|  |
| --- |
|  |
| NCSA Report on User Requirements |
| High Level Review of Scientist Feedback for NDS and MDF |

|  |  |  |  |
| --- | --- | --- | --- |
| **Revision** | **Date** | **Author** | **Comments** |
| 1 | 7/28/14 | Mario Felarca | Initial compilation of notes and requirements |
| 2 | 7/30/14 | Mario Felarca | Continued refinement and analysis of notes |
| 3 | 7/31/14 | Mario Felarca | Incorporation of some initial reviews and final touch ups. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[**1.0** **Overview** 3](#_Toc394589215)

[**2.0 General Requirements** 3](#_Toc394589216)

[2.1 Ease of Use 3](#_Toc394589217)

[2.2 Flexibility in Integration 3](#_Toc394589218)

[2.3 Security and Access Control 4](#_Toc394589219)

[**3.0 Collaboration Requirements** 4](#_Toc394589220)

[3.1 Project Tracking 4](#_Toc394589221)

[3.2 Comments and Review 4](#_Toc394589222)

[3.3 Contributions 4](#_Toc394589223)

[**4.0 Data Management Requirements** 5](#_Toc394589224)

[4.1 Versioning and Metadata 5](#_Toc394589225)

[4.2 Accessibility and Watermarking 5](#_Toc394589226)

[4.3 Publication and DOIs 5](#_Toc394589227)

[4.4 Long Term Curation 6](#_Toc394589228)

[4.5 Extensibility in Data Types and Sizes 6](#_Toc394589229)

[**5.0 Workflow** 6](#_Toc394589230)

[5.1 Data Manipulation and Analysis 6](#_Toc394589231)

[5.2 Organization 7](#_Toc394589232)

# **Overview**

This is a document that will review the results of discussions with Dane Morgan, from the University of Wisconsin at Madison, and Bill Wilson, from the University of Illinois at Urbana-Champaign. The goal of these discussions was to gather information regarding how materials scientists conduct their operations, what their typical workflows look like, and how their data is managed during research, publication, and post-publication.

Elements discussed included what tools and processes they currently use, what works well, what doesn’t work, and how things could be improved. Additionally, there was a focus on what they saw as their immediate needs, in terms of efficiency gains and other low-hanging fruit, as well as a higher level perspective that looked at the long term strategy, and where they want to see data management moving in the future.

These discussions occurred in one on one meetings and they are seen as simply the first step in what will be an iterative process to refine high level requirements and user stories out of this data.

# **2.0 General Requirements**

These items pertain to generic elements that are seen as necessary for getting scientists to use the tools and services that NDS/MDF may provide.

## 2.1 Ease of Use

Currently, the tools that scientists use to collaborate or to otherwise handle their data fall into one of two categories. Specialized tools, that are designed to specifically interact with materials science data, or basic everyday tools that a scientist would be using anyway, even outside of their research.

In order to entice a scientist to use our software rather than their existing processes, there should be a very low barrier to entry for the system. Examples of this include single sign-on, the ability to drag and drop files, and intuitive controls and interfaces.

Additionally, the system must also provide functionality beyond what currently exists, in order to justify a user switching from their existing systems to the NDS/MDF.

## 2.2 Flexibility in Integration

The workflow that a given team uses can vary from group to group, or even from project to project within a group. In some cases, it will be more structured, with gated checkpoints at certain intervals, while in other cases it can be quite organic and free-form.

In order for the system to be useful to the largest group possible, it should be designed in such a fashion as to not necessarily dictate a workflow on the users, but instead be flexible enough to adapt to how the users want to leverage it.

Services should be accessible whenever a user or group would find it most useful, and access should not be constrained in such a fashion where the system imposes its own flow, or tries to otherwise change how the existing user or groups operate.

## Security and Access Control

In order for users to embrace something new, they will need to trust that the system will securely and accurately manage access to their data. This includes safely storing it, allowing it to be organized in a way that is intuitive and can be easily found, as well as providing the means to control who has access to the data.

At all times, the owner of the data should be the one in control of access, where it is linked to, and how it can or cannot be referenced or accessed. The system should not impose levels of access, but should be extensible enough to handle the various ways in which curated data could be shared. Examples of this could be by user, by group, or public.

# **3.0 Collaboration Requirements**

Existing collaboration is very ad-hoc in most cases currently. They are stitched together with standard everyday tools and are not always specifically designed to best serve the needs of scientists. In most cases, collaboration within a group, whether local or remote, occurs in the same fashion. There may be slight differences in how data is exchanged, depending on expedience and simplicity. This section details some requirements that could address the weaknesses in this area.

## 3.1 Project Tracking

Currently projects are tracked very loosely, using independent documentation or simply by a given presentation. Oftentimes, this material is not versioned, and there is no way to reliably link it to data and other supplemental information, as the materials evolve leading up to publication.

If the system could provide services that allowed this type of tracking and versioning, the benefits in curation and accessibility would provide ample motivation to move off of the existing basic tools, such as email and Dropbox that projects currently use for this purpose.

## 3.2 Comments and Review

As projects progress towards publication, often collaborators will provide feedback to each other, or review the contributions of other members of the group. Currently, a lot of that information is only exchanged over email, and is often not preserved after the project moves on past publication.

The ability to curate these elements would provide efficiency gains during the actual project collaboration phase. Furthermore, these artifacts could be made available as supplemental and searchable data that is associated with the project after publication.

## 3.3 Contributions

Collaboration often involves contributions by various members of the project group. These items can be calculations, datasets, and analysis among other things. These are usually exchanged via simple methods, such as email or dropbox or physically on a flash drive. Typically, there is no management over these elements, other than by manual accumulation and aggregation.

