ScholarSphere Migration to PCDM

We are moving our data. You can too!

Carolyn Cole

Photo by Mike Lewinski CC BY 2.0

ScholarSphere Migration to PCDM

Data Model Migration

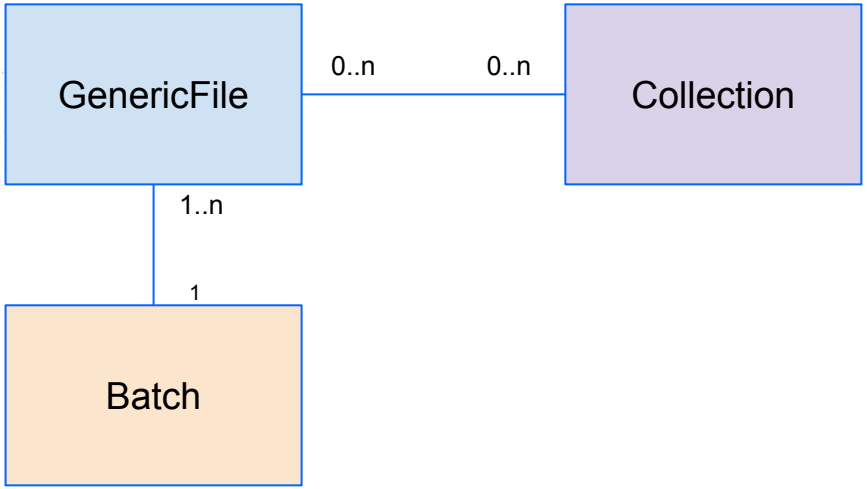
- Simple overview of Sufia 6 & Sufia PCDM models
- Model Migration Options
- Model Migration Decision

