Dryad System Curation Proposal

Introduction

Dryad is being designed as a "catch-all" repository for numerical tables and all other kinds of published data that do not currently have a home. A major design consideration with these data is to avoid placing an undue burden of metadata generation on individual researchers while at the same time capturing sufficient metadata to enable data discovery and reuse. Dryad’s curator will also reduce the burden upon depositors by ensuring the accuracy, completeness, and consistency of author-provided metadata. Furthermore, the curator will provide a service to authors by ensuring the permanence and usability of their data through the use of appropriate tools and preservation metadata.

Purpose of this document

The purpose of this document is to direct the curatorial management of data and metadata in Dryad, and will offer some ideas for overall policy and requirements. This document will guide the curation of individual submissions as they are received, as well as guide the ongoing curation of all objects stored in Dryad. Additionally, this document includes a proposal for the design and functioning of a curator interface that will facilitate the various responsibilities of said curator.

The Digital Curation Centre defines data curation as: “The activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future.” Data curation can also include managing vast data sets for daily use; updating it to keep it readable, etc. Therefore, the term data curator is applicable to a large range of professional backgrounds, from minimal management of digital materials, to the addition of metadata, to managing institutional repositories.¹

Dryad’s curator will be an information specialist who has knowledge of and experience working with information standards. Overall, their responsibility will be the preservation, archiving, and the provision of access to digital data stored in Dryad. Furthermore, the curator will maintain documentation of cataloging and curation practices. Their work, as judged by the quality of the data in Dryad, will be assessed by quality control measures such as: completeness, accuracy, conformance to standards and user expectations, and timeliness (i.e., the use of current terminology).

Current status of DSpace

DSpace is an evolving system, and currently updates and changes are being made to the architecture in order to better facilitate curation tasks. However, the immediate

¹ http://www.dcc.ac.uk/FAQs/data-curato...
The DSpace administration UI is very poor. It doesn't expose all the curatorial functions needed for the system. It provides no way to define local policies for the DSpace-based service or means of enforcing them. It has no serious reporting infrastructure to inform curators about the state of the archive and its contents. It provides poor tools for editing metadata about collections, items and bitstreams, nor does it have good support for withdrawing and deleting items from collections. Some provenance metadata is captured by the History system (which is undergoing significant work at the moment), but this metadata is largely unavailable to curators as a means to help them manage the archive. While this may not have major implications for the DSpace architecture, any significant development effort should include a review of the functionality and user interfaces provided to this critical part of the system.²

Therefore, the typical curation tasks that will be detailed in the following sections will provide system requirements that will also necessitate changes made to the architecture of Dryad.

**Overview of curation in Dryad**

**Prevailing rules and policy**

1. **Issue of ADDING metadata**
   a. Metadata can be added if an optional field is not completed.
      i. If the metadata is KNOWN, for example, if there are clearly keywords associated with an article, and the depositor did not include them, then the curator will add them.
      ii. For metadata that is not explicitly known to the curator, for example, other subject keywords, or temporal/geographical metadata, this will only be added for “high use” or “high quality” data packages, as determined by use and prominence.

2. **Issue of EDITING metadata**
   a. This should be considered an "ownership" issue – the depositor “owns” the metadata they create.
   b. The curator can correct what is "obviously" wrong – for example, misspellings, metadata in the wrong field, etc.
      i. Feedback from management board: author/depositor provided metadata should be edited by the curator.
   c. Significant changes to author-supplied metadata will have to be done AFTER contacting the author. What constitutes a major change may have to be considered on a case-by-case basis. DELETING metadata, for example removing a keyword, will always be a significant change requiring contact with the depositor.

3. **If a file is in a proprietary format, it should be converted to the "least common denominator" format - and something that is ideally open source.**

---
Minimum Criteria for Curation

For BOTH data packages and publications, it is important to double check if the right metadata has gone into the appropriate fields. Also, it is important to check for correct spelling and punctuation in all metadata fields. This task can also be accomplished with the use of an external spell check tool, and this tool should be able to be implemented in batch operations. This is the absolute minimum amount of effort. Changes made to the current system architecture will allow the curator to enhance the value of a data package WITHOUT actually expending much for effort – for example, having a stored name authority file will mean that full author names can be stored in Dryad, but the curator will only have to consult the file.

