Use of the Hydra/Sufia repository and Portland Common Data Model for research data description, organization, and access

Steve Van Tuyl, Digital Repositories Librarian, Oregon State University, steve.vantuyl@oregonstate.edu
Michael Boock, Head of the Center for Digital Scholarship and Services, Oregon State University, michael.boock@gmail.com
Hui Zhang, Digital Applications Librarian, Oregon State University, hui.zhang@oregonstate.edu

Introduction

We'd like to thank NDLTD for giving us this opportunity to present. For this paper, we will provide some background about Oregon State University's repository and ETDs, discuss our project to migrate the repository from DSpace to Sufia/Hydra, our use of the Portland Common Data Model to represent objects and files in the repository, some of the challenges we are facing in this migration, solutions to those challenges, and next steps.

Background

In July 2005, the Oregon State University Libraries began accepting electronic versions of student theses and dissertations into the institutional repository. A pilot project to deposit Electrical Engineering and Computer Science master's theses began in July 2005 and voluntary deposits from other units were accepted. Mandatory deposit of doctoral dissertations began July 2006 and masters theses deposits were required beginning January 2007. We began digitization of the university's print theses and dissertations soon after and completed that internal project in 2013. Every thesis or dissertation ever produced at the university is now available online, except for two whose authors asked that they not be available. The repository also includes technical reports, research articles, presentations and posters, and datasets (Table 1).

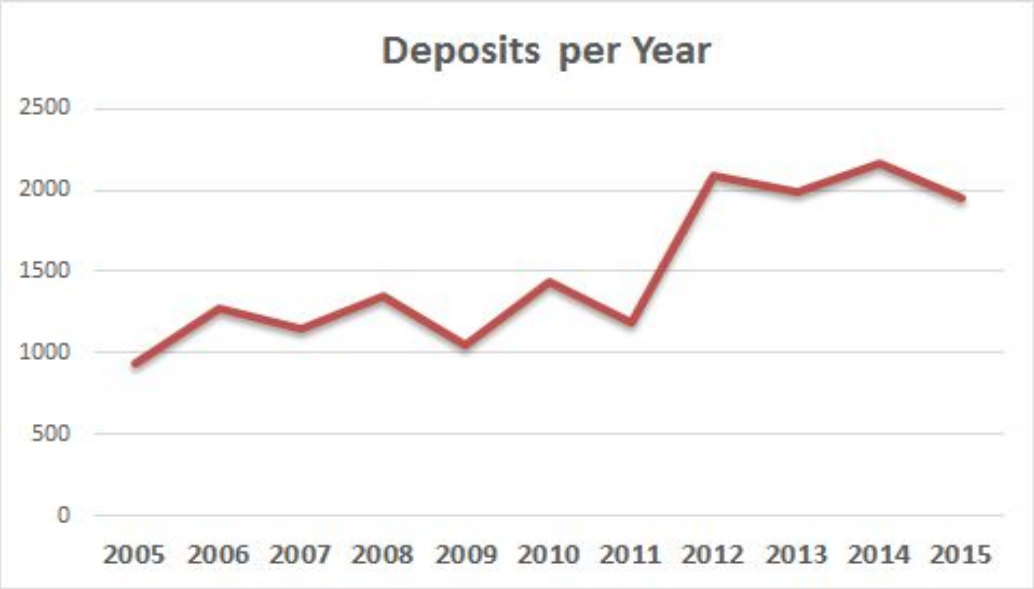
Table 1

Content Types in SA@OSU	
Theses and Dissertations	24,842
Technical Reports	11,185
Articles	7,788
Presentations and Posters	1,041
Audio and Video	159

Datasets	59

The number of annual deposits has increased to around 2,000 items per year since 2012, 500-600 of which are theses and dissertations (Table 2).

