JOM, Vol. 68, No. 8, 2016
DOI: 10.1007/s11837-016-2001-3

The Materials Data Facility: Data Services to Advance Materials Science Research

B. BLAISZIK,[1,4] K. CHARD,[1] J. PRUYNE,[1] R. ANANTHAKRISHNAN,[1] S. TUECKE,[1] and I. FOSTER[1,2,3,5]

1.—Computation Institute, University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637, USA. 2.—Department of Computer Science, University of Chicago, Chicago, IL, USA. 3.—Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439, USA. 4.—e-mail: blaiszik@uchicago.edu. 5.—e-mail: foster@anl.gov

With increasingly strict data management requirements from funding agencies and institutions, expanding focus on the challenges of research replicability, and growing data sizes and heterogeneity, new data needs are emerging in the materials community. The materials data facility (MDF) operates two cloud-hosted services, data publication and data discovery, with features to promote open data sharing, self-service data publication and curation, and encourage data reuse, layered with powerful data discovery tools. The data publication service simplifies the process of copying data to a secure storage location, assigning data a citable persistent identifier, and recording custom (e.g., material, technique, or instrument specific) and automatically-extracted metadata in a registry while the data discovery service will provide advanced search capabilities (e.g., faceting, free text range querying, and full text search) against the registered data and metadata. The MDF services empower individual researchers, research projects, and institutions to (I) publish research datasets, regardless of size, from local storage, institutional data stores, or cloud storage, without involvement of third-party publishers; (II) build, share, and enforce extensible domain-specific custom metadata schemas; (III) interact with published data and metadata via representational state transfer (REST) application program interfaces (APIs) to facilitate automation, analysis, and feedback; and (IV) access a data discovery model that allows researchers to search, interrogate, and eventually build on existing published data. We describe MDF's design, current status, and future plans.

Key words: Materials, data publication, data management, data preservation, software as a service

INTRODUCTION

Materials innovation is a critical, well-recognized component and driver of economic prosperity and global competitiveness.[1] Yet, the pipeline through which new materials are designed, developed, manufactured, and deployed—a pipeline that spans from early laboratory research and discovery to market adoption—remains slow, costly, and inefficient.[2] Nationally coordinated efforts, via The Materials Genome Initiative (MGI), are tasked with developing the computational, experimental, and data innovation infrastructure needed to halve the time from materials discovery to deployment.[1] Achieving such results will require not only deeper understanding of materials processing, process–structure–property relationships, industrial techniques, synthesis, simulation, theory, and combinations thereof but also new shared data infrastructure and services to support open materials-specific data sharing and exchange, structured materials data cataloging, and tools to simplify and promote data discovery, data reuse, and development of advanced materials informatics.[3]

Access to a wide variety of materials data is critical to transforming the research-to-market adoption pipeline. To enable access, we need methods for making such data discoverable and

(Published online July 6, 2016)

accessible: in other words, for *publishing* that data. To this end, a variety of materials-related databases and data repositories have been established, for example, the Materials Project,[4] the Open Quantum Materials Database (OQMD),[5] the NIST Materials Data Repository,[6,7] NREL MatDB,[8] NIMS MatNavi,[9] Automatic-FLOW for Materials Discovery,[10] Novel Materials Discovery (NoMaD) repository,[11] Computational Materials Data Network,[12] Citrine Informatics' Citrination platform,[13] and AiiDA.[14] So too have more general scientific repositories such as Zenodo,[15] Dryad,[16] Figshare,[17] and Dataverse.[18] There are also emerging activities in e-science and improved collaboration spaces for scientists such as the PRISMS Materials Commons.[19,20] For further reading, see the review by Kalidindi and De Graef.[21]

Although each of these efforts contributes to data publication and sharing, they neither individually nor collectively represent a complete solution to the materials data-sharing problem. For example, many existing systems are aligned with a particular field of materials science or type of data and, thus, are not general enough to meet the needs of the broad materials community; are designed to store small (on the order of GB) and/or derived data rather than the much larger data (often many TB) that underpins many materials studies; and/or lack support for sufficiently general policies regarding metadata, curation, access control, and persistent identification.