Having these actively managed by the system would provide another level of provenance for the project, allowing for better reproducibility and facilitating the ability of a collaborator to see where they have made contributions and what they have contributed. This information can often become lost over time, especially as a project moves through publication.

# **4.0 Data Management Requirements**

There are many different types of raw data that are used when initially collaborating on a project and prior to publication, however during actual publication, typically very little of it is associated directly. The data is typically turned into graphs or tables or represented as images that summarize the data and results. How to enhance the usage of the raw data is what this section covers.

## 4.1 Versioning and Metadata

Raw data sets, as well as the figures and tables that are part of a publication, often are not versioned, nor do they have much in the way of description outside of the context of the publication itself. Services that would allow for the addition of metadata would aid in providing greater context and information about the elements.

Versioning services would also allow for greater tracking of how the publication was developed, who contributed specific elements, and when those elements were changed. This would allow for greater reproducibility, as well as enhancing collaboration.

The combination of these two features would also make it much easier for the scientist to decide which elements are important enough to make accessible on a larger scale after publication.

## 4.2 Accessibility and Watermarking

Currently, very little raw data is actually linked to a project when it is going through publication. The data elements that are directly associated are graphs, figures, tables, and images.

Access to those elements, both raw and processed, is currently extremely limited and very manual. Services that would allow these elements to be accessed easily after publication is something that would be very useful, both for the original project members as they move on to other projects, as well as to other researchers who would be led there through the publication.

In both cases, the important part will be ensuring that the system allows for simple and intuitive access. Elements should be easily found by searching for community elements, metadata, or keyword for the specific elements themselves.

Additionally, since the only data that will be accessible will be elements that the owners have decided to share, there must be a method in place that will prevent the data from being simply copied and re-purposed with minor alterations. This can involve auditing who has viewed and accessed the data elements, as well as potentially coming up with a solution that will mark the data in such a way that the source is identifiable.

## 4.3 Publication and DOIs

DOIs are usually assigned by publishers, and the manner in which they are used sometimes varies from project to project. With data elements being collected and managed actively by the system, services can be provided that will allow for greater usage of DOIs.

The granularity at which DOIs will be assigned may need to be reviewed, however this type of interface would allow for specific datasets to be tagged with DOIs, instead of the more generic case of tagging a whole collection of data.

Having the data in a centralized repository that is linked to the publication will facilitate the scenarios where a researcher needs to update the data elements, but doesn’t know where they have placed it or how many copies of the data exist. This will also allow the researcher to decide if the old data should be preserved alongside the new data.

Discussion still needs to take place with respect to how DOIs should be used when it comes to versioning or updating data that is already assigned to a DOI. Should the data attached be immutable, or should it simply be a reference that links a publication to the elements.

## 4.4 Long Term Curation

In most current cases, after publication, the only elements that are actively preserved are the items that were directly linked, or the items that were uploaded as supplemental information. Much of the process and raw data that was used in creating the publication is typically lost or not uploaded with the publication.

A researcher could decide that these data elements would be useful if the services were in place to allow for them to queried and accessed in relation to the publication. This functionality would also provide an organized archiving system which is lacking in most current projects.

## 4.5 Extensibility in Data Types and Sizes

A key feature of the NDS/MDF system is that it should be able to handle the storage and curation of a variety of data, both in terms of size as well as format. The services should be able to handle large data sets, on the magnitude associated with astronomical data, as well as handling smaller, more fine grained collections of data, that are more typical with materials sciences.

This type of functionality will also allow for greater usage of the services during earlier stages of a workflow, when collaboration and organization of the data and their derived elements may involve much smaller collections.

Flexibility in terms of the types of data that can be aggregated, managed, and securely stored as a collection that is linked to a project or publication is also a high priority. It was stressed that different domains will have potentially much different requirements in this regard. One example that was given was how diverse and “messy” materials science data is in this regard, compared to genomic data.

# **5.0 Workflow**

The software and services provided by NDS/MDF should integrate with or enhance the existing domain’s workflow, and not try to shape it beyond cases where there are clearly large gains to be made in terms of efficiency. Many of the previous requirements detailed probably overlap with this section at some level, but there are a couple of specific cases that merit their own points in this section.

## 5.1 Data Manipulation and Analysis

During collaboration, as the project is progressing towards publication, there is raw data that is generated that is often operated on independently by collaborators, utilizing domain specific tools, to perform analysis and transform the data into different formats.

These types of actions are not audited currently, and it can become difficult to follow the chain of events backwards to see how certain elements were derived. The ability to incorporate some services that can be used on-demand, or out-of-band, with respect to the main publication flow, would allow these elements to be aggregated and managed along with the main publication elements.

Where possible, the analysis itself could be integrated with the service, thus allowing for a more seamless transition from the existing workflow, while adding the benefits of security, curation, and metadata collection to these intermediary stages.

## 5.2 Organization

How data is kept within a project will vary from group to group, but the typical focus points for organization and management revolve around publications or presentations. Providing services and interfaces that allow for a user or group to easily prioritize and re-organize data at these key times should be a baseline requirement.

Functionality such as versioning or taking a snapshot of the project’s current repository, adjusting how the data and presentation/publication are linked together, controlling the scope of who will have access, and specifying how that access will be monitored, are all on the high priority list for these types of gating events.

An additional feature could be the requirement of a minimum level of metadata information when elements are added to the project repository. This will allow for greater context whenever the project hits one of this points and the elements need to be re-organized.

The ability to manage raw data and the derived elements in a simple yet fine grained manner should promote the usage of the system and services. In this way, larger and larger amounts of the supplemental information will be made available for curation and storage, thus making it accessible to a wider amount of people through the other features that are part of the NDS/MDF system.