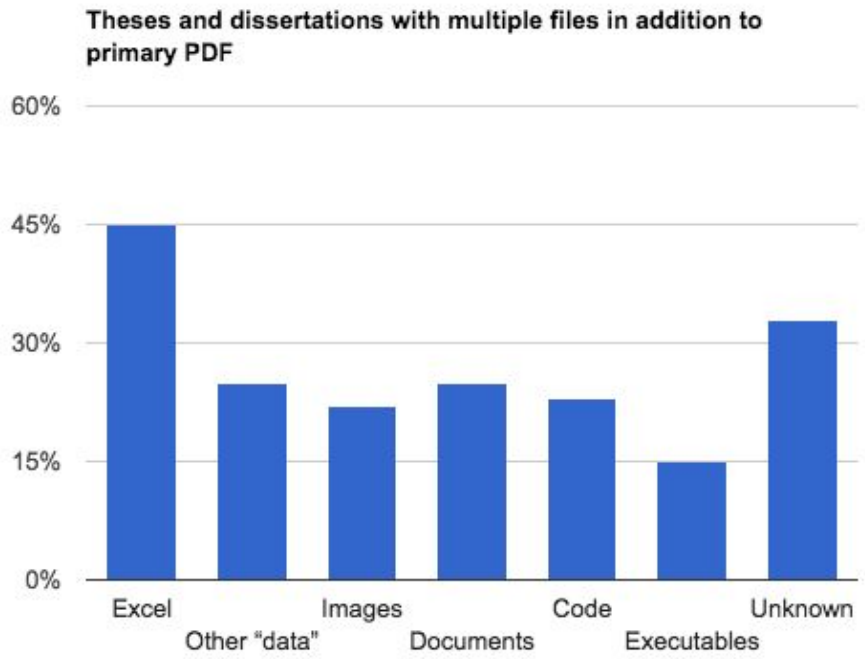
Table 2



Multi-part files in ScholarsArchive@OSU

We know that there are many more datasets in our repository that are not classified as datasets because they are attached to other DSpace items like articles, presentations, and theses and dissertations. Currently, as many as 25% of the total items in ScholarsArchive@OSU have multiple files and 11% of theses and dissertations in the repository have multiple files associated with them (Table 3). These include theses with supplementary maps, images, datasets, and software. 45% of the dissertations in the repository that have supplementary files have excel files. Other additional files attached to the same Dspace dissertation item records include content that wasn't usable, couldn't be opened, or crashed the computer. Or, we were able to identify what it was (for example, a Lotus spreadsheet) but weren't able to curate. Steve Van Tuyl did this work shortly after he arrived at OSU to identify the types of supplementary data housed in our repository. I won't talk about this much more except to mention that I know Gail Mcmillan has done similar research and will be talking about that in another paper today.

Table 3



The reason I make note of this is that OSU will need to decide what to do with these files when we migrate. Do we touch each one of these files to determine what to migrate, how they are each described and related to other items in the new repository? Probably we will migrate these objects and worry about them later.

Reasons for PCDM and Migration

ScholarsArchive@OSU is currently hosted on the DSpace repository platform, as it has been since the repository's inception in 2004. There are elements of DSpace that prevent us from fully engaging with digital objects the way we would like to. One key example is the strict organizational structure. DSpace has a hierarchical and linear form of information organization (communities and collections) that prevents more modern and granular types of interaction with the data in the repository. These restrictions have proven especially difficult in the face of continued proliferation of digital object types we collect, or could collect, in SA@OSU.

In addition to more flexible and reliable statistical reporting--something that many of our