With these challenges in mind, Materials Data Facility (MDF)[22] services are uniquely differentiated from these efforts by our distributed data publication[23] and discovery models (described in the next section) that build on and leverage production services provided by Globus, a non profit software-as-a-service (SaaS) provider affiliated with the Computation Institute, a University of Chicago and Argonne National Laboratory partnership.[24–26] A key focus of MDF is to enable materials researchers to publish, discover, and access datasets regardless of size and source. We expect that MDF will allow researchers to make raw and derived data available, share, discover, and access big datasets (e.g., datasets gathered via x-ray scattering and tomography, high-energy diffraction microscopy, neutron scattering datasets and more) as well as smaller datasets (e.g., datasets gathered via atomic force (AFM), scanning electron (SEM), transmission electron (TEM) microscopy, and more). In so doing, MDF facilitates the data exchange that is needed to avoid redundant work, encourage broader data synthesis, and allow researchers to find new data resources to support their work. We suggest that MDF is an important step towards the common data infrastructure components necessary to unleash materials informatics capabilities toward the realization of in silico materials discovery.

Traditionally, materials data publication has focused primarily on the carefully curated standard reference data (SRD). These highly-valuable, refined and verified data represent the results of dozens of experiments and integrate results obtained by various means. But SRD constitute only a tiny fraction of the vast quantities of data that are produced by experiment, simulation, and analysis every year. Many other data, of different sizes and types, can be valuable in different situations. Nevertheless, these other materials datasets are immensely heterogeneous in terms of source, quality, purpose, format, descriptive metadata, and size. Even within a discipline, data sizes for different experiment types may vary by many orders of magnitude, and the metadata and terms that are used to describe the data and detail the process by which it was obtained may be entirely different. The methods, scripts, and codes used to operate on data to produce derived datasets are equally diverse, ranging from simple Python scripts and Microsoft Excel worksheets to complex codebases and toolkits maintained by researchers from many institutions. Thus, the materials community also needs methods to enable exchange at any level of the data generation process and, more importantly, the ability to review, for any published result, the raw data used to derive the result along with provenance information that describes the process (protocols, experiment conditions, models and analysis codes, etc.) by which the result was obtained. MDF allows this model of rapid data exchange and simplified access to be realized, enabling data to be published, curated, interlinked, and shared at any point of the data lifecycle.

MATERIALS DATA FACILITY

To meet the complex data needs of the materials community, simply allocating storage for research datasets or continuously implementing ad hoc solutions at the research group scale are insufficient and suboptimal solutions. We aim to build scalable, shared data services and provide common data infrastructure components and resources (e.g., reliable, professionally-maintained scalable storage has been allocated at the National Center for Supercomputing Applications (NCSA) to support initial MDF users) to address the emergent needs of the broad materials community. MDF provides intuitive interfaces through which any researcher can access a growing set of advanced capabilities. Initially our focus was on two services, data publication and data discovery (Fig. 1).

Figure 2 provides a high-level view of MDF's data publication and discovery service architecture. We outsource responsibility for operating the publication and discovery services themselves to Globus, which supports them as cloud-hosted services as part of the Globus Publication system.[23] This outsourcing to cloud services, following a SaaS model,[24,25] allows us to reduce friction to initial user adoption, to make advanced capabilities available to a broad community of users via simple Web-

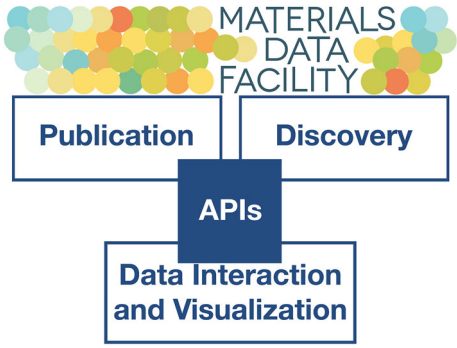

Fig. 1. MDF operates two cloud-hosted core services, data publication and data discovery. APIs for the core services will allow researchers to develop more specific tools for data interaction and visualization to fit particular materials science use cases.

based interfaces, and to reduce end-user costs and maintenance burden as no local software needs to be installed. Users publish and discover data via a Web user interface (UI) or representational state transfer (REST) interfaces. MDF builds on and leverages Globus services to provide authentication, group management, and data access and transfer. It leverages DSpace[27] to provide workflow management and policy enforcement as well as the user interfaces for publication and discovery.

MDF implements a publication model (Fig. 3) in which data are organized in terms of collections and datasets. A collection comprises a set of policies and a set of logically similar datasets. Configurable collection policies allow the collection administrator to specify the curation workflows, metadata schemas, input forms, storage endpoint(s), dataset identifiers, and other policies that apply to datasets within the collection. A dataset is a bundle of data URIs and metadata as prescribed by the collection meta data schema(s). Importantly, although initial storage for the MDF data publication has been allocated at NCSA (UIUC), the data storage for each collection can use any one Globus endpoint:[6] i.e., any locally-hosted storage, repository-specific storage, institutional storage, or even cloud storage (e.g., Amazon S3) running a Globus Connect client. After a dataset is assembled, the associated metadata are indexed to a registry to facilitate materials dataset discovery by MDF users.

