Research Data Management Implementations Workshop

Crædl: Collaborative Research Administration Environment and Data Library

Adam J. Sierakowski, Ph.D.*
Maryland Advanced Research Computing Center
Johns Hopkins University

14 – 15 September 2017

1 Background

The Hopkins Extreme Materials Institute (HEMI[1]), a premier research institute at the Johns Hopkins University, provides global intellectual leadership to advance the fundamental science associated with materials and structures under extreme conditions. HEMI leads the multi-institution collaborative Center for Materials in Extreme Dynamic Environments (CMEDE[2]), which brings together the work of hundreds of researchers at twelve research institutions across the country. The CMEDE consortium faces three key challenges beyond its primary scientific mission:

1. Managing the research efforts of hundreds of researchers distributed across the country;
2. Sharing large data sets across institutional boundaries;
3. Igniting collaborative efforts through data discovery.

In this document, we introduce Crædl, a new tool being developed to solve CMEDE's data management challenges. However, we recognize that the challenges faced by CMEDE are by no means limited to this particular consortium, and are constructing Crædl to assist all academic research groups, including those that operate at a far smaller scale than CMEDE.

2 Crædl

Crædl—the Collaborative Research Administration Environment and Data Library—is a research data management environment built with one guiding principle in mind:

> *Empower researchers to focus on their domain science by providing an environment for data management that expedites task administration, project collaboration, and community interaction.*

Crædl does this using the following features:

- Research group management: organizes researcher networks and their grants, projects, and publications;

*Contact: adam@craedl.org
[1] `http://hemi.jhu.edu`
[2] `http://hemi.jhu.edu/cmede/`

- Automatic metadata population: combines research group administrative information with researcher-augmented metadata templates to pre-populate metadata during file upload;
- File sharing: allows researchers to control read and/or write permissions to share their files with individuals or groups within Crædl;
- Data discovery: indexes metadata for searching;
- Federated login: allows researchers to use their institutional credentials for access.

Crædl balances structure and flexibility to enable researchers to incorporate it directly into their daily workflow. By doing so, researchers can manage their data in small increments spread over the course of their project instead of returning to it at the end when there is less likelihood that they will accurately document the work (if at all).

This documentation is not only beneficial to the researcher during the project—they will be able to find the data they are searching for more easily—but is also beneficial to advancing the scientific knowledge in their research group, institution, and wider community through improved organization, planning, documentation, and knowledge transfer. Further, after collecting all of this administrative and research data, Crædl can easily provide activity reports to support institutional decision making with hard data.

Crædl has been under development since the beginning of 2017 and it will be rolled out to HEMI's researchers in this Fall. You can learn more about Crædl online at `craedl.org`, and you can reach out to `info@craedl.org` for more information. Presently, we will discuss Crædl's features and workflow in detail.

2.1 Research group management

A crucial piece of Crædl is the research group management feature. Crædl not only keeps track of personnel, but also grants, projects, and other research activities such as publications and presentations. This feature serves two main purposes:

1. Provide an environment in which principal investigators can track the data generated by their researchers;
2. Assimilate information crucial to the daily operation of the research group that Crædl can use to pre-populate file metadata.

At face value, Purpose 1 may seem to go beyond the scope of the core application, and in many ways that is an accurate assessment. However, the importance of Purpose 2 necessitates Purpose 1. We will discuss both simultaneously since they are so intimately related.

Researchers have something that we want: their metadata. Through first-hand experience, we have found that the promise that Crædl will index this metadata and allow researchers to search through it is often insufficient to convince them to spend their time filling out webforms describing their work. This brings us to the lower-order reasoning behind the above Purposes, in reverse order:

2. When a researcher puts a file into Crædl, we want to minimize the number of clicks required to input the metadata by inferring as much information about the work as possible;
1. The more information that we know about the daily operation of the research group, the more accurately we can infer details about the work.

Incorporating the research group management feature into Crædl not only assists the principal investigator directly, but also through secondary and tertiary features that can only exist because of the information that Crædl is subsequently able to correlate between researchers and their work. This feature is the key to the ability of Crædl to make file upload—and, crucially, metadata input—efficient. Without this efficiency, researchers will not use the system. Additionally, with researchers working within the system, Crædl can ease the administrative burden of research by assisting in project coordination, organization, planning, documentation, and knowledge transfer.

2.2 Automatic metadata population

Crædl is battling for the attention of the researcher and it must never get in the way. It must instead work alongside the researcher as an assistant. As such, one of the core features of Crædl—automatic metadata population—is designed to minimize amount of time it takes a researcher to upload a file and populate its metadata. Metadata recording, e.g., which researcher contributed the file (the researcher currently logged in) or file modification date (now) are basic pieces of information that Crædl can easily specify automatically. Further, Crædl is able to pre-populate some less obvious pieces of metadata with information provided through the group management feature, such as which grant a file is associated with or which principal investigator oversees a project.

There remains a significant amount of metadata that is unique to each data set that only the researcher uploading the file can provide. In order for Crædl to minimize the number of clicks needed to populate the metadata, it provides an extensive templating framework, which is best described through the complete workflow required to upload a file. Here, we describe that workflow.

2.2.1 Community metadata schema

First, Crædl creates communities where researchers can gather to work on related problems. For example, a researcher may be a member of a computational fluid dynamics (CFD) community. Crædl enables researchers within this community to decide on a community-wide metadata schema that should be applied to all files associated with the community. In our example CFD community, we may decide to specify a metadata field that contains the type of simulation being run, e.g., a pressure-driven channel flow or a wall-driven shear flow. Only specifically appointed researchers have the authority to edit the community metadata schema.