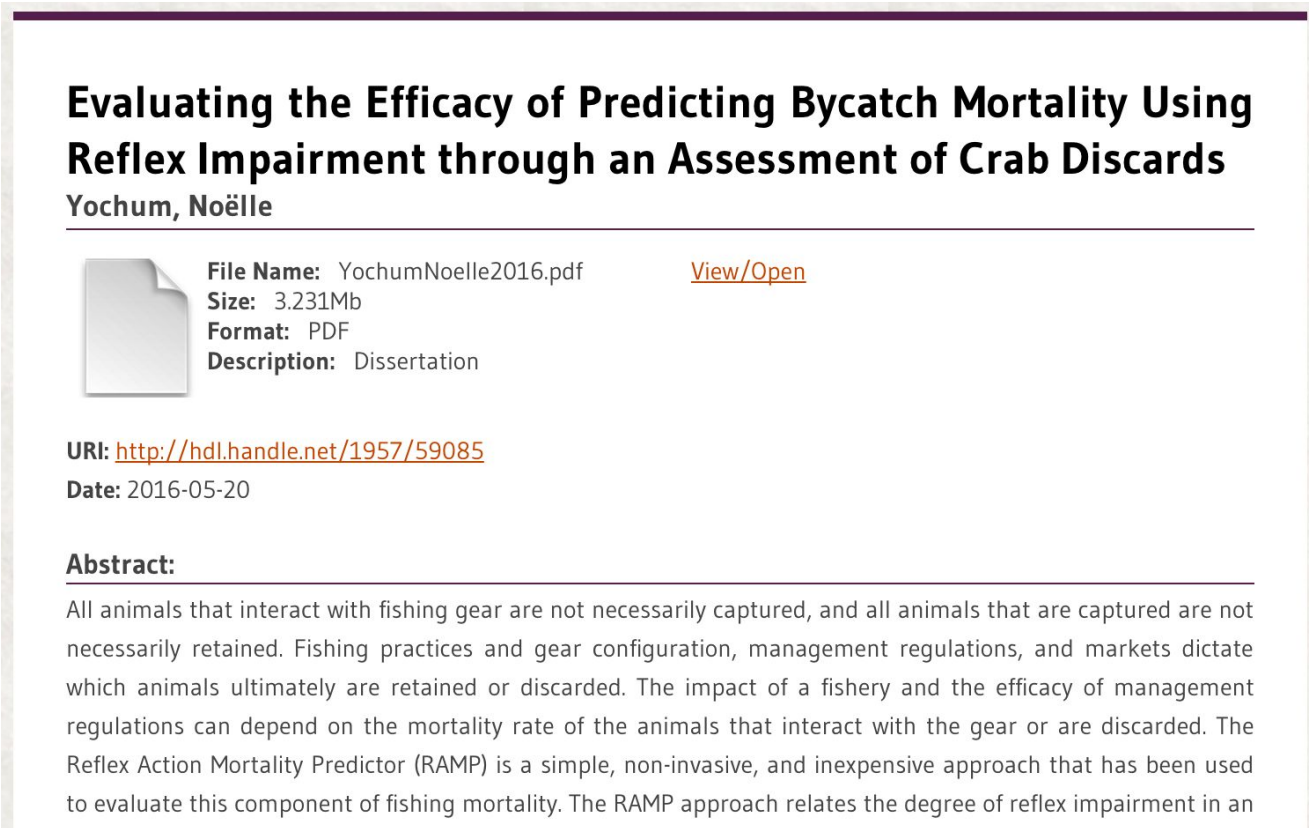
stakeholders require--we also wanted an infrastructure that allows us to better represent the complex structure and interrelatedness of digital objects in the new repository. Deposit of faculty research articles that build on the research contained in ETDs and supplementary datasets is becoming increasingly common. In our current repository, supplementary data and related content files for ETDs are co-located with the thesis document without regard for representation of the relationships among file types or differentiation in the description of files.

The metadata primarily describes the thesis document (almost always a PDF, currently) but does not adequately describe accompanying supplementary files and related documents. This creates a problem for reporting, description, discovery, and reuse of those supplementary files and for contextualizing the research contained in the ETD with other content in the repository and outside of the repository.

Most library development at OSU is in Ruby on Rails. Our digital collections are already on a hydra/fedora platform so our development team has experience with that platform and code base. They have far less experience with the java and xslt code base upon which DSpace is built.

In DSpace, a simple dissertation looks like this (Figure 1). There is a single PDF file. As you see, there is very little descriptive metadata associated with the bitstream itself: file name, size, format (automatically generated by the system), and a description that is optional. This example will look very similar in Hydra/Sufia.

Figure 1



Evaluating the Efficacy of Predicting Bycatch Mortality Using Reflex Impairment through an Assessment of Crab Discards

Yochum, Noëlle

File Name: YochumNoelle2016.pdf View/Open
Size: 3.231Mb
Format: PDF
Description: Dissertation