MDF Data Publication Capabilities

Data publication goes beyond simply making data accessible on the Web; it also encompasses associating a core set of descriptive attributes with the published data. For scientific data to be broadly useful beyond its original purpose, it should be well described, discoverable, verifiable, and accessible. Ideally, improving the quality of each of these of these features allows the underlying study (either experimental or computational) to be understood, replicated, and built on in future studies. MDF services attempt to address these key features at various levels through a flexible collection-oriented policy model.[7] Figure 4 shows current (and planned capabilities) of MDF along several important axes.

Identification MDF data publications are uniquely identified with a collection-specific persistent identifier (PID). Although data need not remain in the same physical location for all time, the PID is resolvable to a landing page that defines the current location of the underlying data and metadata throughout the data lifetime. Association of a persistent identifier enables proper citation, as well as linkage between, for example, papers and the underlying dataset, and improves subsequent data discovery and metric tracking. MDF currently supports Digital Object Identifier (DOI) and Handle PIDs.[28] Collection owners may specify and configure the PID provider used, based on their individual publication and institutional requirements, whereas MDF also provides a default DOI namespace that collection owners may elect to use.

Description Without descriptive metadata, published datasets are difficult to discover, and even when discovered, it is challenging for others to understand the meaning and the conditions under which the data were collected and processed. It is therefore desirable that published data are described such that others can discover and understand that data in the future. MDF supports the description of published datasets by using arbitrary and extensible metadata. Collection administrators may define schemas, consisting of optional and required metadata fields, for all data published to their collection. As MDF grows, users will be able to discover and reuse schemas developed by other researchers to describe their own datasets as appropriate. Where possible, domain-specific schemas, ontologies, and taxonomies will be used to describe contents following consistent standards that promote cross-repository sharing.

Curation MDF supports the publication of data into user-managed collections. Depending on the context of these collections, there may be requirements (expressed or implicit) that data meet particular levels of completeness before being made available. For example, requirements may include that data are complete (i.e., the data are sufficient to reproduce a study), well described after a prescribed metadata schema, or meeting various institutional or funding agency policies. MDF supports the definition of optional user-driven curation workflows in which configurable groups of users must approve datasets before they are published. In these workflows, curators may be allowed view or modify metadata or data within the dataset.

Access to data must be both intuitive and secure. MDF provides a Web user interface and will provide REST interfaces to access published data. It also enables flexible authorization policies to be expressed and enforced on collection management, data publication, and data access. Importantly, user identity is established via Globus Auth using users'