Data Migration Tools

- Tool Design
- Example of Extension

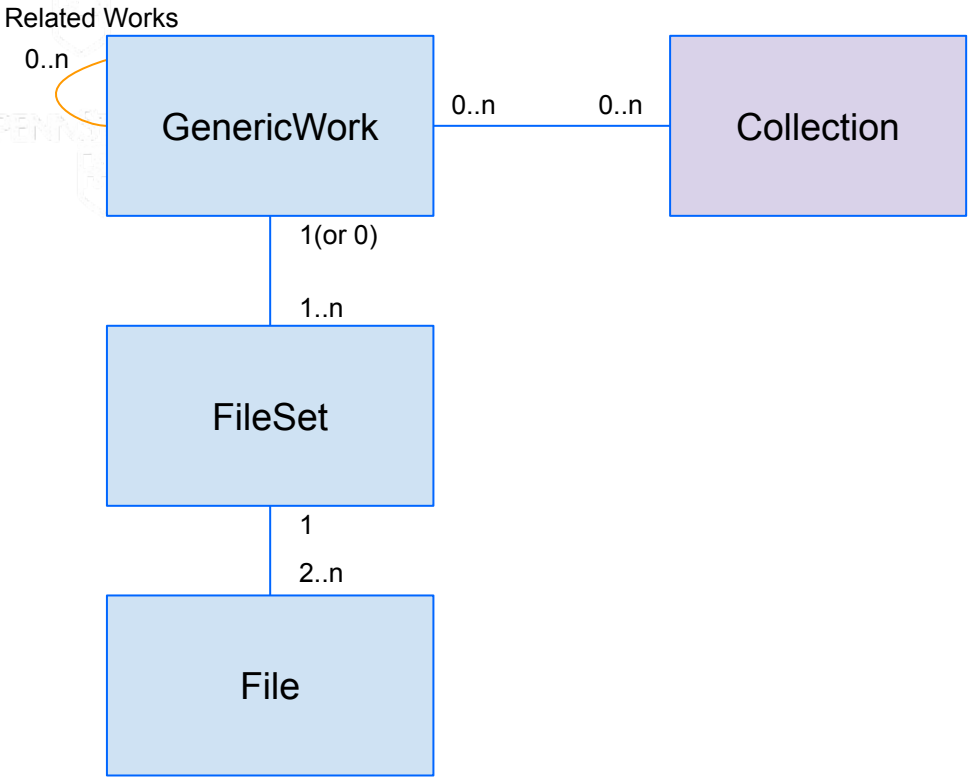
Photo By Nasa on the Commons

Sufia 6 Model



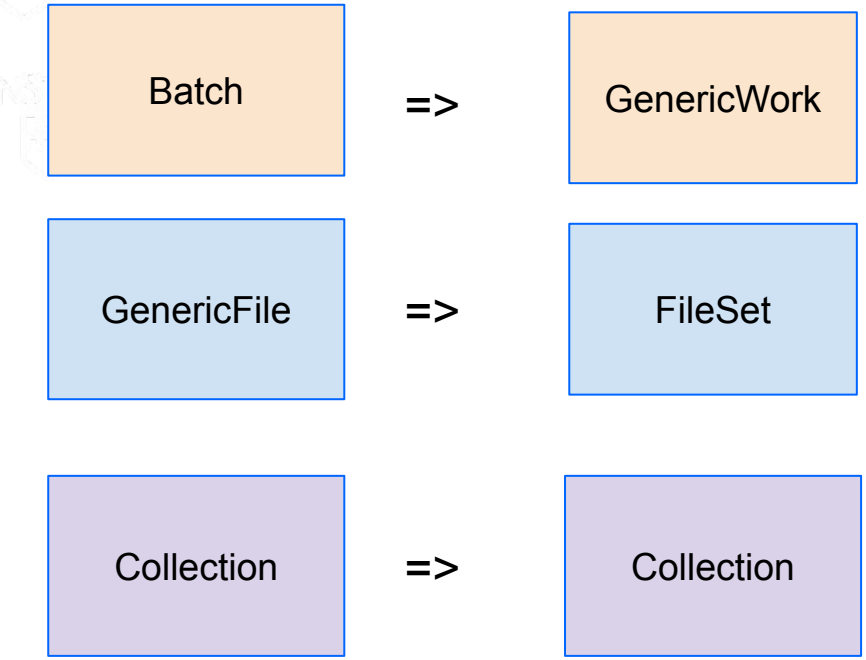
- A very simple file centric model
- GenericFile is the main object in the system
  - Contains data streams with metadata, content, and derivatives
- Batch contains 1 or more GenericFiles uploaded at the same time
  - A way to apply similar metadata to multiple files
- Collection is a user created object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

Sufia PCDM Model



- 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

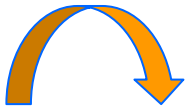


- All files uploaded at one time (the Batch) would be considered a work
- Each GenericFile would be a FileSet contained by the work
- Each Collection would map to a PCDM collection

Data Model Migration - Option 2

GenericFile => GenericWork & FileSet

Batch =>



Relation Between GenericWorks

Collection => Collection

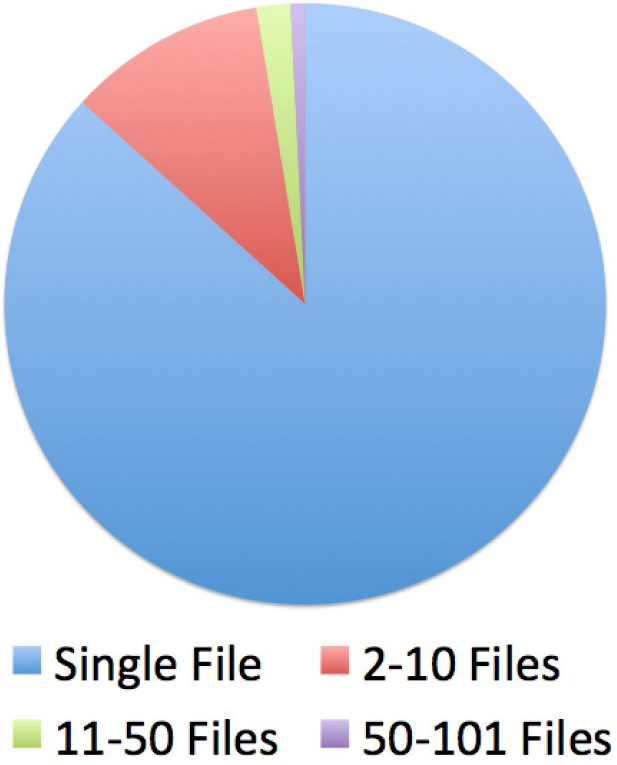
- All files uploaded at one time (the Batch) would be related to each other
- Each GenericFile would be a GenericWork and a FileSet
- Each Collection would map to a PCDM Collection

Data Model Migration - Questions

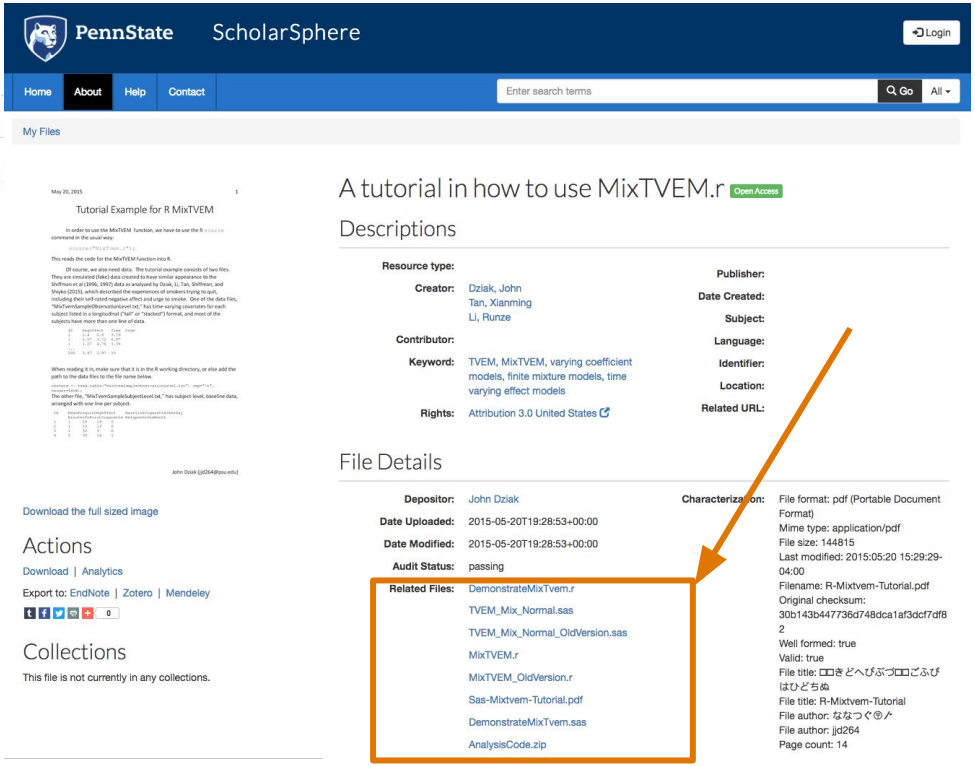
- How are the ScholarSphere users uploading data into ScholarSphere currently?
- What do the ScholarSphere users currently see when they upload files?
- Which option would make the most sense for the majority of ScholarSphere users?

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



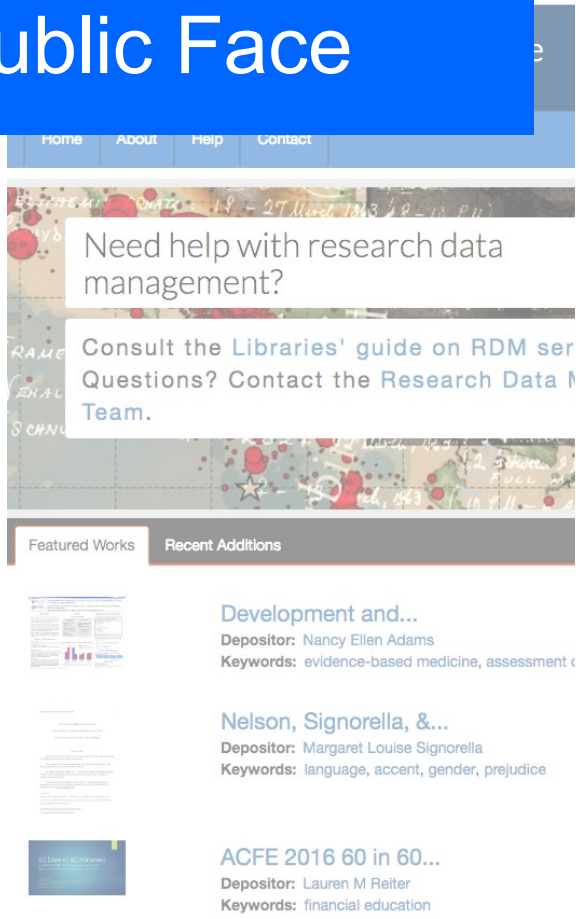
Current Visualization of Batches



After Initial upload batches are only displayed as related files.

Search Results - The Public Face

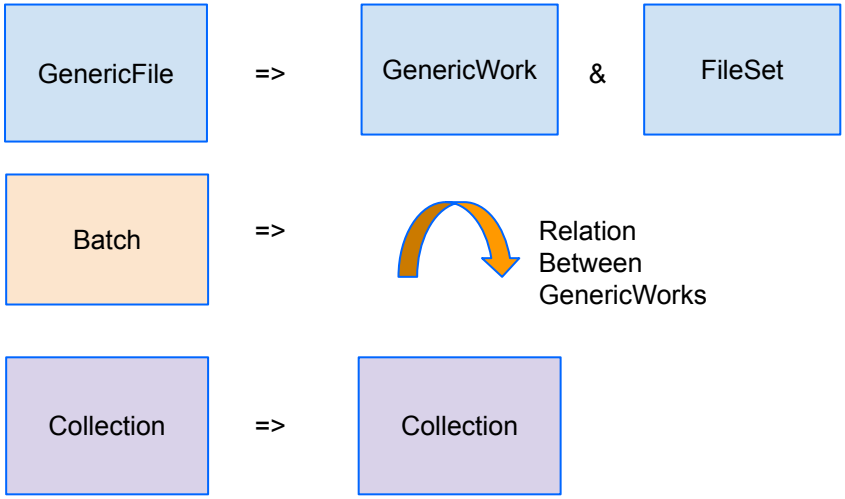
- In ScholarSphere before the migration 100% of Public GenericFiles are available in search results
  - Each GenericFile is a hit in the result list
- In Sufia PCDM FileSets are not displayed in search results
  - Each GenericWork is a hit in the results list