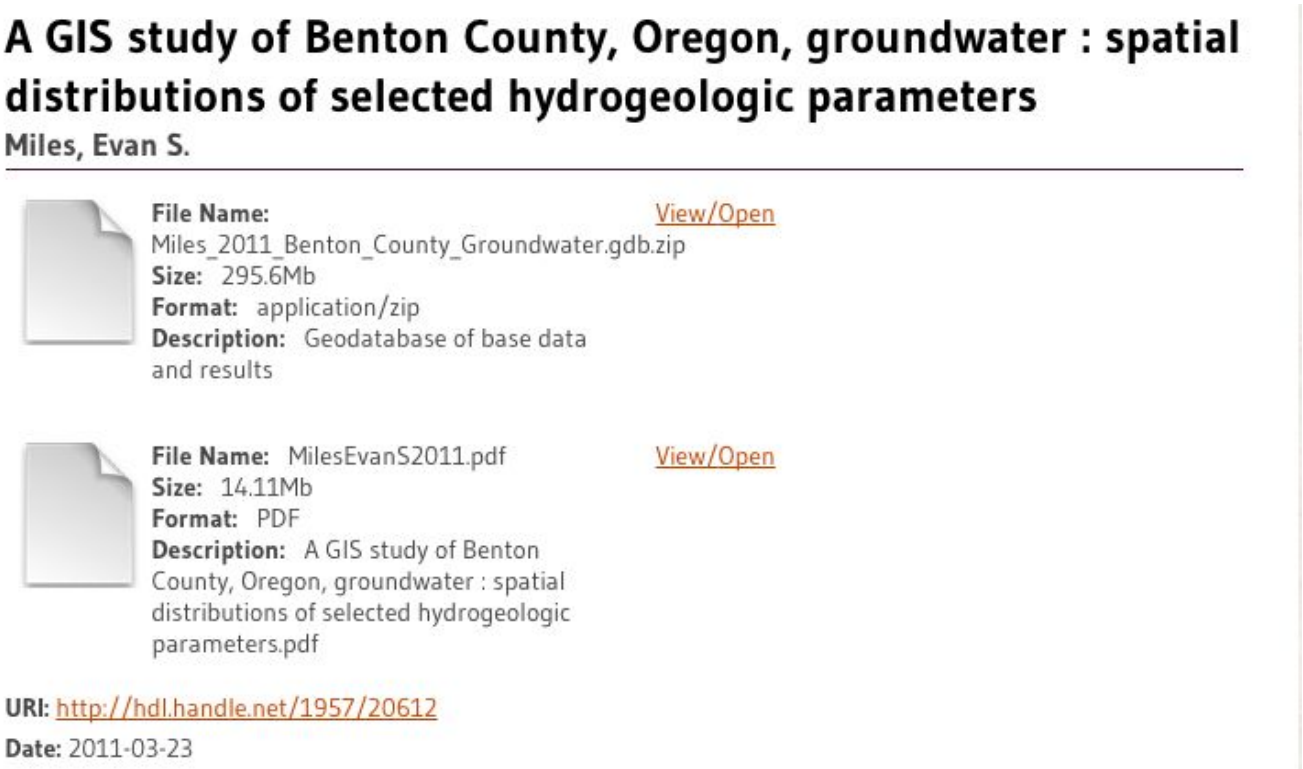
URI: http://hdl.handle.net/1957/59085
Date: 2016-05-20

Abstract:

All animals that interact with fishing gear are not necessarily captured, and all animals that are captured are not necessarily retained. Fishing practices and gear configuration, management regulations, and markets dictate which animals ultimately are retained or discarded. The impact of a fishery and the efficacy of management regulations can depend on the mortality rate of the animals that interact with the gear or are discarded. The Reflex Action Mortality Predictor (RAMP) is a simple, non-invasive, and inexpensive approach that has been used to evaluate this component of fishing mortality. The RAMP approach relates the degree of reflex impairment in an

Ordinarily in DSpace, datasets are included on the same record with the PDF as with this one in Figure 2. A geodatabase is contained in a 300 mb zip file alongside the dissertation PDF. We provide some minimal description to that file here but it sits in the same item record with the PDF and the dataset is only findable if the user gets to this record by way of search or browse of the dissertation metadata.

Figure 2



A GIS study of Benton County, Oregon, groundwater : spatial distributions of selected hydrogeologic parameters

Miles, Evan S.

File Name: Miles_2011_Benton_County_Groundwater.gdb.zip View/Open
Size: 295.6Mb
Format: application/zip
Description: Geodatabase of base data and results

File Name: MilesEvanS2011.pdf View/Open
Size: 14.11Mb
Format: PDF
Description: A GIS study of Benton County, Oregon, groundwater : spatial distributions of selected hydrogeologic parameters.pdf

URI: http://hdl.handle.net/1957/20612
Date: 2011-03-23

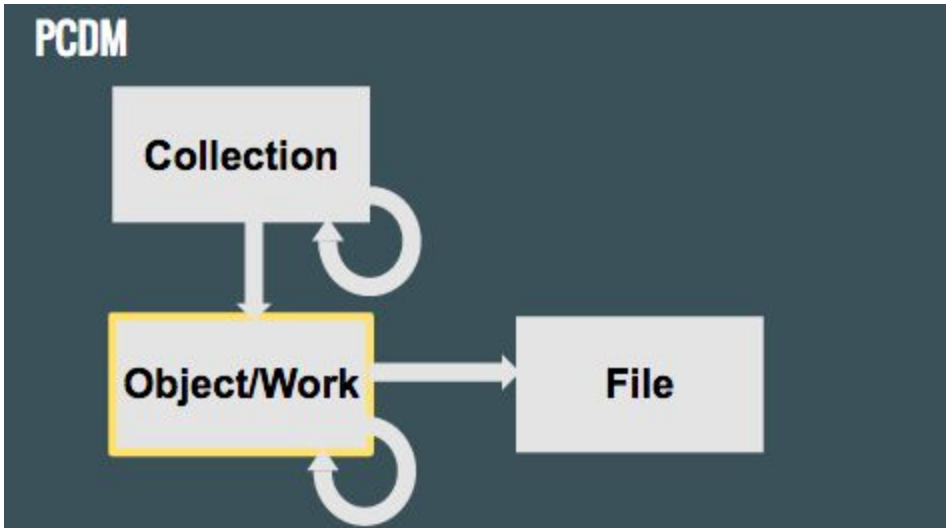
Other times, as in Figure 3, we have a dissertation and a multi-file dataset upon which the dissertation as well as other research (e.g. articles, book chapters) was based. These items all exist in different DSpace collections, without links between them or anything to tell the user how the different items relate to each other. One element of this dataset may have been used for a particular faculty article, another element of the dataset for this dissertation. That cannot be represented except informally with mentions in the metadata. We see PCDM, the Portland Common Data Model, as the answer. Briefly, PCDM is a customizable and flexible domain model that provides an opportunity to describe multipart or compound objects separately, to demonstrate their relationships, and provide links between them.