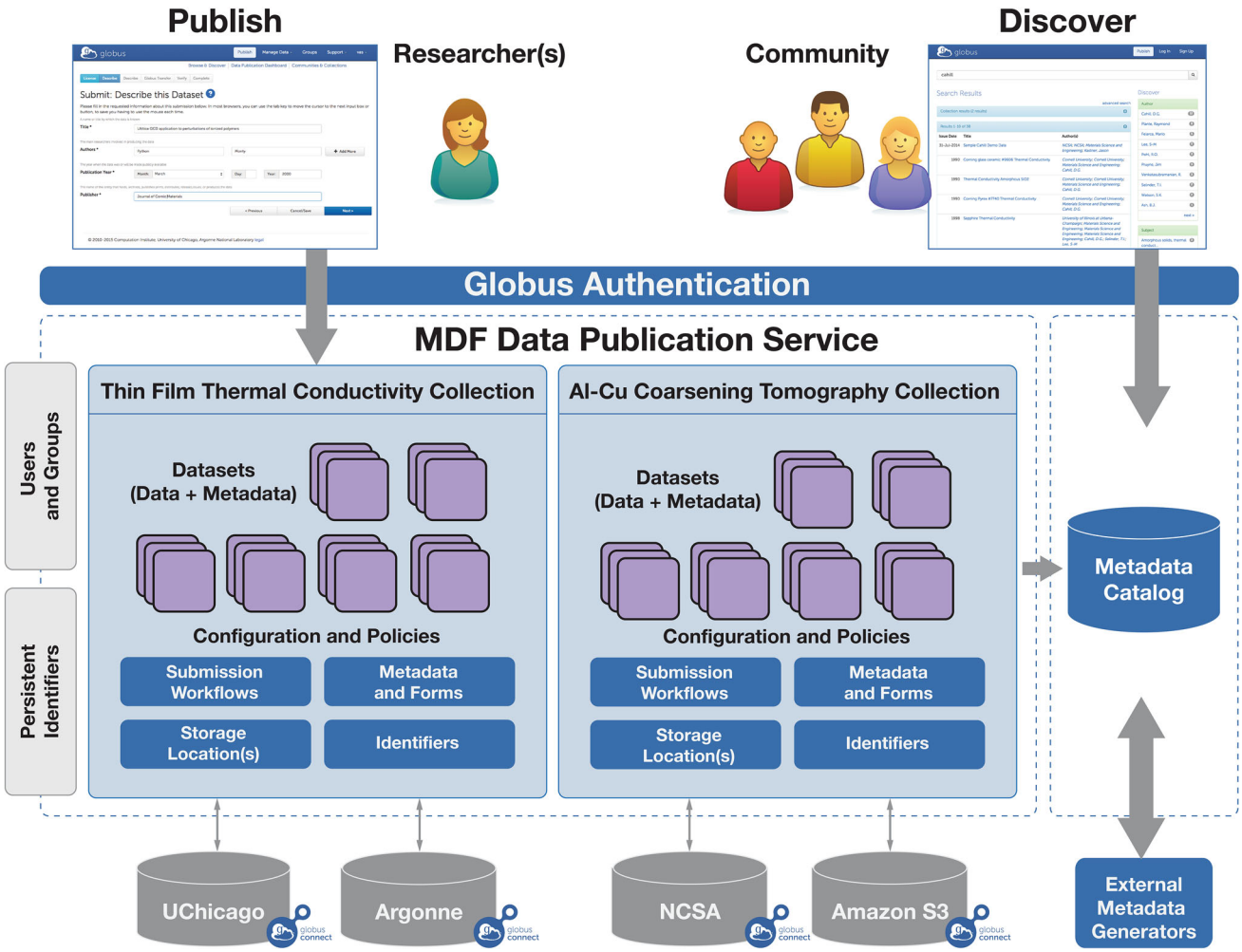

Fig. 2. MDF architecture diagram showing (top) the user interfaces for publication and discovery; (center) a logical view of the system, with multiple distinct collections and the metadata catalog; (left) users and groups handling, and persistent identifier minting; (bottom) the different repository storage endpoints (both local and cloud) used to store data. The metadata catalog indexes dataset metadata and file-level data (in development). It will soon also support ingest of external metadata (e.g., from other repositories).

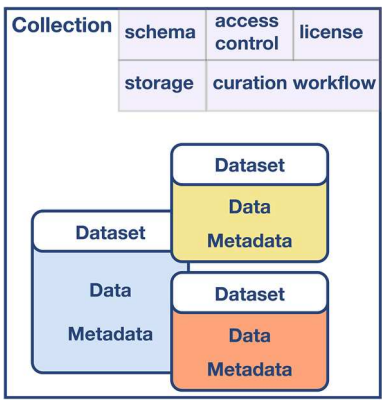

Fig. 3. Collection data model used for publications within MDF. When a user creates a collection, he or she defines its metadata schemas, access control policy (either public or restricted to a user group), license, storage endpoint, and curation workflow. Each collection can contain many datasets, with each dataset comprising a data-metadata bundle.

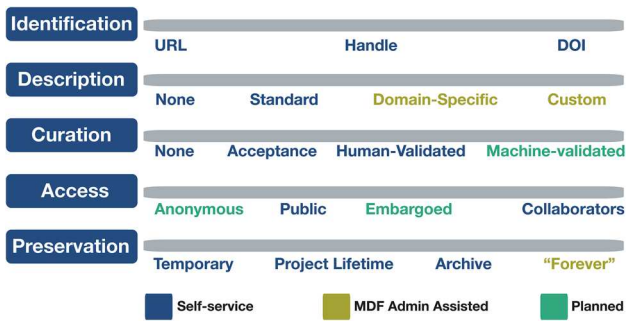

Fig. 4. Publishing materials data require flexible policies in the areas shown here. Identification policies can range from a short-term URL to a long-term unique DOI; description policies can range from none (i.e., a raw dataset) to a dataset described with domain-specific metadata; curation policies can range from none to machine-validated content; access policies can range from broad public access to access only by specified collaborators; and preservation policies can range from temporary to decades.

external identities. Since MDF is built with Globus technology (which will be discussed in more detail), users are also able to use Globus python API clients for data access and transfer. In addition, the MDF data publication service includes capabilities to toggle sharing settings associated with individual collections, to allow read access to any group of users (e.g., the creator alone or a research group or effort) or to the public. We are currently investigating methods to support temporal dataset embargoes.

