The Materials Data Facility

A prototype for the National Data Service

Overview

The Materials Data Facility (MDF) is a suite of services that enable materials scientists, particularly those engaged in the Materials Genome Initiative (MGI), to share, archive, and publish digital data resulting from their research. This facility is also intended to serve as a prototype for a National Data Service (NDS) that will provide similar services for data from any research discipline. We will use the requirements, use cases, and experience of materials scientists to drive the design and implementation of the broader NDS. The overall purpose of the Facility is to make the data from materials research more widely and easily available for re-use, re-analysis, and verification. Removing the barriers to data will hasten the flow of data, information, and results between researchers and engineers across academia and industry and will accelerate the process of bringing new materials to market.

The Facility will provide capabilities that will be useful throughout the research and publishing process:

1. Private storage of data and sharing prior to publication
    - Researchers will be able to log into the MDF to securely access personal storage space where they can upload and download their data files.
    - Researchers will be able to organize their data products and files into logical collections, and upload files into those collections.
    - Researchers will be able to share data files and collections with their collaborators, either by simply sharing a URL to the data or by managing group permissions on the data collections.
2. Preparing data for publication
    - The MDF will provide tools that allow the user to extract or assign metadata to the data files in their collections. Validators will ensure that datasets have the metadata needed to be referenced from the literature.
    - When the paper that describes the results based on a data collection is ready for publication, the researcher may engage a data-publishing tool to release the data to the public. In particular,
        - Digital Object Identifiers (DOIs) are assigned automatically to the collection as a whole and to the individual files as appropriate.

        - The publisher of the related article is notified of the data's availability and is provided with all information necessary for linking the paper to the data products.
    - When the paper goes to publication, the associated data products *may* be shipped to another repository for long-term preservation.

3. Enabling data re-use
    - The MDF will provide services for resolving data product identifiers to their download-able files
    - As new collections are published, the MDF will issue electronic notifications to subscribers.
    - Data transfer services will allow individual products or whole collections to be transferred robustly and efficiently to compliant processing and analysis platforms (such as campus clusters). Special access may be enabled for compute systems near the repository storage.

Initially, the MDF will support only rudimentary data discovery, based on basic author, keyword, and publication metadata. However, as collections become more numerous and diverse, richer discovery mechanisms may be necessary.

An important NDS the NDS goal is to establish an open framework for the data services it provides, including repository services. An open framework will be allow the community at large not only to bring in and integrate additional resources into the NDS but also to build specialized tools that interoperates with the NDS. We believe such an open framework will be just as important for the MDF and the MGI community. We imagine, for instance, that commercial users might continuously monitor new collection notifications and automatically download and analyze products matching criteria of interest. Thus, in addition to building user-oriented web interfaces to the MDF services, we will also create REST-ful web interfaces for programmatic access.

Other topics to address:

- Anticipated usage: how many datasets, of what size, from how many people, are we expecting? (This can help us scale things.)
- Anticipated policies: how much space will we give people for private storage? For published data?
- Success metrics: How will we judge success? (E.g., number of uploads, number of visitors, number of downloads, number of citations to DOIs - amount of data re-use)
- Who shall we recruit to provide initial data? To maximize chances of success, we need to be strategic here. Dane has some good ideas!

- Note that when MGI people talk about data repositories, they are mostly thinking about large numbers of computed or experimental values stored in databases, not large experimental datasets. Do we think that we want to address the former as well?