For preservation purposes, the minimum effort entails checking to see if the formats are valid, that the files can be opened and used, and that proprietary formats are converted to the “lowest common denominator.”

Deposition Process - July 2009

1. Describe the publication.
2. Upload and describe the associated data packages.
3. Approve data packages for publication.

Deposition Process - future

1. Select the journal in which the article appears using the dropdown menu.
2. If the journal is partner journal, enter the manuscript number. This will automatically prepopulate the metadata fields for the publication corresponding to the manuscript number.
3. If the article is not from a partner journal, and "Other" is chosen, the next step is to describe the article in as much detail as possible.
4. Upload data packages and describe them. They can be uploaded individually, or all together as a zip file.
5. Edit the publication metadata, or choose to finalize the submission.

There are two current submission workflows that the author/depositor would undertake: one, where the publication metadata is automatically imported, and two, where the publication metadata has to be manually entered before upload of data.

Required fields to describe the publication: title, authors, journal name. Required fields for the data package: title (authors, keywords, etc. inherited from publication).

Places where curator enhances metadata creation, and tasks that would be beyond the MINIMUM effort of curation:

1. Confirm accuracy and correct author-created metadata, metadata from journal – could require EDITING depositor metadata
2. ADD optional metadata
3. SUPPLEMENT author-created metadata, i.e., add more keywords, etc.
4. Use tools, controlled vocabularies to enhance metadata

Curation tasks

Steps to Curate a NEW Submission

The curator is not going to be ADDING metadata to the majority of objects in Dryad unless they are deemed of high use (i.e., popular), or of educational value, rather they are simply checking for accuracy and completeness. Ideally, metadata from partner journals will not need to be double-checked unless an error is brought to the curator’s attention. The curator will, however, need to find the DOI for articles once they are published, and add this information to the publication metadata. New submissions will remain in a queue until published by the curator.

The integration of HIVE will also reduce the curator’s time significantly – not only will multiple controlled vocabularies be on hand for the curator, but the accuracy of depositor-supplied metadata will be ensured.

There will be two major workflows for new submissions: one for the publications, and one for the data packages. However, they are both connected since data packages belong to publications, and therefore belong to one large workflow. Since the publication metadata is either created manually first, or metadata is automatically received from journals first, the publication curation workflow will necessarily take place before the data curation workflow.

PUBLICATIONS – articles that are NOT from partner journals

Steps that the curator will follow:

1. Confirm metadata supplied by depositor with original publication metadata.
a. Currently, the DOI must be used to find the original publication metadata, and the process would have to be accomplished manually. If the DOI could be used to query ISI Web of Science and/or PubMed, then the metadata could be provided to the curator and compared in an automatic fashion. Please see system requirements.
   i. Double check the author's names - if there is only a first initial, other sources will need to be referenced in order to find the full author's name. *This should be taken care of by a NAME AUTHORITY FILE.
   ii. Double-check the TITLE, JOURNAL NAME, the SUBJECT KEYWORDS, CITATION information (year, volume, issue, pages, etc.) - check for accuracy.
   iii. THE CITATION STRING for the ORIGINAL ARTICLE: we have not yet chosen a standard for this - in the meantime, we will be using the AMERICAN NATURALIST's citation format, for example: Belyea, L. R., and J. Lancaster. 1999. Assembly rules within a contingent ecology. Oikos 86(7):402–416. The name of the journal must be spelled out. *This should be taken care of by having ALL citation metadata stored in Dryad, so that we can automatically generate a sample citation string.
   iv. Is there a corresponding author? This information needs to be confirmed, and added if necessary.

* NOTE: currently, if changes are made to the publication metadata, it will not automatically be inherited by the associated data packages. Any changes, then, must be made to the data object's metadata as well. This should be changed ASAP, and is listed in the SYSTEM REQUIREMENTS list at the end of this document.

