The Materials Commons Data Repository

Glenn Tarcea*, Brian Puchala*, Tracy Berman*, Giorgio Scorzelli†,
Valerio Pascucci†, Michela Taufer‡, John Allison*

* University of Michigan, Ann Arbor, MI USA.
† University of Utah, Salt Lake City, UT USA
‡ University of Tennessee, Knoxville, TN USA

2022 IEEE 18th International Conference on e-Science (e-Science) | 978-1-6654-6124-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ESCIENCE55777.2022.00060

***Abstract*—Repositories are increasingly used for publishing and sharing scientific data. The Materials Commons is a data repository that follows the FAIR (Findable, Accessible, Interoperable, Reusable) principles. We demonstrate the challenges with FAIR and how Materials Commons solves them. We also discuss the Nationals Science Data Fabric (NSDF) [1], a project that is democratizing data access, and show how Materials Commons with the NSDF software stack accelerates data access and scientific research.**

***Index Terms*—Materials Science, FAIR data, Repository, MGI (Materials Genome Initiative), Open Research, Open Access**

I. Motivation and Contributions

To accelerate the pace of scientific advances through the preservation, sharing, and reuse of data, there has been a significant community-wide focus on the development of scientific repositories and collaboration platforms. From making it easy to access the systems, to discovering what data is contained, and understanding how it was obtained and can be used, the development of these repositories faces a number of challenges. Developing a system that makes it easy for users to discover what data is available, access the data, and understand how it was collected and can be reused is a very significant challenge, especially for general purpose repositories that are not restricted to a single type of data. The FAIR data principles [2], standing for findability, accessibility, interoperability and reusability, are commonly used guidelines for systematically considering how to enhance the reusability of data stored in a repository.

In this poster we describe, using the FAIR principles as a guide, the design and implementation of the Materials Commons, a data repository and collaboration platform for the materials science community developed by the PRedictive Integrated Materials Science (PRISMS) Center at the University of Michigan.

II. Materials Commons

Materials Commons is a system that provides researchers with space to store and share project data privately, tools for analyzing and understanding materials data, and the ability to publish data for reuse by the broader community. Along with a website [3], Materials Commons projects and datasets can be accessed via a command line tool and a Python API for custom integration into a user's workflow. Materials Commons was designed and implemented with these key features in mind: findability, data accessibility; interoperability; and reusability.

A. Findability

In Materials Commons, findability is enabled through our meta-data and workflow extraction tool, relationship linkage, and text based search. Materials Commons provides tools that help to capture a user's workflow and important attributes of the materials or models they are working on. These Action/Entity or Process/Sample pages bring together relationships, important attributes, link and display files, and provide access to tools. Through the use of Excel spreadsheets, our CLI tool, and the website we are making it easier to capture and display this information. Materials Commons also provides MQL (Materials Query Language). This interface allows researchers to search on different attributes and their values, and supports complex queries based on contained-in relationships, boolean and comparison operators. The interface provides the user with the range and frequency of values for attributes to help in query construction. The results include process and sample data. Meta data, summaries, relationships and links, present ample space for Materials Commons to further improve its search capabilities and surface relevant data to the user. Figure 1 shows a snapshot of the Materials Common with a sample and process attributes, the workflow, and relationships for the coupling thermomechanical processing and alloy design to improve textures in Mg-Zn-Ca sheet alloys (ZX00_d31) in Materials Commons [4].

B. Data Accessibility

To address data accessibility, Materials Commons makes its data available on the web and easily downloadable using standard and widely used methods, including through standard SCP/SFTP file transfer protocols, web based downloads, and the Globus file transfer service. Datasets are given digital object identifiers (DOIs) so that they can be easily included in papers and shared via permanent links. By providing many different download mechanisms users can get at the data using the tools they are comfortable with and that work in their environment.

C. Interoperability

Enabling interoperability in a general use repository is particularly challenging because the development and adoption of formal schemas and ontologies tends to be a large barrier to entry. Therefore, Materials Commons attempts to take a more social and incremental approach to interoperability. To help