Preservation requirements can vary by project, funding agency, and institution. MDF's distributed data model allows collection owners to select a storage location most appropriate for their application. For example, the distributed model allows for the dataset pointers to the underlying data to move as the data move over time, e.g., because of hardware migrations or data migration to cold storage. Critically, MDF also enables storage owners to implement redundant storage deployments, arbitrary backup procedures, and multi-data center or disaster recovery models.

Verifiability of data, especially with a distributed data model, is important so that users have an assurance that a specific dataset, once published, has not been changed. The MDF data publication service relies on the Globus data access infrastructure[24–26] to prohibit modification of published data. We are also working to enable long-term verifiability of all published data via a computed and cloud-stored checksum. Thus, users who access published data will be able to verify that the contents have not changed between initial publication and access, regardless of data location.

After data are published with the data publication service, as described previously, the datasets should be readily discoverable. The second component of MDF focuses on building capabilities to discover and access published data more easily. For data to be discoverable, it is critical that published data be intuitively searchable, browsable, and retrievable:

Searchable Datasets are only useful to other researchers if they can find them. As the number and diversity of published datasets published with MDF increases, methods are needed to discover data of interest quickly and easily. Users are now accustomed to powerful search interfaces that provide high-quality results even when the user has only incomplete information about what he or she is trying to find. As such, MDF has been designed with powerful search methods to filter quickly through large collections of data, to sort collections by arbitrary metadata attributes, and to find results based on both structured queries (e.g., "beam_energy $>$ 0.1 kV") or full-text search queries (e.g., "beam_energy Cu tomography").

Browsable Simple browsing interfaces enhance discovery of known data, improve discovery of similar but non-exact matches, and facilitate serendipitous discovery. The MDF data publication service makes published data available via intuitive and standardized Web UI or REST interfaces, organized in standardized ways, and described following custom or standardized metadata formats. These methods enable users to browse many datasets quickly, e.g., faceted by author, topics, or date; inspect or explore dataset attributes; find similar dataset publications; and explore connections between datasets.

Retrievable All data and metadata that a user are authorized to access should be easily retrievable (e.g., via Web UI, REST API, etc.). MDF supports standard Globus data access models that allow authorized users to download and transfer published datasets and encoded metadata descriptions directly from the publication storage system to the user's storage system.

BUILDING THE MATERIALS DATA FACILITY

Rather than implement the MDF data publication and discovery services from scratch, we instead developed the first version of the MDF data publication service by adapting the commonly-used DSpace institutional repository system and augmenting its capabilities with enhanced search and large-scale data management functionality provided by Globus.[24–26]

DSpace[27] provides the workflows for publishing, curating, and accessing data. It supports a self-service collection management model via which individual researchers can create and manage collections as well as associated policies. Importantly, it offers a granular access control model that enables permissions to be set and enforced on collections, workflows, datasets, and metadata. Although designed for document-based publication, local storage, static workflows and metadata, global collection policies (e.g., identifiers), and non-scalable data management via HTTP, DSpace is built on a modular architecture that we have deeply modified to enable large and remote data management as well as self-service customization of almost every collection policy.

We leverage *Globus Auth* to allow MDF to accept identities from any Globus Auth-supported identity provider, a group that includes many academic institutions. Through a standard OAuth2 authentication and authorization model, MDF supports secure authentication via campus identities, commercial identities (e.g., Google), or other supported identity providers. Globus Auth also allows collections of identities to be linked, thus, enabling users to create a set of identities (e.g., their institutional credential, a Google credential, and a credential associated with a national computing facility) via which they are presented with the same state in MDF, irrespective of which identity is used to log in.