PUBLICATIONS – articles that are from partner journals

In most instances, the DOI will not yet be available, and the publication date will not be known. This information will have to be found and added by the curator. Ideally, the journals will send the curator a table of contents with a list of DOIs on a regular basis. This may have to be done manually by the journal (i.e., a person has to remember to do this), and the curator will also have to maintain this aspect manually, although the list sent to the curator should be able to be parsed using a script. The publication date has to be known so that embargoes can be set, and the setting of such dates will be manual. This process will only have to take place, of course, if the depositor chooses an embargo date – the default will be no embargo.

Even if the DOI is not available, or if there is an embargo, the METADATA will still be available to the public and should be published by the curator, and therefore the items will have to be checked by the curator for accuracy and quality. To the public, messages should be displayed that convey the correct status of the item, for example:

- available now
- under embargo, embargo will end at a certain date
- under embargo, date available is unknown

DATA FILES
Data submitted to Dryad will need to go into a queue until the curator approves it for publication. Metadata is inherited by from the publication, but the depositor can change or add to it. Depositors also have the option to create a unique title and add a description to a data file. Furthermore, the embargo is applied to the data file, if the depositor chooses one.

1. For each data package, double check that the same metadata from the publication has been inherited correctly. *This step will not be necessary when inheritance is implemented properly.
2. Double check the FORMAT of the file - is this correct? Can it be downloaded/opened/etc.? *Here is where TOOLS come into play (see below).
3. Embargoes – the default will be no embargo. If the depositor chooses an embargo date, it will have to be based on the publication date. If the article is not yet published, the data file will have to stay in a queue until published. The journal must contact the curator with tables of contents that show publication dates. Upon receipt, the curator will manually change the embargo date, and then publish the data. Currently, the embargo is set in the “Edit this item” feature, and the display is as follows:

New submission – publication curation workflow
New submission – data curation workflow
Ongoing Curatorial Responsibilities – all objects in Dryad

Data packages of special educational value will receive extra curatorial attention and be presented for student use through a dedicated education section of the repository. The curators will also target a limited number of data files of special data packages (i.e. those that are frequently downloaded by users, or those that are particularly suitable for educational purposes) for a higher-level of curatorial attention.

Data curators will select a limited number of data packages (1-2 per year) to receive extra curatorial attention, based on popularity or thematic area. Preference will be given to data likely to have strong resonance with students (on topics such as the evolution of
antibiotic resistance or viral pathogenicity, domestication of companion animals, human origins, origin of life, etc). Curators will work with authors, and with the NESCent Education and Outreach Group, to provide detailed metadata, more extensive background and related material, and a set of suggested exercises appropriate for each data package. Resources will be targeted at the Advanced Placement, college, and graduate levels.

In this case, the curator will be likely ADDING metadata to the submission. This will entail the utilization of in-house tools that identify keywords, and the utilization of controlled vocabularies. HIVE will provide easy access to these controlled vocabularies.

Ongoing tasks will also involve the utilization of tools for preservation purposes. Please see below for details about curation tools. Curation will be assisted by custom software for metadata quality assessment, as well as existing software such as JHOVE and Xena for format validation and migration. The curator will be notified by the system when it is time to implement migration, for example, because the task will have been put on a schedule.

The curator will also study and incorporate methods for automatic measurement of metadata quality by drawing on and extending work of the AMeGA and Infomine projects. Dryad will provide an empirical measure of metadata quality for each metadata record using a variety of metrics (for instance, the match rate between a document and a controlled vocabulary). The rating will help the curator determine which metadata records require review and whether the original depositor needs to be contacted.

**Other curatorial responsibilities**

1. Participate in the metadata generation and quality evaluation studies.
2. Communication with authors/journals when problems arise, helping to verify the usability of metadata, and presenting tutorials on the use of the repository at the annual meetings of the consortium societies.
3. Dryad tutorials, designed for active investigators in the field, will be prepared by the data curators with the assistance of other project personnel, and presented at the scientific conferences deemed most appropriate by the MB (2-3 conferences/yr). The aim of the tutorials will be to explain the role of NESCent relative to the journals and specialized databases, to demonstrate the deposition and retrieval interface, and to assist authors in increasing the extent and quality of the metadata provided by raising their awareness of metadata in general.