Figure 3

Portland Common Data Model (PCDM) and Data Modeling

The primary element of PCDM, from the perspective of this conversation, is the Object (or Work). Files are connected to works, but are not embedded in the work itself (as in DSpace and other repository systems) (Figure 4). This provides for flexibility in access controls and structure, so you can set a different embargo or access control on particular files, for example. One of the key elements of PCDM, for our purposes, is the ability to nest works within other works. This will allow us to treat those individual elements of multi part objects individually, describe them separately with their own metadata, but still be able to represent them as part of a whole, intellectual entity in the repository.

Figure 4

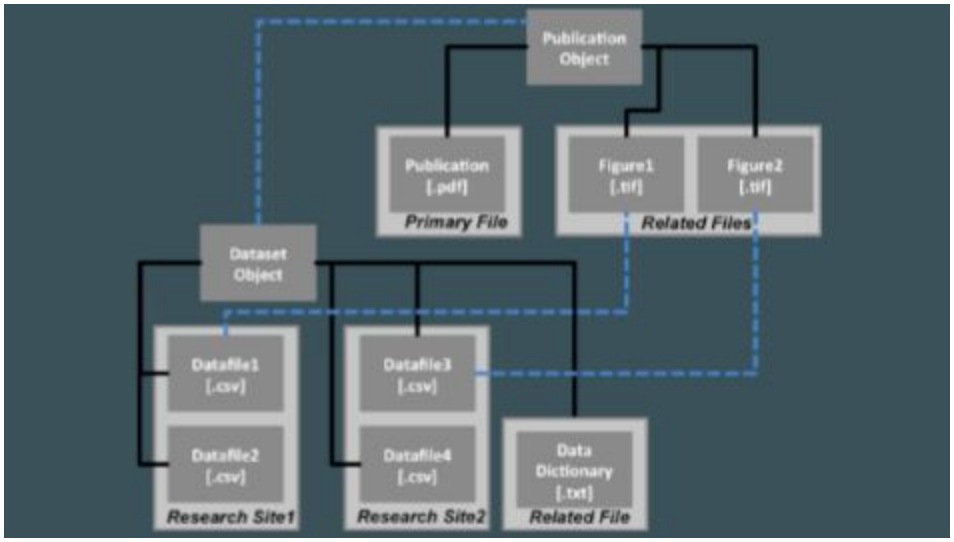


As part of our repository migration process, we identified a number of case studies of compound objects in our DSpace instance in order to conceptually model their structure using PCDM. Figure 5 is a PCDM conceptual model where you have your publication object, let's say for our purposes this is a dissertation. Included in that object you will usually have a pdf as the primary file. You may also have some related files such as figures and tables. This is more common for research article post prints where tables and figures are separate from the text.

In PCDM, the dataset associated with the dissertation is described separately as a dataset object and can consist of any number of files including a readme file and a data dictionary. Each of these files can be described separately or together.

There is a lot of flexibility - in fact, infinite flexibility. We can infinitely nest works within works and represent complex objects in a huge variety of ways. The problem, though, is that there is some value in representing similar objects in the repository in a consistent manner. Practically speaking, this means we need to constrain the flexibility of PCDM to a set of common models for the different types of content in our repository (ETDs, datasets, books, etc.) to ensure consistent representation.