The community may specify any number of fields of various type, and each field can be marked as being required. Importantly, a default value can be saved for each field, and this value will be inherited as the default as we proceed deeper into the workflow.

2.2.2 Project metadata schema

When a principal investigator receives a new grant, s/he enters it into Crædl and attaches it to an existing community. Subsequently, when a researcher begins a new project under this grant, the project inherits the community metadata schema and allows the researcher to edit the default metadata values set by the community.

Under each project, perhaps analogous to a new computational study in our CFD research example, the researcher is able to specify additional metadata fields that are unique to the project in the same manner that the community metadata schema was specified above. Each field may be marked as required, and default metadata values may be saved. As an example, our CFD researcher might be simulating fluid/particle interaction in this project and it may be valuable to store, e.g., the radius of a particle as a metadata field or the code version being used.

A project (and, if desired, any of its contents) may be shared with any other Crædl researcher with either read and/or write permissions in order to facilitate organization, collaboration, and knowledge transfer.

2.3 File sharing

Within each project exists a file browser with the project representing the root directory. Researchers can create directories to assist in their own organization, and those directories may each be shared independently with any other Crædl researcher with read and/or write permissions.

When a researcher is ready to upload a file, s/he navigates to the desired directory in the file browser and selects upload. Crædl pre-populates the webform presented to the researcher from the two levels of metadata schemas, including all default metadata values that have been defined in the project, and supplements it with metadata inferred from the research group management system. The user may choose to modify these fields or, if they accurately describe the file, accept the defaults as they are. All the user must do is select the file to upload and submit the webform. The researcher may share each file independently with any other Crædl researcher with read and/or write permissions.

2.3.1 File transfer

The file that the researcher is uploading must be transferred to the Crædl file server. If the file is small, it can be uploaded directly through the browser using the HTTP transfer protocol. If the file is large, it must be transferred using Globus, which is coordinated directly within Crædl using the Globus transfer API[3]. The file server requires authorization from the web server (and, thus, the researcher, via the file permissions s/he has set) before performing any incoming or outgoing file transfer.

2.3.2 File server

The back end file server on which the data is stored is currently a piece of hardware owned by HEMI. This provides peace of mind to HEMI as an institution as they maintain control of the system on which their researchers' data resides. We have designed Crædl to easily adapt to any institutional data storage facility to afford a level of flexibility both to the client institution and the Crædl team. Additional storage may be added as needed and future instances of Crædl for other institutions will likely follow a similar formula.

2.4 Data discovery

Crædl indexes all metadata and exposes it to researchers through a search form. Crucially, a researcher can only see search results to which s/he has been granted read and/or write permissions. Once a researcher selects a file, s/he is presented with the associated metadata and is provided the option to download the file if the researcher that shared the file requested that it be openly accessible. To encourage openness, researchers may share their metadata while requiring that they be contacted for confirmation before another researcher may download the file.

2.5 Federated login

Institutional firewalls present substantial barriers to data sharing between collaborators, especially when file sizes become large. Crædl removes this barrier by providing federated login through Globus's authentication API[4] (which uses CILogon) so that researchers connect to Crædl using their institutional credentials at their institutional login page. Once authenticated with their institution, the authentication is passed to Crædl to complete the login process. Importantly, Crædl never receives any user credentials; they remain safely behind their institutional firewalls. Once authenticated, researchers have logged into Crædl with direct access to their own data as well as the data that their collaborators have made available to them.

3 Future development

3.1 Data archival and publishing

Crædl is currently designed for the management of operational data, but it is a rather obvious extension for researchers to archive some data for cold storage and to publish other data for public consumption. We plan to work towards implementing these capabilities by partnering with existing institutional research data management centers so that the institution will maintain control over and responsibility for the data.

Specifically, we are currently working with the Johns Hopkins University's Sheridan Libraries' Data Management Services (DMS[5]) group to build a pipeline between the two services. Crædl will provide to the researchers a system for auditing the files in their project and choosing which files to archive, which to publish, and which may be safely deleted. Then, Crædl will pass these data sets to the DMS group for long-term storage. DMS will make published data immutable, mint a DOI, and adopt the metadata that Crædl passes along.

[3] `https://docs.globus.org/api/transfer/`
[4] `https://docs.globus.org/api/auth/`
[5] `http://dms.data.jhu.edu`

3.2 Bulk file upload and client synchronization

Crædl is currently capable of uploading one file at a time, but we plan to develop a bulk upload workflow. This will require careful design to ensure the researcher is provided the opportunity to efficiently verify the integrity of the metadata on each individual file.

As a related note, we aim to develop a Crædl client application that will synchronize a chosen directory on the researcher's computer with the Crædl servers. There is clearly much room to expand automation capabilities to more efficiently assist researchers with their data management in this way.

4 Crædl as a service

While Crædl is currently a prototype designed specifically for HEMI, we have written it so that it may someday benefit other research institutions. Depending on the rate of adoption of Crædl within HEMI in the final quarter of 2017, we will begin to assess the interest in providing Crædl as a service to the larger academic research community. We anticipate that Crædl will help researchers stay better organized and that better organized researchers will be more productive; we hope to soon begin to investigate this hypothesis to quantify the effects that Crædl has on researcher productivity.

I would be excited to hear from you should you find value in the ideas you have read here. Please contact me directly at `adam@craedl.org` to provide any feedback you might have.

Crædl

Collaborative Research
Administration Environment
and Data Library