Curators will **not** be expected to validate the biological correctness of the data itself, or to determine the completeness of each data package. Furthermore, the curator will not be checking each individual file to validate it, unless the curator is notified by the system, a Dryad user, or by a tool that there is a problem.

**Other notes**

As of July 2009, the plan is to hire both part-time student curators and a professional curator. Therefore, certain tasks will be delegated to the students that are simpler and easier, while more complicated tasks will be given to the professional curator. This will
entail splitting up the workflow further, and will also have implications for the design of the curation interface. Please see system requirements below.

**Curation tools**

The Dryad curator will need to have a “toolbox” on hand to assist them with their tasks. The tools they will need must help with the following:

1. Identifying data (for example, where it is located, what formats it is in)
2. Describing data (for example, batch metadata creation)
3. Manipulating data (for example, data management, data storage)
4. Preserving data (for example, migration)
5. Data registration
6. Documentation of commonly used terms and concepts

**Tools to integrate with Dryad**

Other tools are listed on the wiki, but the most essential or potentially essential tools are listed here for consideration.

**JHOVE – to detect and verify file formats**

It should run on any operating system that supports Java 1.4 and has a directory-based file system. Currently supported formats are AIFF, ASCII, Bytestream, GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, WAV, and XML. Documents are analyzed and checked for being well-formed (consistent with the basic requirements of the format) and valid (generally signifying internal consistency). JHOVE notes when a file satisfies specific profiles within formats (e.g., PDF/X, HTML 4.0).

**USE CASES for JHOVE**

1. Identification
   a. "I have an object; what format is it?"
2. Validation
   a. "I have an object that purports to of format F; is it?"
   b. "I have an object of format F; does it meet profile P of F?"
   c. "I have an object of format F and external metadata about F in schema S; are they consistent?"
3. Characterization
   a. "I have an object of format F; what are its salient properties (given in schema S)?"

It is recommended that JHOVE be integrated with Dryad, however the DSpace community has seen some issues with its performance. Please take into account comments from the DSpace wiki regarding the integration of JHOVE:

Since JHOVE’s format identification functionality seems somewhat unreliable, for now we are sticking with DSpace’s identification based on file extensions, and

---

3 https://www.datadryad.org/wiki/Summer_2009_Curation_Workflow#Aids_to_Assist_the_Curator
4 http://hul.harvard.edu/jhove/
will use JHOVE only for format validation on ingest. Technical metadata extraction will be available only via a command-line tool. Hopefully there will be another tool in the not-too-distant future that will provide more reliable format identification (either Jhove2 or the UK National Archive's DROID tool).

**DROID (Digital Record Object Identification)**

To be used in conjunction with JHOVE. DROID (Digital Record Object Identification)\(^6\) is a software tool developed by The National Archives to perform automated batch identification of file formats. DROID is a platform-independent Java tool, which is freely available to download under an open source license.

**AONS – file format obsolescence**

AONS (Automated Obsolescence Notification System) II\(^7\) notifies repository managers about formats within digital resources in their repositories and alerts them to potential problems relevant to obsolescence and long-term usage. A tool like this is essential for the purposes of migration/conversion (see below).

AONS provides information from authoritative international registries such as LC DFW (Library of Congress Digital Formats Web Page) and PRONOM and the future GDFR (Global Digital Formats Registry).

**LOCKKS – preservation tool**\(^8\)

LOCKSS is a freely available preservation service that works on the principle that by persistently caching multiple copies of a web serials over multiple sites, the chances of that particular object being preserved are greatly increased. It is used by libraries to preserve their content over the long-term. The software is cheap and easy to use, and any institution can get involved.