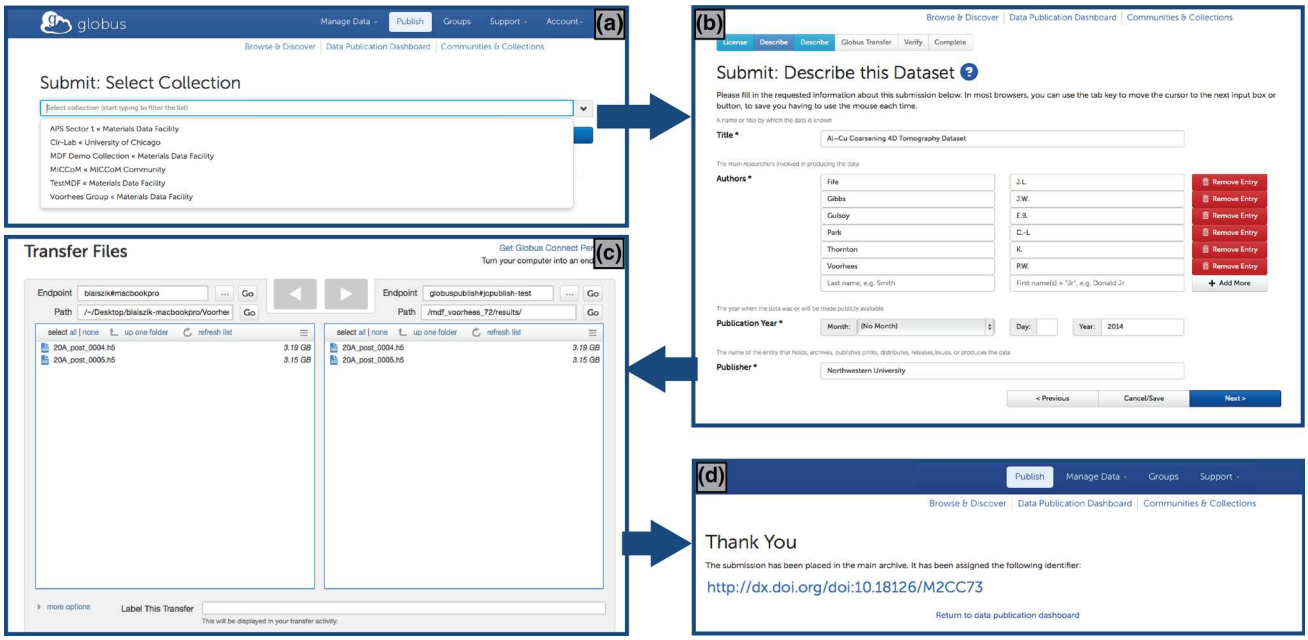

Fig. 5. Dataset submission workflow (following the arrows from the upper left) showing how a data publication is assembled with the Web user interface. Steps include: (a) choosing a collection for the submission (collections have associated policies, including final data storage location, curation requirements, metadata schema, type of unique identifier); (b) describing the item with customizable metadata; (c) assembling the underlying dataset using Globus transfer from one or more source data endpoints; and (d) obtaining the unique identifier for the dataset.

Globus Data Management provides high-performance remote data access, transfer, and sharing services. Globus presents a global namespace through which distributed data can be uniquely referenced and managed. Through Globus's data access APIs, MDF can control (in real-time) permission to deposit and access data in remote, distributed storage locations. Globus data access APIs are implemented on a wide variety of data storage platforms, from cloud storage to specialized high-performance data storage. Finally, Globus provides a reliable, fault-tolerant, and high-performance method for transferring large amounts of data (reaching TB and beyond), enabling MDF users to publish, move, mirror, and download large datasets.

Globus Groups provides self-service group management that underpins the self-service authorization model used in MDF. Users control aspects of group management ranging from membership criteria, to membership workflows, to approval processes, and to group and membership visibility. These groups can then be associated with MDF policies for collection administration, submission authorization, data access rights, and curation responsibilities.

Publishing, Curating and Discovering Data with MDF

Data publication in MDF follows the flow shown in Fig. 5. As an example, we describe the steps involved when a researcher from a university at a scientific facility (e.g., the Advanced Photon Source) publishes a four-dimensional (4D) tomographic dataset many TB in size to an existing tomography-related collection on MDF.

1. Ensure that the data are in a location with a Globus endpoint installed. If no endpoint is installed at the data storage location, the required client software (similar to a Dropbox or Google Drive client) can be installed on MacOS, Windows, or Linux.
2. Authenticate to the MDF data publication service with institutional credentials (e.g., Northwestern ID). Once authentication is successful, the researcher is presented with his or her personalized data publication dashboard.
3. From the dashboard, select "submit a new dataset", and then choose the appropriate collection, for example, the x-ray tomography collection Fig. 5a, for submission and accept required collection licenses.
4. Describe the dataset with standard metadata (e.g., authors, title, affiliation, funding details—Fig. 5b) and custom metadata associated with the chosen collection (e.g., beam energy, sample material composition, tomographic scan time, identifiers associated with other data logs or notebooks, etc.).
5. Assemble the data for the dataset from any number of storage endpoints (perhaps including multimodal data from multiple beamlines or from geographically separate experimental facilities) using Globus transfer tools to copy selected files and folders (Fig. 5c). This assembly step is