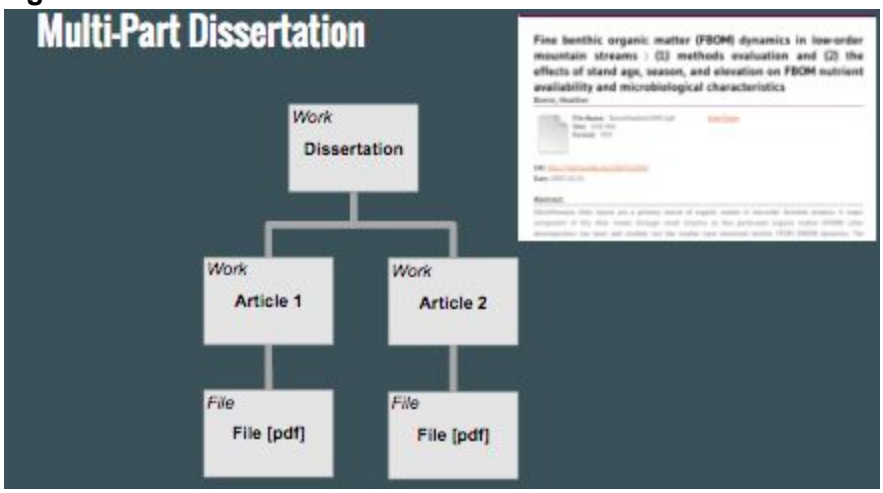
Figure 5



Again, simple objects such as PDFs won't look a lot different in Hydra/Sufia. With more complex objects, such as the one in Figure 6, we'll be able to represent objects currently described as one, separately. We will assign metadata more granularly and represent relationships between objects that formerly either had to be packaged together as a single item, or described separately without being able to demonstrate or express relationships between them (except in notes or by putting items together in collections--which still doesn't make it clear how each item relates to the other)

This is a specific case of a dissertation that is presented to the graduate school as two separate articles, a form of dissertation that many of you will be familiar with and is increasingly common at OSU. In DSpace we are forced to describe this as a single object. Using PCDM, we'll describe the dissertation as a work that includes both articles, and each article as works, describing the articles separately, and be able to show the relationships between them.

Figure 6

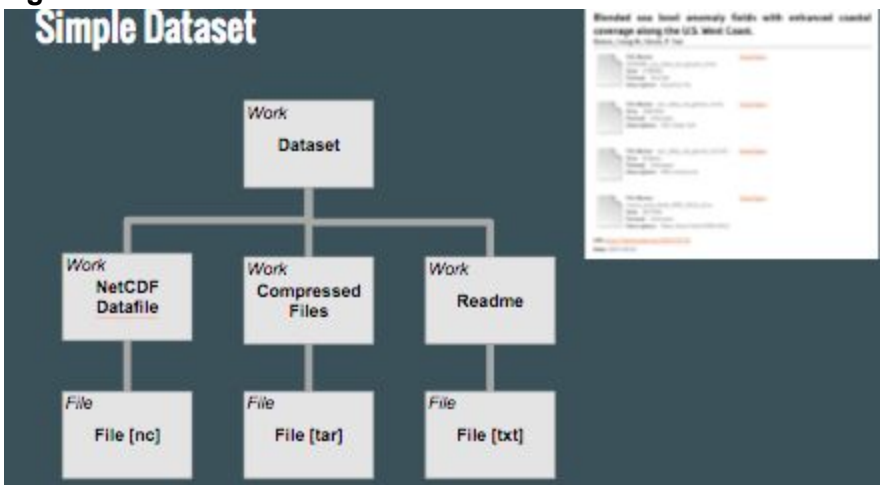


For datasets, in DSpace one is unable to represent hierarchy at the object level. You can only list a bunch of files with very minimal description for each (the automatically generated file

name, size, and format, and a brief description).

The modeling of a “simple” dataset in Figure 7 is a good start towards representing the complexity of a typical dataset in our DSpace repository. In DSpace we were, essentially, forced to package some of the data in the repository in compressed files in order to retain the directory structure the depositor wanted to see. In essence, we package data together to respect the intentions of the depositor with respect to the structure of their data.

Figure 7

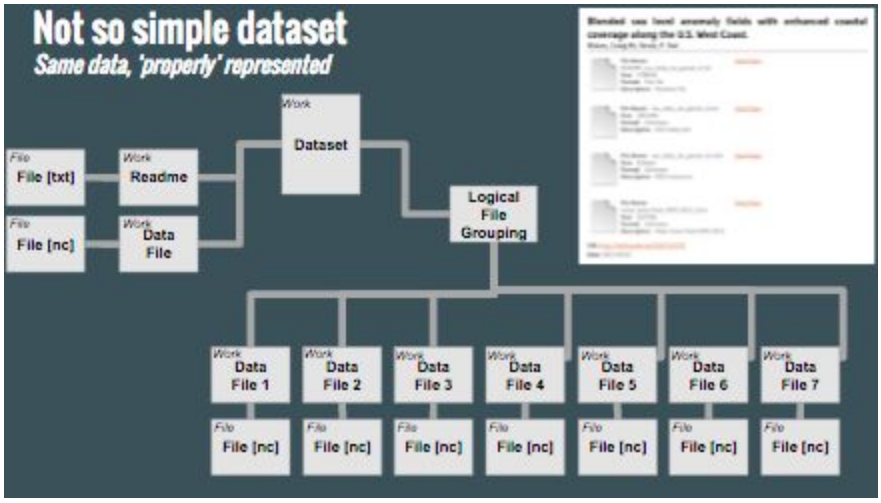


Using PCDM in Sufia, we’ll be able to describe different works separately and demonstrate their relatedness to the dataset as a whole and to other works. In the case of this dataset, we’ll have a record for the entire dataset, a record for the readme, a record for the data file, and a record for each of the data files, with specific metadata assigned to each. If you get to one of the items, through a search or browse, you’ll see the other related items and how each of them relate to the other.

This is the same dataset that we saw on the previous slide. The data is packaged as a compressed file in DSpace which you can see here in the upper right of the slide. PCDM allows us to represent the packaged information in a manner that is more consistent with the depositor's intended structure and in a way that allows flexibility for relating it to other objects in the future at a more granular level.

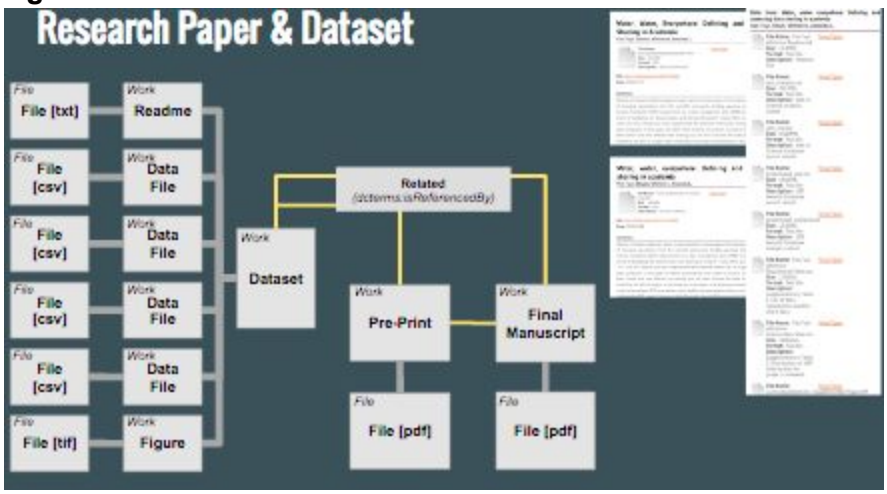
PCDM allows us to apply descriptive metadata to any of the Work level items here - so we can individually describe all of the elements of the dataset (and do things like set access controls on each file) but we can also describe the intellectual entity that is the entire dataset - the central Work in this example in Figure 8.

Figure 8



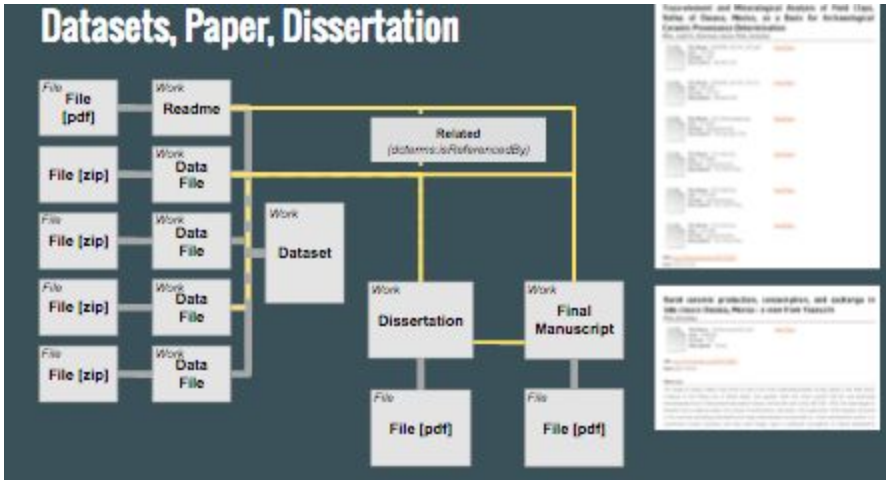
An example of why this is important can be seen in Figure 9 - we have the ability to relate the dataset to two separate documents in the repository, in this case a pre-print version of an article and a final manuscript version of an article. We could also link it to a thesis or dissertation that used the dataset.