- It collects content from the target web sites using a web crawler similar to those used by search engines.
- It continually compares the content it has collected with the same content collected by other LOCKSS Boxes, and repairs any differences.
- It acts as a web proxy or cache, providing browsers in the library's community with access to the publisher's content or the preserved content as appropriate. It can also serve content by Metadata (Open URLs) via resolvers.
- It provides a web-based administrative interface that allows the [library staff] to target new [journals] for preservation, monitor the state of the [journals] being preserved, and control access to the preserved [journals].\(^9\)

**Other tools for installing on local machines**

---

\(^8\) [http://www.lockss.org/lockss/Home](http://www.lockss.org/lockss/Home)
\(^9\) LOCKSS has been developed in a library context, hence the language
Migration and conversion

The migration of digital information refers to the “periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation.”\(^{10}\) COST can be a significant factor here. For example, “While it is difficult to predict the frequency at which digital information will need to be migrated, or to accurately predict the costs, the Yale Project Open Book planned for data to be migrated \textit{each five years}. Costs will vary depending on the nature of the digital resource and the aspects that must be maintained. Migration of information raises intellectual property issues and there may be costs associated with this.”

This process will likely be done on a schedule for all objects in Dryad, while simple conversion of a file on an individual level will be taken care of by the curator on a case-by-case basis. What is created is called a “derivative file,” and will be stored in the system in addition to the original file. For example, if a Word Excel file is submitted, it will be converted to tab delimited, and stored. “Storage media refreshment,”\(^{11}\) or the copying of data from one long-term storage medium to another, should take place for all objects in Dryad on a schedule.

The recommended file formats are:\(^{12}\)

<table>
<thead>
<tr>
<th>Textual Formats</th>
<th>File Extensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrobat PDF/A</td>
<td>.pdf</td>
</tr>
<tr>
<td>Comma-Separated Values</td>
<td>.csv</td>
</tr>
<tr>
<td>Open Office Formats</td>
<td>.odt, .ods, .odp</td>
</tr>
<tr>
<td>Plain Text (US-ASCII, UTF-8)</td>
<td>.txt</td>
</tr>
<tr>
<td>XML</td>
<td>.xml</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image/Graphic Formats</th>
<th>File Extensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>.jpg</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>.jp2</td>
</tr>
<tr>
<td>PNG</td>
<td>.png</td>
</tr>
<tr>
<td>SVG 1.1 (no Java binding)</td>
<td>.svg</td>
</tr>
<tr>
<td>TIFF</td>
<td>.tif, .tiff</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Audio Formats</th>
<th>File Extensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIFF</td>
<td>.aif, .aiff</td>
</tr>
<tr>
<td>WAVE</td>
<td>.wav</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video Formats</th>
<th>File Extensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVI (uncompressed)</td>
<td>.avi</td>
</tr>
</tbody>
</table>


\(^{11}\) Copying data from one long-term storage medium to another of the same type, with no change whatsoever in the bitstream (the binary form of the data).

\(^{12}\) From UT Digital Repositories http://repositories.lib.utexas.edu/recommended_fileFormats
Migration can be time-consuming and complex – there will likely be a manual component to this process that is unavoidable. There are numerous tools that are available for specific file formats – for example, image files, music files. These tools may have to be installed and tested to test efficacy, and therefore, the curator may need to utilize numerous tools to accomplish migration.

A helpful tool in this context is **XENA preservation tool**\(^\text{13}\) - Xena can convert any data object into an ASCII representation containing XML metadata, via Base64 encoding. This is known as 'binary normalisation' and is fully reversible when there is a need to re-create an original data object. Xena can also convert data objects into openly specified file formats, such as XML or PNG, in a process known as 'normalisation.' These normalised files may be accessed via the Xena viewer, or exported for use with other applications. It is recommended that XENA be installed for use by the curator.

**External spell checking**

A tool should be able to be called via the Dryad interface for the curator to use to check spelling. This tool does not have to be integrated, but should be able to be available as an external tool or should be installed on a local machine. There are numerous spell checkers available online, and the best one will have to be determined via testing.

Another issue is that of diacritics and special characters. It should be noted that copying and pasting can create problems in display that the curator may have to correct.

**Other preservation tools**

The Metadata Extraction Tool\(^\text{14}\) developed by the National Library of New Zealand. It is designed to:

1. automatically extracts preservation-related metadata from digital files
2. output that metadata in a standard format (XML) for use in preservation activities

The Metadata Extract Tool includes a number of 'adapters' that extract metadata from specific file types. Extractors are currently provided for:

1. Images: BMP, GIF, JPEG and TIFF.
2. Office documents: MS Word (version 2, 6), Word Perfect, Open Office (version 1), MS Works, MS Excel, MS PowerPoint, and PDF.
3. Audio and Video: WAV and MP3.
4. Markup languages: HTML and XML.

If a file type is unknown the tool applies a generic adapter, which extracts data that the host system 'knows' about any given file (such as size, filename, and date created).


Tools to build in house

A metadata extractor that looks for key phrases in the article PDF (e.g., "locations of the specimens") which may indicate data packages underlying the article. The extractor can be configured to search for arbitrary phrases. The extractor can be run by a curator pressing a button on the publication page. The results page contains a list of matching phrases in one column, with a list of data package titles in the second column. Basic implementation: convert PDF to text, search within the text for phrases matching the list of target phrases.

OTHER – controlled vocabularies

Access to controlled vocabularies should be given to the curator – the ability to search them/link to them from the curator interface. Some are as follows:

1. ITIS: Integrated Taxonomic Information System. Here you will find authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world.
   a. See also the Catalogue of Life.
2. uBio: uBio is an initiative within the science library community to join international efforts to create and utilize a comprehensive and collaborative catalog of known names of all living (and once-living) organisms. The Taxonomic Name Server (TNS) catalogs names and classifications to enable tools that can help users find information on living things using any of the names that may be related to an organism.
3. The Biodiversity Heritage Library has a suite of name services available.
4. MeSH Medical Subject Headings
5. NBII Biocomplexity Thesaurus
7. Plant Ontology, The Plant Ontology (PO) has been developed and maintained with the primary goal to facilitate and accommodate functional annotation efforts in plant databases and by the plant research community at large.
8. ERIC Thesaurus
9. Library of Congress Authorities
10. European Distributed Institute of Taxonomy
11. Encyclopedia of Life

SYSTEM REQUIREMENTS

What follows is a list of system requirements necessary to help the curator in their responsibilities. Some of them are “obvious” but they are listed because they do not exist in the Dryad interface as of yet.

1. Ways for the depositor to contact the curator.
   a. The best way for a depositor to contact the curator is through a standard form that the curator can quickly and efficiently parse. This form will be available via the Dryad interface. Form contents will be displayed to the curator in the “In Tray” under “notifications” (see below).
2. A more streamlined, less time-consuming way to ADD a data file to a publication that has already been published. Currently, a "fake" publication must be made, the data file uploaded, and then the fake publication deleted, and the metadata (dc.relation) manually changed to create the relationship between the new data file and the real publication.

3. We need to be able to mimic the inheritance that takes place during deposition - if we make changes to the publication metadata, we should have the option to apply it to all associated data packages. The curator can be asked, "Do you want to apply these changes to associated data packages?", yes or no. This is essential - there will be some data packages that have 30 some files, and if we make a change to one, we should be able to cascade those changes. It is important to note that references, not strings, will be used for this feature.

4. For non-partner journals, the curator must double-check depositor-supplied metadata. Currently, the DOI must be used to find the original publication metadata, and the process would have to be accomplished manually. If the DOI could be used to query ISI Web of Science and/or PubMed, then the metadata could be provided to the curator and compared in an automatic fashion.

5. As of July 2009, the plan is to hire both part-time student curators and a professional curator. Therefore, certain tasks will be delegated to the students that are simpler and easier, while more complicated tasks will be given to the professional curator. This will entail splitting up the workflow further, and will also have implications for the design of the curation interface. Ideally, a student curator would have a unique version of "My Dryad" that will list tasks unique to their expertise, and the professional curator will have something similar.
   a. For the student curator, in particular, there should be a checklist for each object in Dryad that is being curated, where they can confirm that certain tasks have been completed. This will be useful for them, so that they know their progress, but also for tracking purposes, and to maintain quality.