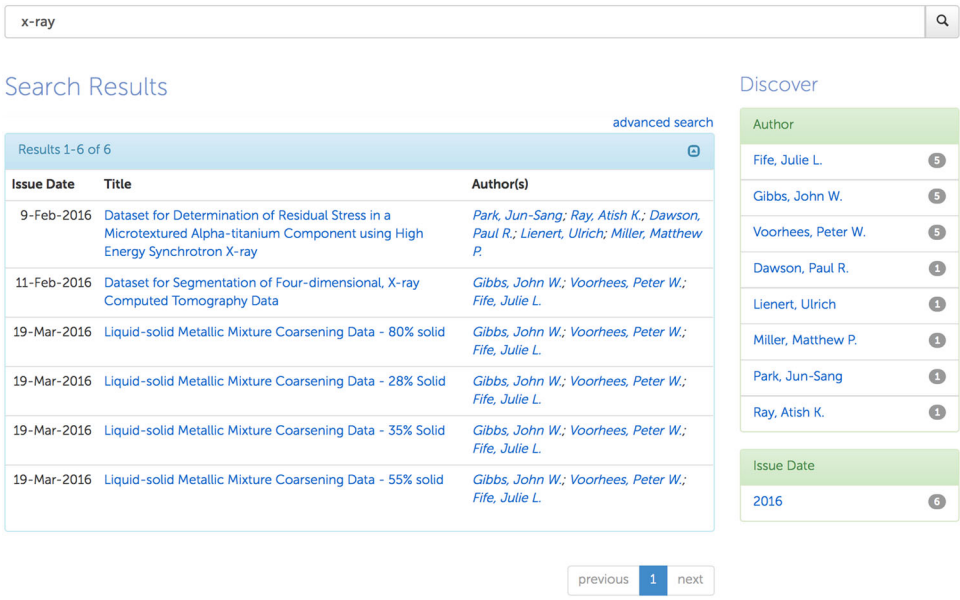

Fig. 6. Free-text search across MDF collections (here, for "x-ray"). Search results, comprising matching published datasets to which the user has access, are shown on the left; on the right, facets on subject and author allow quick navigation through the results. MDF free text searches are performed against all indexed metadata fields.

asynchronous, allowing the researcher to assemble the dataset over time while continuing to update and refine the metadata.

6. Upon completion of dataset assembly, the user is presented with a summary page to confirm publication details. The researcher has the opportunity to verify the assembled data and metadata before finalizing.
7. If a curation workflow is defined for the collection, the collection's predefined group of curators is notified via email of the pending submission. (A curator can also view any publications awaiting curation on his or her publication dashboard.) Any authorized curator may "claim" a curation task. He or she then guided through a process by which the curatory may check the metadata and underlying data to ensure that it meets the requirements of the collection. Curators are finally presented with the option to publish or reject the dataset. In the case of rejection, the curator may optionally specify a reason for rejection which the original submitter can use to address any issues and resubmit the dataset at a later date.
8. After the dataset curation workflow is complete, the dataset and its metadata are indexed and made available following the authorizations defined by the collection and a permanent unique identifier is minted (Fig. 5d). Free text search and automatically generated facets on the MDF portal and collection pages help users find the datasets across collections. An example of a free text MDF search, with cross-collection results, is shown in Fig. 6. Matching datasets are returned as result summaries (left), and additional facets on author and subject (right) are generated to help users navigate through the result set.

If no collection is a good fit for the dataset publication, the researcher may submit the dataset to an "open" MDF collection or define his or her own collection. Creating a new collection involves designating a storage endpoint, defining a curation workflow (optional), specifying data access policies (public or shared with a group of users), and defining the metadata schema and requirements for the new collection.

SUMMARY AND FUTURE WORK

MDF provides scalable and robust data publication and discovery services for the materials science community. It is designed to meet a wide range of requirements derived from this diverse community, including the ability to publish small and large data of varying types, employ arbitrary metadata schemas that can semantically describe the data across disciplines, and implement secure and dynamic access control and workflow models. With 100 TB of reliable and secure storage available in its first instantiation (scalable to PB), MDF is ready for production use. Adoption of MDF data publication has been swift. Over the first month of availability, it has been used by several groups of materials scientists to publish studies including: dynamic x-ray tomography of Al-Cu alloy solidification and coarsening, multimodal characterization of MoS_2 epitaxial growth, and wide-angle x-ray scattering

(WAXS) of various Ti alloys with datasets that individually exceed 600 GB and collectively exceed 2.5 TB in size. Over the coming year, we will continue our outreach to the materials community, build relationships with materials research centers and national user facilities, and publish a wide variety of datasets to improve our process understanding.

Our current work is focused on supporting adoption of MDF and on enhancing the capabilities provided to users. We continue to work with materials scientists across the Unites States and worldwide to further understand their data publication requirements and refine their associated use cases. In the process, we are working to develop new, reusable metadata schemas with collaborators at NIST and to, explore deploying additional data storage locations. We are also investigating integration with complementary materials data and tool repositories with the goal of developing an integrated ecosystem of materials capabilities.

Based on user feedback, we are enhancing MDF capabilities with increased metadata flexibility, customizable search, automated metadata extraction, and extensible APIs and python client libraries. We are also investigating integration with the publication processes of academic publishers and a variety of institutions. Collectively, these capabilities will further increase the value provided to materials scientists, while reducing overheads and enabling new integration scenarios. We welcome interest and feedback from the broad materials science community to help shape and drive what we hope will be a transformative initiative.

ACKNOWLEDGEMENTS

MDF is a collaboration among the University of Chicago (UC), Argonne National Laboratory (ANL), the National Center for Supercomputing Applications (UIUC), and the NIST-funded Center for Hierarchical Materials Design (CHiMaD) (Northwestern, ANL, UC, NIST). The authors would like to acknowledge the ongoing contributions to MDF of John Towns, Kenton McHenry, and Michael Ondrejcek at the University of Illinois at Urbana-Champaign and the National Data Service. This research was supported in part by DOE under Contract DE-AC02-06CH11357; and by NIST as part of the CHiMAD project funded by the U.S. Department of Commerce, National Institute of Standards and Technology, under financial assistance Award Number 70NANB14H012.

REFERENCES

1. J.P. Holdren, *Materials Genome Initiative for Global Competitiveness* (National Science and Technology Council OSTP, Washington, USA, 2011).
2. A. White, *MRS Bull.* 37, 715 (2012).
3. J.J. de Pablo, B. Jones, C.L. Kovacs, V. Ozolins, and A.P. Ramirez, *Curr. Opin. Solid State Mater. Sci.* 18, 99 (2014).
4. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, *Appl. Mater.* 1, 011002 (2013).
5. J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, *JOM* 65, 1501 (2013).
6. T.N. Bhat, L.M. Bartolo, U.R. Kattner, C.E. Campbell, and J.T. Elliott, *JOM* 67, 1866 (2015).
7. NIST Repositories (National Institute of Standards and Technology, 2016), https://materialsdata.nist.gov/dspace/xmlui/. Accessed 31 March 2016.
8. NREL MatDB (National Renewable Energy Laboratory, 2016), http://materials.nrel.gov. Accessed 31 March 2016.
9. MatNavi (National Institute for Materials Science, 2016), http://mits.nims.go.jp/index_en.html. Accessed 31 March 2016.
10. S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L. Hart, S. Sanvito, M. Buongiorno-Nardelli, and N. Mingo, *Comput. Mater. Sci.* 58, 227 (2012).
11. The NoMaD Repository (NoMaD Team, 2016), http://nomad-repository.eu/cms/. Accessed 31 March 2016.
12. CMDNetwork (ASM International, 2016), http://www.asminternational.org/web/cmdnetwork. Accessed 31 March 2016.
13. Citrine Informatics, http://citrination.com. Accessed 31 March 2016.
14. G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, *Comput. Mater. Sci.* 111, 218 (2016).
15. Zenodo, https://www.zenodo.org. Accessed 31 March 2016.
16. H. Tran, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, *Sci. Data* 3, (2016).
17. Figshare (Digital Science & Research, 2016), https://figshare.com. Accessed 31 March 2016.
18. The Dataverse Project, http://dataverse.org. Accessed 31 March 2016.
19. PRedictive Integrated Structural Materials Science, http://www.prisms-center.org. Accessed 31 March 2016.
20. Materials Commons, https://materialscommons.org. Accessed 31 March 2016.
21. S.R. Kalidindi and M. De Graef, *Annu. Rev. Mater. Res.* 45, 171 (2015).
22. Materials Data Facility, https://www.materialsdatafacility.org. Accessed 31 March 2016.
23. K. Chard, J. Pruyne, B. Blaiszik, R. Ananthakrishnan, S. Tuecke, and I. Foster, in E-Sci. E-Sci. 2015 IEEE 11th Int. Conf. On (IEEE, 2015), pp. 401–410.
24. I. Foster, *IEEE Internet Comput.* 15, 70 (2011).
25. B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, and T. Steven, *Commun. ACM* 55, 81 (2012).
26. K. Chard, S. Tuecke, and I. Foster, *Cloud Comput. IEEE* 1, 46 (2014).
27. M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, and J.H. Walker, DSpace: An open source dynamic digital repository (2003).
28. N. Paskin, Digital Object Identifier (DOI®) System, Third Edition, 2011.