978-1-6654-6124-5/22/$31.00 ©2022 IEEE
DOI 10.1109/eScience55777.2022.00060

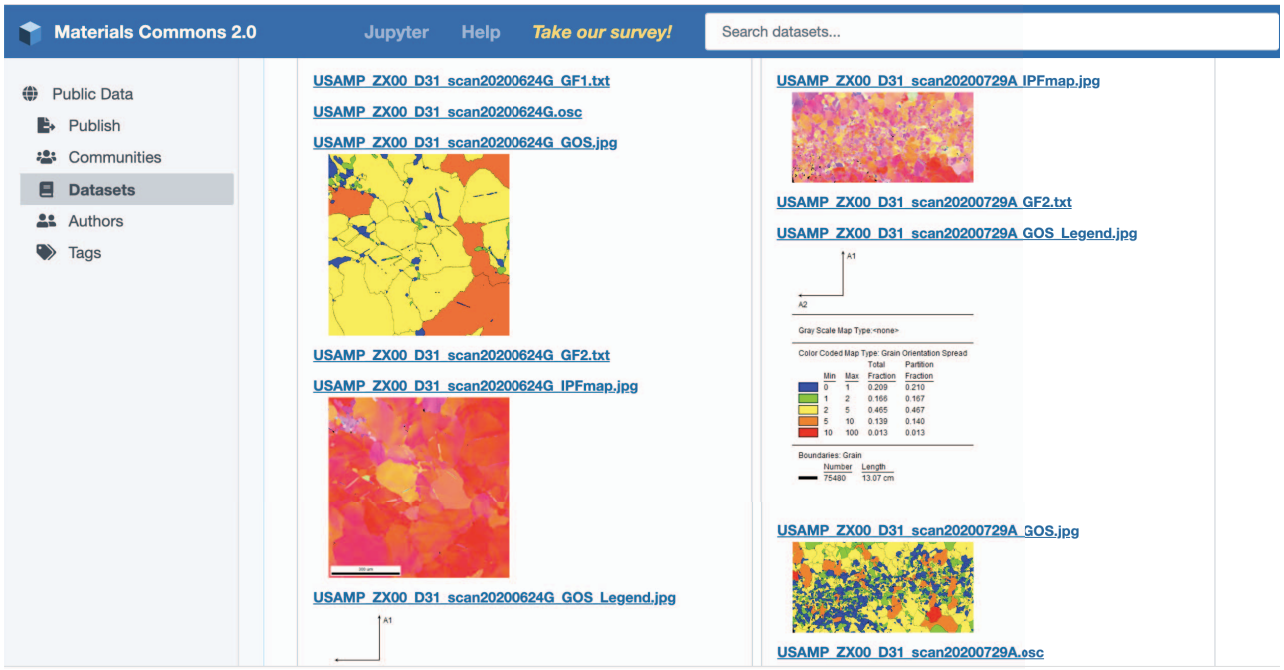

Fig. 1. Sample of process attributes, Workflow, and relationships in Materials Commons.

researchers self organize and encourage the development of de facto standards, Materials Commons includes "Communities of Practice" where users can find and share related datasets and describe best practices for particular types of datasets. Materials Commons datasets can be cloned into a new project, encouraging reuse of both the data and the data format, structure, and metadata conventions. Communities of Practice encourage a social and incremental approach to interoperability. Though still early in development, our thinking is that much like Github created a social environment for sharing code, features like Communities of Practice may be a way to create a similar environment for research data.

D. Reusability

Finally, to enhance data reusability, Materials Commons' workflow presentation helps researchers to explore published data, gain insight and understanding into it, and build their confidence that they can reuse it. We are developing tools that allow for visualization and generation of data on Materials Commons. These tools will allow users who publish data on Materials Commons to provide links in their papers to give readers the ability to explore results and data interactively. We believe this combination of workflow, tools, live exploration and ability to embed live links into papers will change how users interact with and re-use published data.

III. Cross-cutting Collaborations

The PRiSMS center is working with the National Science Data Fabric (NSDF) to democratize data access. We are working to integrate Materials Commons into the NSDF offerings. One of our first projects is with the Cornell High Energy Synchrotron Source (CHESS) to provide high speed access to the beamline data that CHESS produces.

IV. Conclusions

By going beyond traditional repository roles as a place to just store files and some descriptive text, Materials Commons bridges the gap between publishing a set of files and data re-use by helping researchers to explore and understand the published data. We believe that the next generation of repositories will prioritize exploration and understanding of data and as well as provide services for communities to come together and develop standards.

Acknowledgment

Development of Materials Commons is funded by the US Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under Award No. DE-SC0008637 as part of the Center for PRedictive Integrated Structural Materials Science (PRISMS Center) at University of Michigan. Research using Materials Commons is supported by the National Science Foundation, awards #2138811.

References

[1] V. Pascucci and et al, "NSDF: Nationals Science Data Fabric," http://nationalsciencedatafabric.org/.
[2] M. Wilkinson, M. Dumontier, I. Aalbersberg, and et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 160018, 2016.
[3] J. Allison and et al, "The Materials Commons Database," https://materialscommons.org.
[4] T. Berman and J. Allison, "Dataset: Coupling thermomechanical processing and alloy design to improve textures in Mg-Zn-Ca sheet alloys," 10.13011/m3-hvkb-rm86.