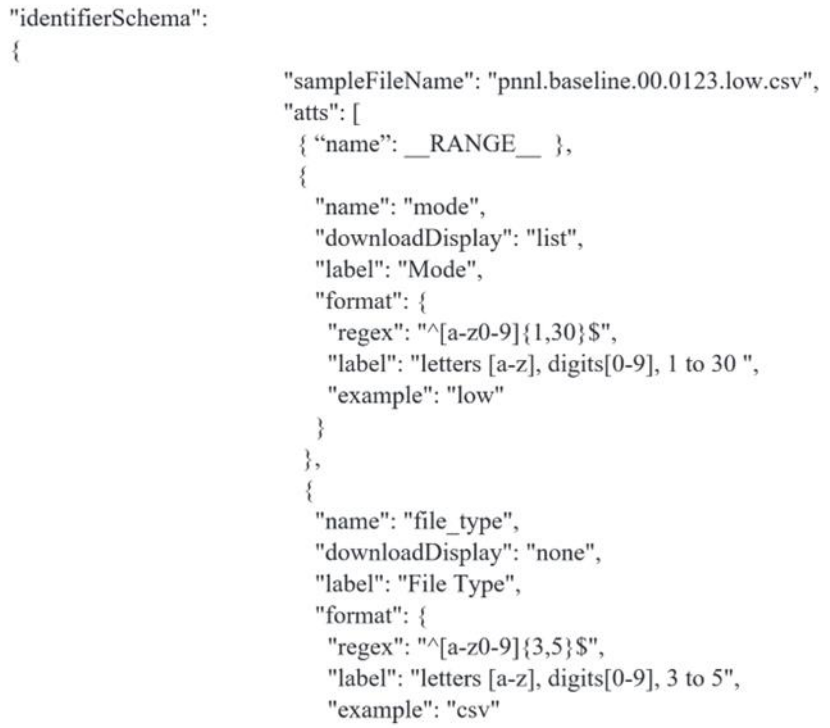
The file naming metadata is unique to each Distribution in a Dataset. This enables flexibility to support the various needs of each dataset. Additionally, 2 special types are supported by the DAP platform: Dates and Ranges. Time based data can easily produce a very large number of files. Similarly, as in some of the SPP data a single collection sample might produce 1000 files that are just distinguished by their order in a range. The DAP indexes these Dates or Ranges for optimal search performance. For large numbers of files, determining the file naming standard is critical for anyone making effective use of the data. This file naming metadata is captured in an attribute called “identifierSchema”. This is used at all levels of the metadata to fully describe the attributes of a file name. This structure is used by the DAP platform to provide searchable labels for different parts of the filename as well as put limits on what is valid. In Figure 8 this attribute is used to describe and limit the valid datasets in a project.



```
"identifierSchema":
{
                              "raw": {
                               "atts": [
                                 {"name":"source", "label":"Data Source"},
                                 {"name": "case","label":"Case Number"},
                                 {"name": "datalevel","label":"Data Level"}
                               ],
                               "valid":[
                                 {
                                  "source": [ “ornl”, “anl”, “pnnl” ],
                                  "case":["baseline","case0","case1","case2","case3"],
                                  "datalevel":["00"]
```

Figure 8. An example of file naming metadata schema.

Then in a distribution this attribute is used to describe and limit the rest of the file name. In Figure 9, a sample file name is included. Additionally, the “Range” attribute is used as part of the file name. Capturing the file naming convention in this way requires the data producers to be clear and specific about the files being stored to ensure later usability.



```
"identifierSchema":
{
                              "sampleFileName": "pnnl.baseline.00.0123.low.csv",
                              "atts": [
                                { “name”: __RANGE__  },
                                {
                                  "name": "mode",
                                  "downloadDisplay": "list",
                                  "label": "Mode",
                                  "format": {
                                    "regex": "^[a-z0-9]{1,30}$",
                                    "label": "letters [a-z], digits[0-9], 1 to 30 ",
                                    "example": "low"
                                  }
                                },
                                {
                                  "name": "file_type",
                                  "downloadDisplay": "none",
                                  "label": "File Type",
                                  "format": {
                                    "regex": "^[a-z0-9]{3,5}$",
                                    "label": "letters [a-z], digits[0-9], 3 to 5",
                                    "example": "csv"
```

Figure 9. An example of file naming schema with “Range” attribute.

3.0 Conclusion

Using the Project Open Data metadata standard enables the DAP to support the FAIR principles of a data repository. The ability to extend and include domain specific metadata is a natural part of the system.   Supporting the effective organization and access to very large numbers of files is the key feature of supporting the complex needs of SPP domains and is supported by the DAP and its metadata framework.

4.0 References

[1] Roy, A., Taufique, M.F.N., Khakurel, H., Devanathan, R., Johnson, D.D. and Balasubramanian, G., 2022. Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys. npj Materials Degradation, 6(1), pp.1-10.

[2] Khakurel, H., Taufique, M.F.N., Roy, A., Balasubramanian, G., Ouyang, G., Cui, J., Johnson, D.D. and Devanathan, R., 2021. Machine learning assisted prediction of the Young's modulus of compositionally complex alloys. Scientific reports, 11(1), pp.1-10.

[3] Mamun, O., Wenzlick, M., Sathanur, A., Hawk, J. and Devanathan, R., 2021. Machine learning augmented predictive and generative model for rupture life in ferritic and austenitic steels. npj Materials Degradation, 5(1), pp.1-10.

[4] Himanen, Lauri, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. "Data-driven materials science: status, challenges, and perspectives." Advanced Science 6, no. 21 (2019): 1900808.

[5] Ward, C., Brinson, L.C., Galli, G., Kalidindi, S.R. and MehtaA, M.B., 2019. Building a materials data infrastructure: opening new pathways to discovery and innovation in science and engineering [Internet]. Pittsburgh: The Minerals. Metals & Materials Society.

[6] The Minerals, Metals & Materials Society (TMS), Employing Artificial Intelligence to Accelerate Development and Implementation of Materials and Manufacturing Innovations (Pittsburgh, PA: TMS 2022). Electronic copies available at www.tms.org/AIStudy.

[7] https://oselp.org/2021-think-pieces.

[8] Hanisch, Robert J., Debra L. Kaiser, Bonnie C. Carroll, Callie Higgins, Jason Killgore, Dianne Poster, Marian Merritt et al. "Research Data Framework (RDaF): Motivation, Development, and A Preliminary Framework Core." NIST Special Publication 1500 (2021): 18.

[9] Whalen, Scott, Matthew Olszta, Md Reza-E-Rabby, Timothy Roosendaal, Tianhao Wang, Darrell Herling, Brandon Scott Taysom, Sarah Suffield, and Nicole Overman. "High speed manufacturing of aluminum alloy 7075 tubing by Shear Assisted Processing and Extrusion (ShAPE)." Journal of Manufacturing Processes 71 (2021): 699-710.

Appendix A – Acronyms and Abbreviations

DAP	Data Archive and Portal
SPP	Solid Phase Processing
SPPSi	Solid Phase Processing Science Initiatives
ShAPE	Shear Assisted Extrusion and Processing
FAIR	Findable, Accessible, Interoperable, Reusable
ML	Machine Learning
AI	Artificial Intelligence

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov