



PennState

ScholarSphere Migration to PCDM

We are moving our data. You can too!

Carolyn Cole



PennState

ScholarSphere Migration to PCDM

Data Model Migration

- ❖ Simple overview of Sufia 6 & Sufia PCDM models
- ❖ Model Migration Options
- ❖ Model Migration Decision

Data Migration Tools

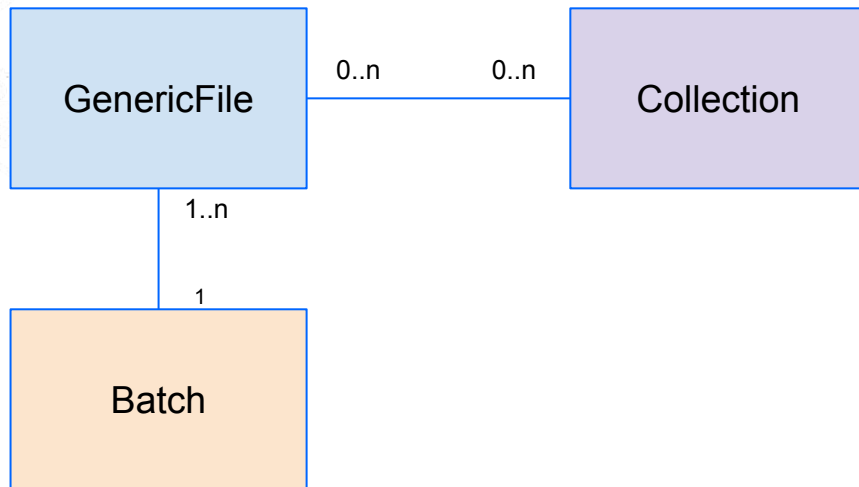
- ❖ Tool Design
- ❖ Example of Extension



Photo By Nasa on the Commons



Sufia 6 Model

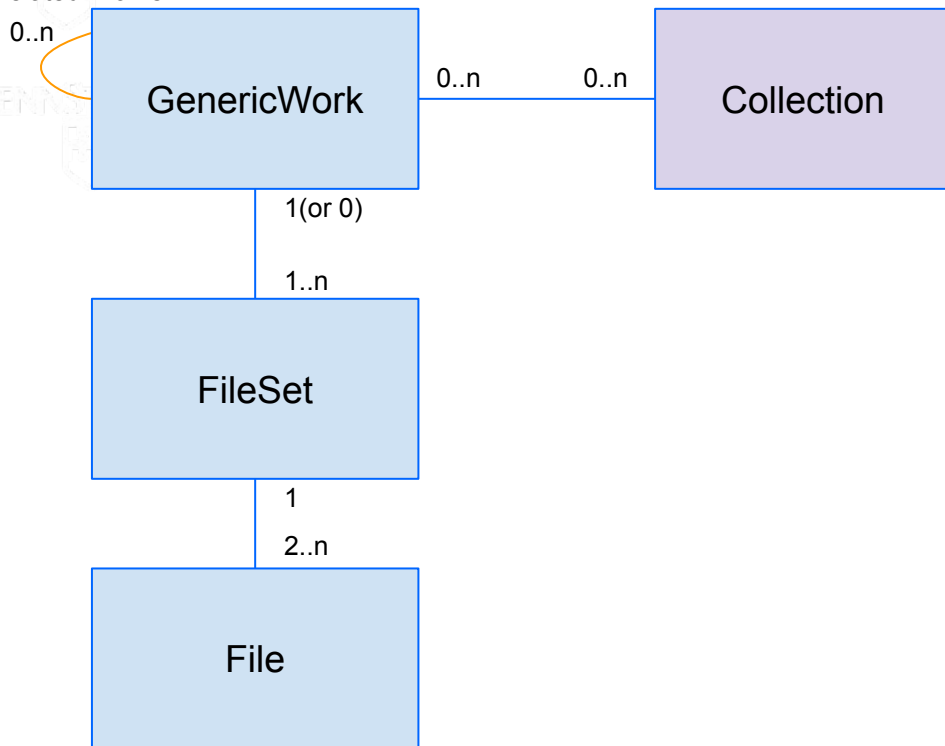


- A very simple file centric model
- GenericFile is the main object in the system
 - Contains data streams with metadata, content, and derivatives
- Batch contains 1 or more GenericFiles uploaded at the same time
 - A way to apply similar metadata to multiple files
- Collection is a user created object that contains zero or more related GenericFiles
 - Has a title and a description



Sufia PCDM Model

Related Works



- A fairly simple Work centric model
- GenericWork contains metadata and one or more uploaded File from the user
 - Descriptive metadata
 - Pointers to contained FileSets
- A FileSet contains the original File and one or more derivatives of the original
 - Original Content
 - Thumbnail
 - Technical metadata
- A file is the binary representation
 - Binary content
- A Collection contains zero or more works



Data Model Migration - Option 1

Batch

=>

GenericWork

GenericFile

=>

FileSet

Collection

=>

Collection

- All files uploaded at one time (the Batch) would be considered a work
- Each GenericFile would be a FileSet contained by the work
- Each Collection would map to a PCDM collection



PennState

Data Model Migration - Option 2

GenericFile

=>

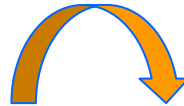
GenericWork

&

FileSet

Batch

=>



Relation Between
GenericWorks

Collection

=>

Collection

- All files uploaded at one time (the Batch) would be related to each other
- Each GenericFile would be a GenericWork and a FileSet
- Each Collection would map to a PCDM Collection



PennState

Data Model Migration - Questions

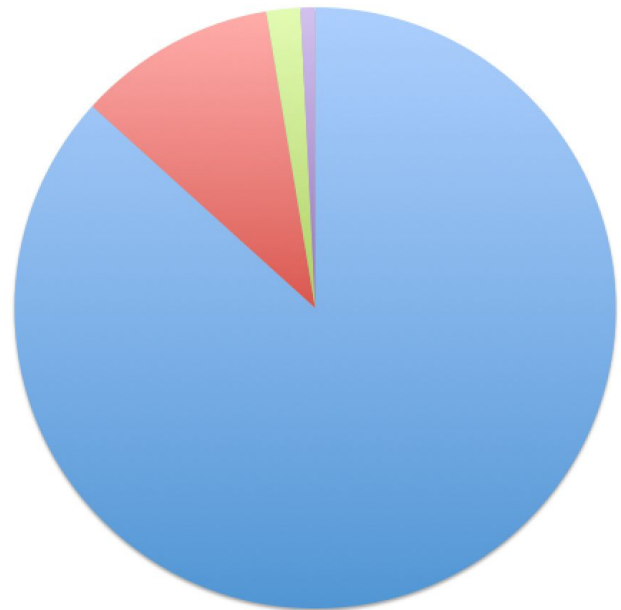
- ❖ How are the ScholarSphere users uploading data into ScholarSphere currently?
- ❖ What do the ScholarSphere users currently see when they upload files?
- ❖ Which option would make the most sense for the majority of ScholarSphere users?



PennState

ScholarSphere User Upload Patterns

- ❖ The majority of ScholarSphere users are not taking advantage of batch upload.
 - 87% of files are being uploaded one at a time
- ❖ Of the ScholarSphere users uploading more than one file at a time most of the batches are small
 - 11% of batches with more than one file have 10 or less files
- ❖ A very small percentage of users are creating large batches
 - Less than 2% of all Batches are more than 10 in size



■ Single File ■ 2-10 Files
■ 11-50 Files ■ 50-101 Files



Current Visualization of Batches

My Files

May 20, 2004

Tutorial Example for R MixTVEM

In order to use the `MaTSEM` function, we have to use the `R` `source` command in the usual way:

This reads the code for the `MatM` function into `B`.

Of course, we also need data. The tutorial example consists of two files. They are simulated (fake) data created to have similar appearance to the Stufman et al (2006, 2007) data as analyzed by Doak, Li, Tan, Stufman, and Shyke (2022), which described the experiences of smokers trying to quit, including their self-rated negative affect and urge to smoke. One of the data files, "MaxTimeExampleObservationsLevel1d", has time-varying covariates for each subject listed in a longitudinal ("tall" or "stacked") format, and most of the subjects have more than one line of data.

When reading it in, make sure that it is in the R working directory, or else add the path to the data files to the file name below.

The other file, "Mia7vcmSampleSubjectLevel.txt," has subject level, baseline data, arranged with one line per subject.

John Dink jd364@msu.edu[Download the full sized image](#)

Actions

[Download](#) | [Analytics](#)Export to: [EndNote](#) | [Zotero](#) | [Mendeley](#)


Collections

This file is not currently in any collections.

A tutorial in how to use MixTVEM.r [Open Access](#)

Open Access

Descriptions

Resource type:		Publisher:
Creator:	Dziak, John Tan, Xianming Li, Runze	Date Created:
Contributor:		Subject:
Keyword:	TVEM, MixTVEM, varying coefficient models, finite mixture models, time varying effect models	Language:
		Identifier:
		Location:
Rights:	Attribution 3.0 United States 	Related URL:

File Details

Depositor:	John Dziak	Characterization:	File format: pdf (Portable Document Format)
Date Uploaded:	2015-05-20T19:28:53+00:00		Mime type: application/pdf
Date Modified:	2015-05-20T19:28:53+00:00		File size: 144815
Audit Status:	passing		Last modified: 2015:05:20 15:29:29-04:00
Related Files:	DemonstrateMixTvm.r TVM_Mix_Normal.sas TVM_Mix_Normal_OldVersion.sas MixTVM.r MixTVM_OldVersion.r Sas-Mixtvm-Tutorial.pdf DemonstrateMixTvm.sas AnalysisCode.zip		File name: R-MixTvm-Tutorial.pdf Original checksum: 30b143b447736d748dca1ef3dcf7df82 Well formed: true Valid: true File title: 口きどへびふぶ口こふひはひどちぬ File title: R-MixTvm-Tutorial File author: ななつで@メ File author: jkd264 Page count: 14



PennState

Search Results - The Public Face

- ❖ In ScholarSphere before the migration 100% of Public GenericFiles are available in search results
 - Each GenericFile is a hit in the result list
- ❖ In Sufia PCDM FileSets are not displayed in search results
 - Each GenericWork is a hit in the results list

The screenshot shows the top navigation bar with links: Home, About, Help, Contact. Below the navigation bar is a banner with a map background and text: "Need help with research data management?" and "Consult the Libraries' guide on RDM ser Questions? Contact the Research Data M Team." Below the banner are two columns: "Featured Works" and "Recent Additions".

Featured Works	Recent Additions
	Development and... Depositor: Nancy Ellen Adams Keywords: evidence-based medicine, assessment c
	Nelson, Signorella, &... Depositor: Margaret Louise Signorella Keywords: language, accent, gender, prejudice
	ACFE 2016 60 in 60... Depositor: Lauren M Reiter Keywords: financial education

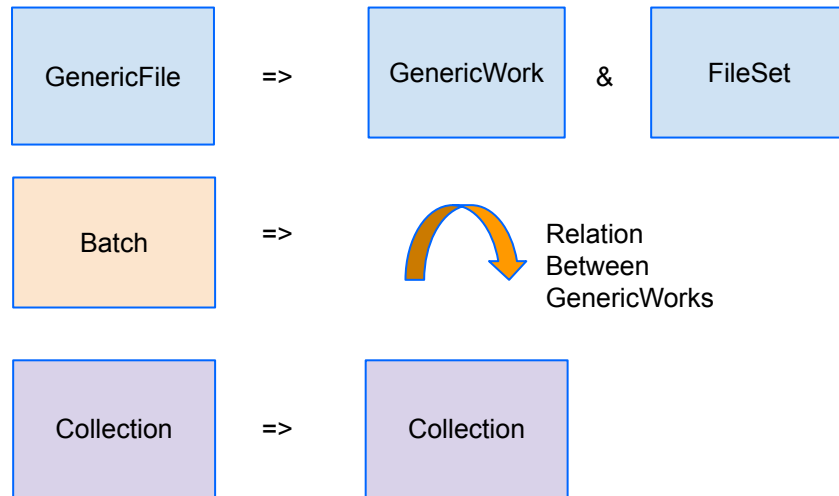


PennState

Data Model Migration - Conclusion

For ScholarSphere we chose Option 2

- ❖ Maintains the relation between search hits
- ❖ Maintains the current visibility of the batch in the UI
- ❖ For 87% of our users option 2 fits their mental model
- ❖ Plan to engage the larger batch users to help with the conversion to works if they would prefer option 1



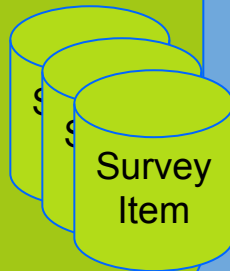


PennState

Migration Design

Sufia 6.7

Survey



Migrate

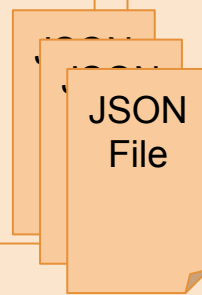
Export



Import

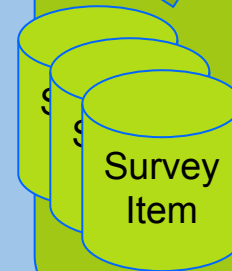


JSON
File



Sufia PCDM

Validate





PennState

Migration Tools

❖ Survey

- List all objects in the repository including the type

❖ Export

- Export all object metadata in the repository with pointers to the content in json files

❖ Import

- Import JSON files mapping objects and metadata as you go

❖ Verify

- Survey the repository and be certain every item was migrated to the correct object type



Photo by Patrick Gray
CC BY-SA 2.0



PennState

Sufia Migration Executables

Sufia 6.7 - sufia_survey & sufia_export

Sufia PCDM - sufia_import & sufia_verify (future)

- ❖ --help Displays all options available to the exe
- ❖ Have built in extension points
- ❖ Have debug options for processing a portion of your repository



Sufia 6 Survey

Survey

sufia_survey.exe

```
--models= (default [Collection, GenericFile] )  
--limit=# (default all)  
--id=[id1, id2]
```

Survey::FedoraIDService

```
call (limit)  
model_registry  
  { Collection.class, GenericFile.class }
```

Survey::Service

```
call ([ids])
```

Survey::Item < ActiveRecord

- ID
- Type
- Title
- Status



Sufia 6 Export

Export

sufia_export.exe

```
--models= (default [Collection, GenericFile] )  
--limit=# (default all)  
--id=[id1, id2]
```

Export::Actor

```
call (model_types, opts)  
register_converter(model_class, converter)  
Registry { Collection.class: CollectionConverter,  
            GenericFile.class: GenericFileConverter}
```

VersionConverter(obj)

to_json ()

GenericFileConverter(obj)

to_json ()

CollectionConverter(obj)

to_json ()

PermissionsConverter(obj)

to_json ()



PennState

Sufia 6 Converter

Works off the Basic
Ruby to_json where
every instance
attribute becomes
an item in the json
output

```
module Sufia
  module Export
    class CollectionConverter < Converter
      def initialize(collection)
        @id = collection.id
        @title = collection.title
        @description = collection.description
        @creator = collection.creator.map { |c| c }
        @members = collection.members.map(&:id)
        @permissions = permissions(collection)
      end
    end
  end
end
```



PennState

Sufia 6 Collection Json

to_json is called recursively on each object so collection includes permission json information

```
{ "id": "wm117p010",  
  "title": "Collection of Wonder",  
  "description": "Wonderful things that we all love",  
  "creator": [ "Cole, Carolyn"],  
  "members": [ "6t053f96k", "9880vr00j"],  
  "permissions": [  
    { "id": "e7cd2089-5112-4a4e-8fd7-9412d975f1fa",  
      "agent": "http://projecthydra.org/ns/auth/person#cam156",  
      "mode": "http://www.w3.org/ns/auth/acl#Write",  
      "access_to": "wm117p010" },  
    { "id": "2f8ddb5e-e92b-4e45-839a-8127fe0a76d0",  
      "agent": "http://projecthydra.org/ns/auth/group#public",  
      "mode": "http://www.w3.org/ns/auth/acl#Read",  
      "access_to": "wm117p010"  
    }  
  ]  
}
```



PennState

Sufia 7 Import & Validate

Vaporware Alert!

Continue at your own risk!

- ❖ Design Exists
- ❖ Tickets Exist



Photo by serragaucha
CC BY-SA 2.0

Sufia PCDM Import

Import

sufia_import.exe

```
--json_directory= directory where files were exported  
--json_locations= [GenericFile:gf_json,Collection:col_json]  
--models= [GenericFile:GenericFileTranslator,Collection:CollectionTranslator]
```

Import::Actor

```
call (model_types, opts)  
register_converter(model_class, convertor)  
Registry { Collection.class: CollectionTranslator,  
            GenericFile.class: GenericFileTranslator}
```

GenericFileTranslator(path)

import ()

CollectionTranslator(path)

import ()

WorkBuilder()

build (metadata)

FileSetBuilder()

build (metadata)

CollectionBuilder()

build (metadata)



Sufia PCDM Validate

Validate

`sufia_validate.exe`

`--model=[Collection:Collection,GenericFile:GenericWork]`

Validate::Service

`call ()`

`mapping_registry { Collection: Collection.class,
GenericFile: GenericWork.class }`

Survey::Item < ActiveRecord

- ID
- Type
- Title
- Status



PennState

Need to Extend the Tools?

**Do you have more metadata
than the vanilla Sufia 6?**

Did you add another model?

**Do you disagree with our
design choices?**

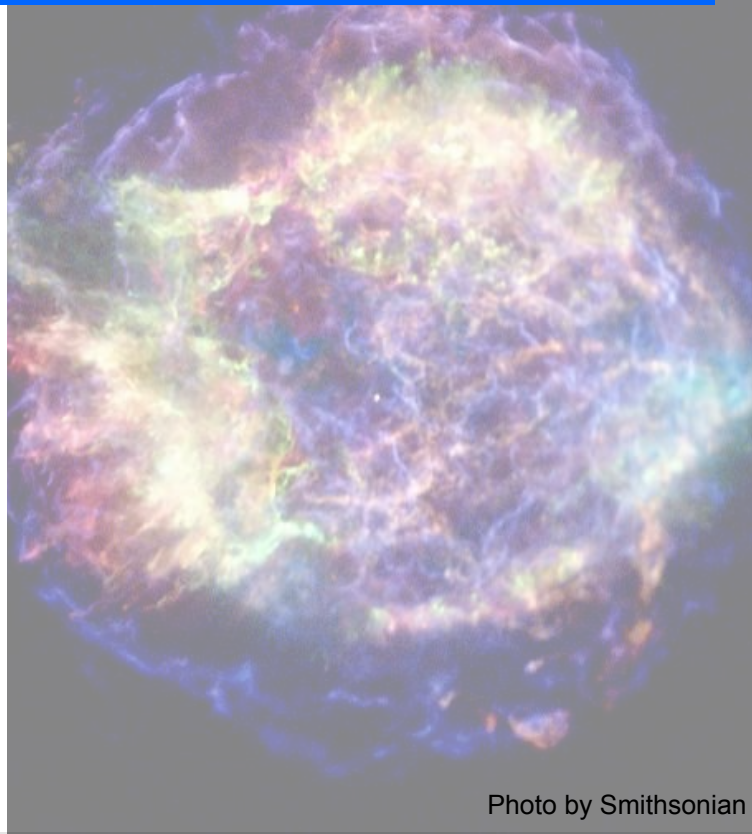


Photo by Smithsonian



PennState

Extending the Survey

1. Still migrating only GenericFiles & Collections?

- a. No Change is needed

2. Have your own model?

- a. Just add the model to the survey

```
> bundle exec sufia_survey --models GenericFile,Collection,MyModel
```



1. Create a new Converter class

```
class MyGenericFileConverter <
  Sufia::Export::GenericFileConverter

  def initialize(generic_file)
    super
    @my_attribute = generic_file.my_attribute
    ...
  end
end
```




2. Pass the convertor to the exporter

```
> bundle exec sufia_export --models  
GenericFile=Mine::MyGenericFileConverter,Collection
```



PennState

Extending the Import

Vaporware Alert!

1. Create a new Translator class
2. Create a new Builder class
3. Pass those classes to
`sufia_import.exe`



Photo by serragaucha
CC BY-SA 2.0

Links

[Export Code in Sufia 6.7](#)

[Import Migration Tickets in Sufia](#)

[ScholarSphere Demo Export
Branch Using Sufia 6.7](#)



Photo by Damian Gadal CC BY 2.0



PennState

Questions?

**Additional questions?
Contact me!**

cam156@psu.edu

Thanks!



Photo by Peter Trimming CC BY 2.0