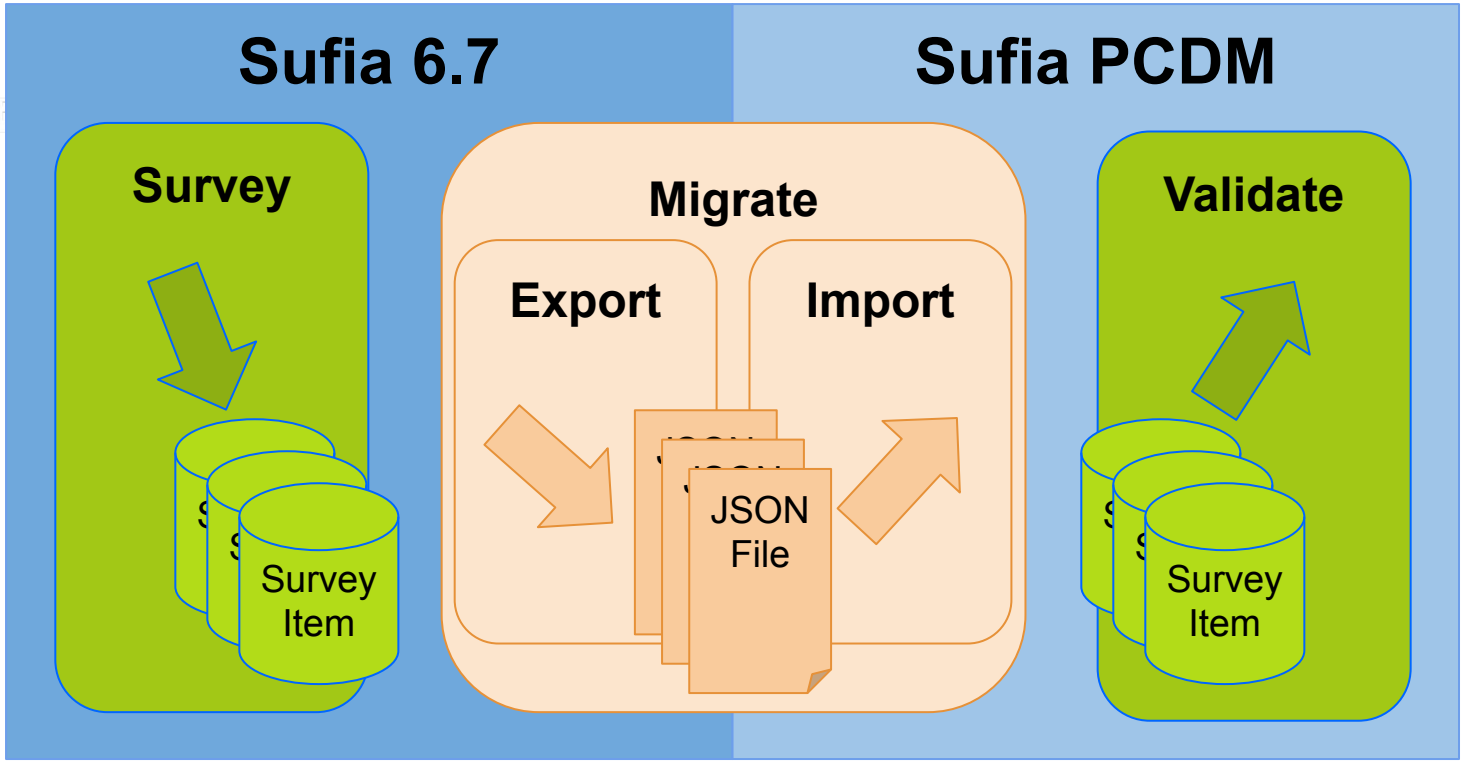
Data Model Migration - Conclusion

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



Migration Design



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

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_convertor(model_class, convertor)
Registry { Collection.class: CollectionConvertor,
GenericFile.class: GenericFileConvertor}

GenericFileConvertor(obj)

to_json ()

CollectionConvertor(obj)

to_json ()

VersionConvertor(obj)

to_json ()

PermissionsConvertor(obj)

to_json ()

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
```

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"
    }
  ]
}
```

Sufia 7 Import & Validate

Vaporware Alert!

Continue at your own risk!

- Design Exists
- Tickets Exist

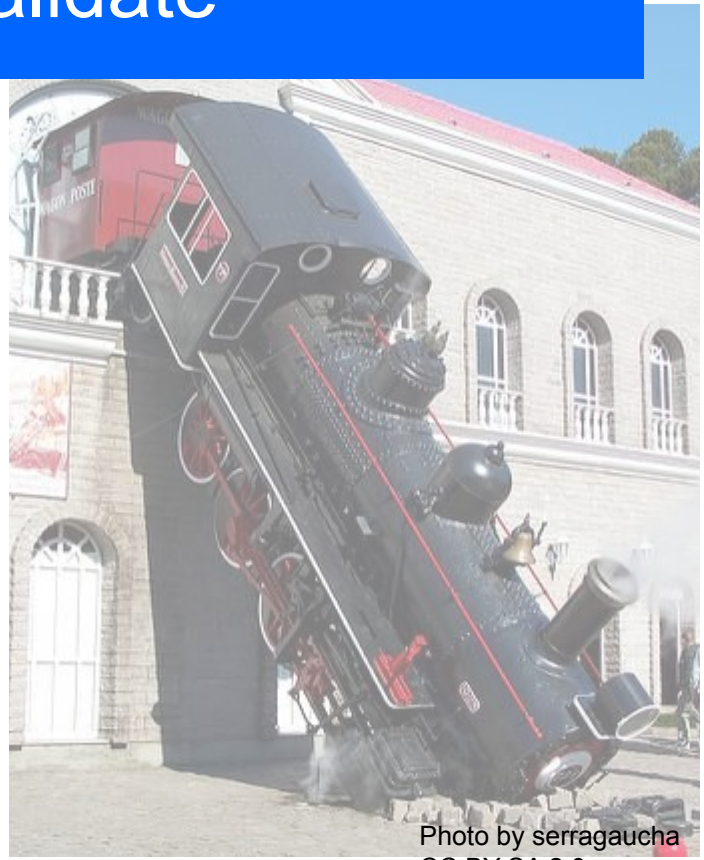


Photo by serragaucha
CC BY-SA 2.0

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_convertor(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

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?