6. Please see the MOCKUPS below for some of the visualizations of the following features:
   a. An "in tray" with tasks and notifications - these will be lists of submissions that are new and requiring curation/approval. Notifications will be like errors, notices of deviant processes, etc. - for example, someone trying to submit an incredibly large data package that is over the storage limit.
      i. New submissions: depositor-submitted metadata for both publication and associated data files should be displayed in its entirety to the curator on one page within the "in tray." This will be useful for metadata verification – since publication metadata is inherited by the data package, but the depositor has the option the edit/delete/add. Since there will likely be multiple data files for a publication, it would be best if each data file metadata set is displayed alongside the publication metadata one at a time. Having the metadata available for comparison in this way will save the curator time.
   b. A batch edit view - this of course will be handled in part or completely by the new version of DSpace. Some other batch/group operations include:
      i. Finding duplicate data objects - right now in the system it is too difficult to find these. It is possible, particularly in the current system, that a file can be uploaded more than once, and once will have to be deleted. A helpful batch process would be to withdraw
a number at once, to also have the option of withdrawing all the
data objects associated with one publication, etc.

ii. Add one READ ME file for a number of data objects - this
SHOULD also be an option for the author/depositor.

iii. Right now, you have a list of data objects with links, and you have
to look at each individual data object in order to edit it.
THEREFORE, batch editing based on metadata field.

c. Curators should be able to view lists of items that need additional
attention
   i. articles that have no associated data packages
   ii. articles that do not have full bibliographic metadata (volume,
       number, DOI)
   iii. etc.

d. Integration with some curator tools - ability to run JHOVE, for example,
from the interface.
   i. Ideally corresponding tools should call each other automatically,
or the curator should be given the option to run an appropriate tool
based on an action.
      1. Example: "File type not recognized by Dryad"
         a. Utilize format verification tools first (i.e., JHOVE)
            and add new file type if it is recognized and valid.
         b. If not valid, other appropriate tools (i.e., XENA, etc.)
to determining whether the file can be converted to
file type that can be used (see also “Tools” mockup
below).

e. List of high profile, high use data files, and those with educational value -
updated continually -> these will require higher curation focus. These can
be listed in the "in tray," and in a "reports" section.
   i. ...we need a section where reports based on stock queries can be
displayed, or can be run and displayed at any time by the curator.

7. Feature: I would like to see a feature like in ContentDM where as new metadata
is entered, it becomes part of a controlled vocabulary - the curator can have the
option to add or delete items from this CV, but this would be a very good interim
solution until HIVE comes into play, and would help with building a name
authority for authors. For example, a depositor adds "Ryan Scherle" as an
author, and the curator sees that this name is in the queue to be added to the
author CV. The curator approves this. The question then is how it is
used/displayed to the user - when they are typing, for example, the name can
appear as a suggestion, as with other CV terms.

8. VERSIONING - here are some of my ideas/recommendations, which would
require system support:
   a. METADATA VERSIONING: The recommendation is to always keep the
original version that is submitted by the depositor - keep all the metadata,
etc. - this can always be reverted back to and/or used as a reference.
Further changes made to the metadata by the curator will not be tracked.
Only the most "up to date" version will be displayed to the users, with the
original version available via the curator interface. CONFIRM: is the
original version the only one that we want to keep, or would be curator
want to look at all the subsequent versions? Can this just be taken care of
by statistics - i.e., is it merely better just to KNOW how many times a
record/field, etc. has been changed?
b. **DATA VERSIONING:** When there are changes/corrections made to the actual contents of the data package, and a version has already been published in Dryad, the NEW version should be considered a new, unique entity, therefore assigned a new unique identifier. The following Dublin Core elements should be used to relate the various versions of the data package: dc.relation.replaces, dc.relation.isreplacedby, dc.relation.isversionof. The actual linking of the data packages via these elements will most likely be done manually, or at least heavily supervised by, the curator if the system can make the links automatically.

**DRYAD’S ADMINISTRATIVE INTERFACE**

This section provides some visualizations of changes that could be made to create a curator interface and are based on the list of system requirements above. Included are some use cases or scenarios that would require the use of certain features and functionalities.

**First page – shown upon login**

The IN TRAY will show the number of submissions needing curation - therefore all of the new submissions. Also, TASKS will include lists of high profile and high use data packages. NOTIFICATIONS will list errors and places where the curator should put their attention. FOR EXAMPLE: a data package has been submitted that is over the storage limit. Also, the notifications section will list messages from users/depositors to the curator.

The REPORTS section will list reports that are the result of batch operations and queries that are run regularly and automatically. Some of them will be basic statistics reports, and some of them will be error reports, and reports that show submissions that are missing metadata, etc.

TOOLS - this is where integrated tools will be listed.

BATCH EDIT is where the curator can go in and search for issues or choose to do unique batch edits - i.e., unique from the batch processes that produce the reports.
“In Tray”

This is probably the most important feature of the curator interface - as stated above, the newest submissions will be put into a "curation inbox" where they will be reviewed before finalizing. I am also envisioning that questions, etc. from depositors will also be displayed in the "notifications" section. "Notifications" would also ideally list errors that were noted during submission, for example, "Publications without DOIs." The listing of high profile and high use data packages will be generated through a query based on use statistics.
Essentially, the various tools that we choose to incorporate will be listed, and ideally they will be run directly from the interface. Available tools come in a wide variety of different languages, platforms, etc. - this has to be taken into account when integrating with DSpace, etc.

1. scenario 1: curator needs to migrate/convert ALL items in a certain (proprietary) format to a non-proprietary one.
   a. Example: there are 20 Microsoft Word documents in Dryad -> convert all at once to TXT.
   b. There are 50 Excel files, they need to be converted all at once to tab delimited.

2. scenario 2: there are a number of file types - those with a certain file extension - that need to be verified/detected, with possible further action taken, depending on the results
   a. Example: are all the .nex files the same as the .nexus files? do they need to be converted?

3. scenario 3: the curator needs to convert metadata about all dryad contents/some dryad contents into XML, and store the XML - for preservation purposes, and this should be done regularly

4. scenario 4: legacy data - are these in outdated formats, can they be converted/migrated? again, look for certain file types/extension - use an external
file format registry to look for outdated formats (i.e., PRONOM, and access should be provided to the curator)
   a. then can we/should we offer emulation services? what do we do with those that can't be migrated/converted?
      i. Recommendation: withdraw (but don’t delete) the submission.

INDIVIDUAL LEVEL scenarios/use cases:

1. scenario 1: a person submits something in Excel or some proprietary format -> the curator can convert it individually
2. scenario 2: a person submits a file with an unknown file extension -> detect and verify this particular file
   a. this should also be a NOTIFICATION for the curator in the "In Tray" - message should be something like "File type not recognized by Dryad"
      i. Utilize format verification tools
         1. ADD new file type if it is recognized and valid.
         2. If not valid, use tools to determining whether the file can be converted to file type that can be used.
         3. If not possible, choose to reject the submission, contact the depositor. If the depositor does not have another file to replace it with, withdraw (but don’t delete) the submission.
         4. IDEALLY this will all be done automatically, or semi-automatically by the system – the system should be able to call appropriate tools in this situation, and prompt the curator with appropriate actions.

The interface as shown here is just one idea – by clicking on the OPEN link, the curator can see a list of various tools at their disposal.
This should function almost like reports function in Access, but not as clunky. Therefore, there will be "stock" queries that are run regularly, or can be run by the curator, and viewed within the interface.

Batch Edit page

Use cases:

1. Bulk edit metadata (e.g. perform an external spell check)
2. Bulk add metadata values (e.g. add an abstract to a set of items)
3. Bulk find and replace metadata values (e.g. correct a mispelled surname across several records)
4. Enable the bulk addition of new items
5. Re-order the values in a list (e.g. authors)

DSpace is working on a batch edit interface, and they asked for input into its functionality. Screenshots are available. Screenshots are available.

---

15 http://wiki.dspace.org/index.php/Batch_Metadata_Editing_Feature
Apparently, this process would involve exporting metadata into a file to be edited on a local machine, and then it would be uploaded again. If this process could be streamlined for Dryad, and the editing takes place within the interface, it would save the curator a lot of time.

An example idea for the interface:

Based on the query, results will be displayed, with further editing options available:

- REMOVE a metadata field OR remove an object
- ADD a new metadata field
- EDIT and existing metadata field
- Etc.