Figure 9



More importantly, we have the ability to relate *elements* of the dataset to different objects in the repository. For instance, we can say that two of the data files in this dataset are related to the dissertation, only one of those datafiles is actually related to the final manuscript (journal article), and that there is a relationship between the dissertation and the final manuscript. Follow the yellow lines in Figure 10.

Figure 10



Data Modelling Consistency Challenges

There are a number of challenges we're facing in terms of the consistency of our data modeling and how we apply our data models. We have the difficulty of determining how best to migrate content efficiently while still taking advantage of the flexibility of PCDM.

As we showed earlier, we have a wide variety of item types with different relationships in our repository. How do we ensure that these items and their relationships are migrated so that they are represented in the new repository consistently. For example, all theses with associated datasets need to be represented according to our predetermined model in the new IR and need to be migrated appropriately to make that happen.

Also, as we start to think about interoperability, the community of repository managers and developers who are using PCDM need to come to some common agreement of how to represent types of objects in their repositories. It would be helpful, for example, if there was a common model for how to represent a dissertation, a multi part dissertation, and how to relate a dissertation to content elsewhere in the repository.

Data Modelling Consistency Solutions

For internal consistency, we will identify major object types in our repository and model them as a starting point. These are the models we showed you earlier for single file, multi-file objects. Our focus will be on finding balance between the flexibility that PCDM offers, resisting the proliferation of data models that are out there, and creating models that we can automate for migration.

We also need to understand that, necessarily, we will need to 'dumb down' some of the data modeling for the objects we migrate - we can't touch every individual object as we migrate, so we will have to make some compromises. That said, as new content is deposited into our Sufia repository, we'll need to strike a balance between consistency (with our migration data models) and proper representation of the deposited content - these may not always be the same.

Last, we need to stay engaged with the wider community of PCDM users and non-users to understand community norms for data modeling and interoperability. How are others in the community modeling complex objects in their repositories?

Intent Challenges

A second challenge, especially with datasets, is understanding the intentions of the depositor when it comes to the structure of their data. How do the depositors want the data structured, but also, how do users want to discover data objects? Does the depositor or user want files to be zipped together or represented as individual files? How do we balance the depositor's preferences vs. how we want to represent it? For example, what do we do with data visualizations that live somewhere else? Do we preserve them in our repository? Do we just link to them. If so, that of course creates persistency questions?

Intent Solutions

As we mentioned earlier, we've decided to strike a balance and respect the depositor's needs as much as possible while retaining as much consistency as we can. We'll want to be transparent about expectations and procedures and try to be consistent to the existing data models.

Discovery, we know, is going to be hard in this space. We need to consider at what level of granularity users want to discover content and how best to help them navigate content and its relationships with other content. Come to the conference next year to find out how we've approached discovery in this space. We haven't gotten there yet.

Migration Challenges

Of course, we are facing a number of challenges in our migration to hydra/sufia. What can we automate and what can't we? How does this impact the end result? 75% of content is single files--that is easy. Which of the remaining 25% can we automate based on our models and which do we need to handle one by one?

Migration Solutions

We have decided to retain all existing metadata, including structural dspace information such as collections and communities in which items currently reside. We have decided not to replicate all of those collections and communities in hydra/sufia.

We have decided to hand migrate all of our datasets and multi-file dissertations with supplementary files. We are also in the process of identifying things that are related based on similar titles. We call these faux relationships. We will decide how much work we do to create

more formal relationships for these kinds of materials. Metadata can solve this problem.

Again, we will capture community & collection structure in item-level metadata to inform creation of PCDM collections (or not). We have manually audited the 400 collections in our repository to create a crosswalk of community/collection structure to item level metadata for migration.

After migration, we will reevaluate data models for new content to allow flexibility but pay attention to consistency: e.g., multi-part dissertation.

Next Steps

For next steps, we're conducting a repository-wide analysis to find logical object types that require specific models. We are setting local model consistency and identifying relatedness of existing content. We are also Identifying items that are too complex to automate (~10%?).

We're conducting a migration automation pilot with static collections that we are not adding content to. We'll migrate majority of contents by data models pertaining to specific item types; i.e. single file items, multi-file items and then either hand migrate remaining content OR modify automation to meet complexity. We will provide specialized service for such critical collections as ETDs. Will be one of the last collections we migrate, probably along with research articles, because those two collections are constantly being added to. Need to ensure a smooth transition with minimum downtime. So we'll continue to use DSpace for electronic theses and dissertations and faculty articles until our Sufia implementation is thoroughly tested.