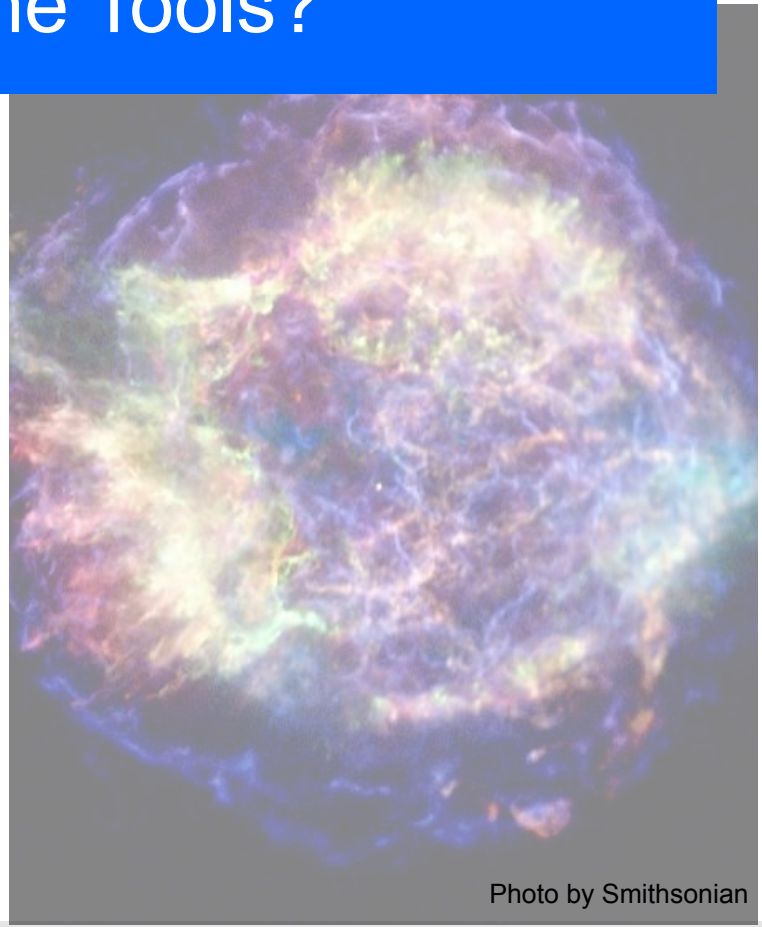


Photo by Smithsonian

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
```

Extending the Export

1. Create a new Convertor class

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

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

Extending the Export (cont.)

2. Pass the convertor to the exporter

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

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

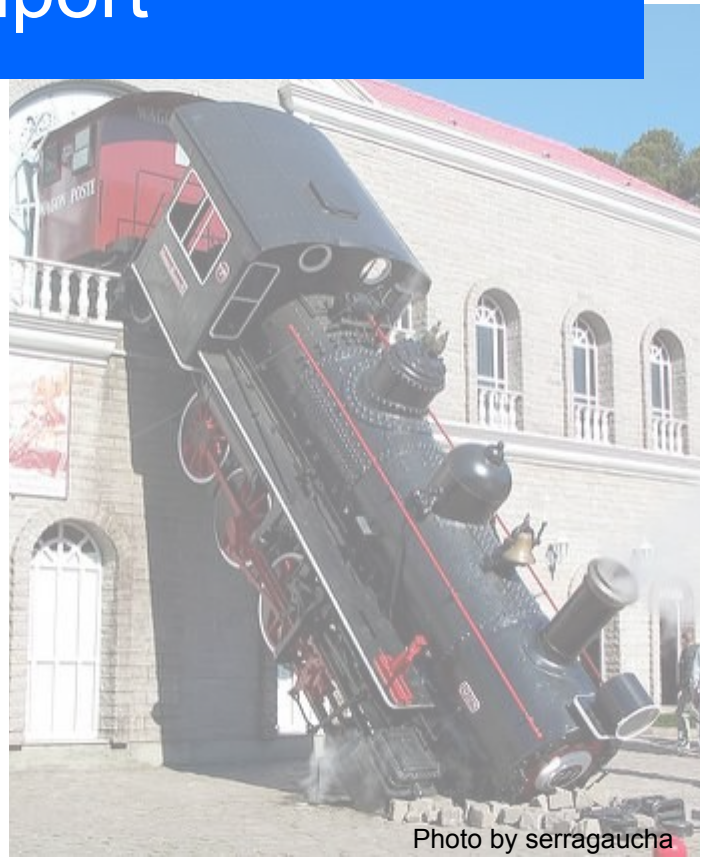


Photo by serragaucha
CC BY-SA 2.0

Links

Export Code in Sufia 6.7

Import Migration Tickets in Sufia

ScholarSphere Demo Export Branch Using Sufia 6.7

Questions?

Additional questions?
Contact me!

cam156@psu.edu

Thanks!

Photo by Peter Trimming CC BY 2.0