Migrating Metadata Digital Collections in RDF, PCDM & Fedora 4

Christina Harlow / cmh329@cornell.edu / @cm_harlow

http://github.com/cmh2166/elag16metadata

Slides, Data, Scripts

github.com/cmh2166/elag16metadata

Exercises, Drawings, Notes

bit.ly/elag16metadata

Lots of Big Important Words!

RDF!

Fedora 4!

PCDM!

Assessment!

Operationalize!

Metadata!

Whoa!

Workshop Goals

- Discuss Real World Use Cases & Work
- Engage in these Topics by Diving In
- Build Out Models, Metadata, Tools, Other Needs
- Get More People (YOU) into Community Discussions

My Caveats

I. AM. NOT. A. DEVELOPER and...

THIS. IS. NOT. A. FEDORA. 4. WORKSHOP.

I am a metadata wrangler, however, and...

& this is a 'what I wish I had known before a Fedora 4 migration' workshop.

Agenda

Day 1: 14:00-15:30

14:00-14:45 Existing Metadata Assessment

14:45-15:30 PCDM & Data Modeling

Day 2: 11:00-13:00

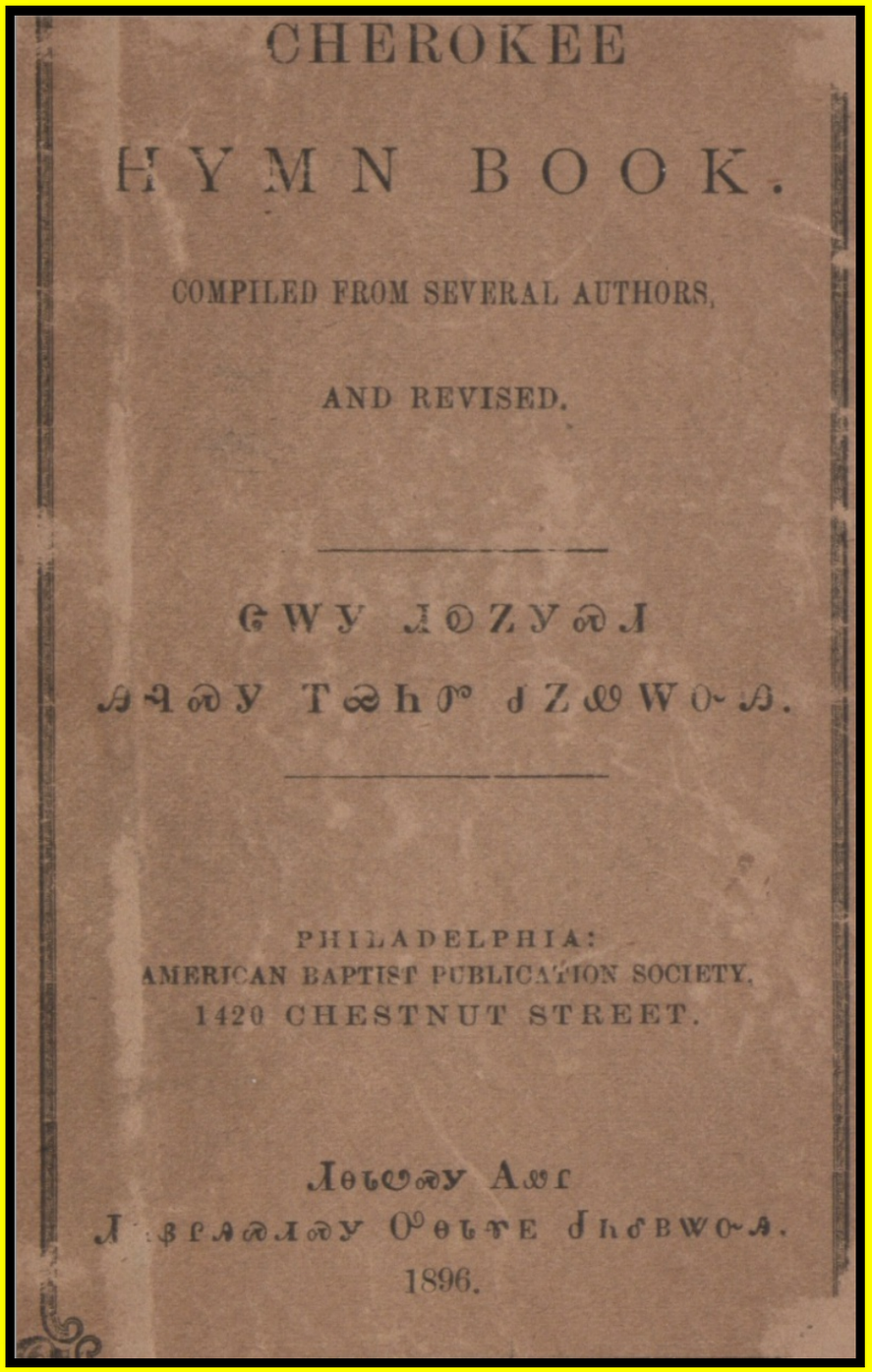
11:00-12:00 Modeling => Fedora 4

12:00-13:00 RDF Assessment & Enhancement

Groups

For the exercises, we'll break out into 4 groups with distinct archival and special collections object types...

Group 1: Digitized Books



CHEROKEE

HYMN BOOK.

COMPILED FROM SEVERAL AUTHORS,

AND REVISED.

ᏣᎳᎩ ᎯᎪᏃᏅᎯᏓ

ᎠᏎᏍᎩ ᏔᏍᏂᏲ ᏧᏃᏪᎳᏅᎯ.

PHILADELPHIA:
AMERICAN BAPTIST PUBLICATION SOCIETY,
1420 CHESTNUT STREET.

ᎯᏍᏪᏍᏅᎩ ᎠᏴᏟ
ᎯᏐᎵᏁᎯᏍᏅᎩ ᎤᏍᏆᏂᎪᎯᏳᎲᏍᎬ ᏧᏂᎸᏫᏍᏓᏁᎰᏒᎯ.
1896.

Group 2: Digitized Photographs

Group 3: Digitized Journals

CORNELL
Reading-Course for Farmers' Wives
PUBLISHED BY THE COLLEGE OF AGRICULTURE OF CORNELL UNIVERSITY, FROM NOVEMBER TO MARCH, AND ENTERED AT ITHACA AS SECOND-CLASS MATTER UNDER ACT OF CONGRESS OF JULY 16, 1894.
MARTHA VAN RENSSELAER, *Supervisor*.

SERIES I. FARMHOUSE AND GARDEN. | ITHACA, N. Y., FEBRUARY, 1903. | No. 4. THE KITCHEN-GARDEN.

THE KITCHEN-GARDEN.

BY JOHN CRAIG.

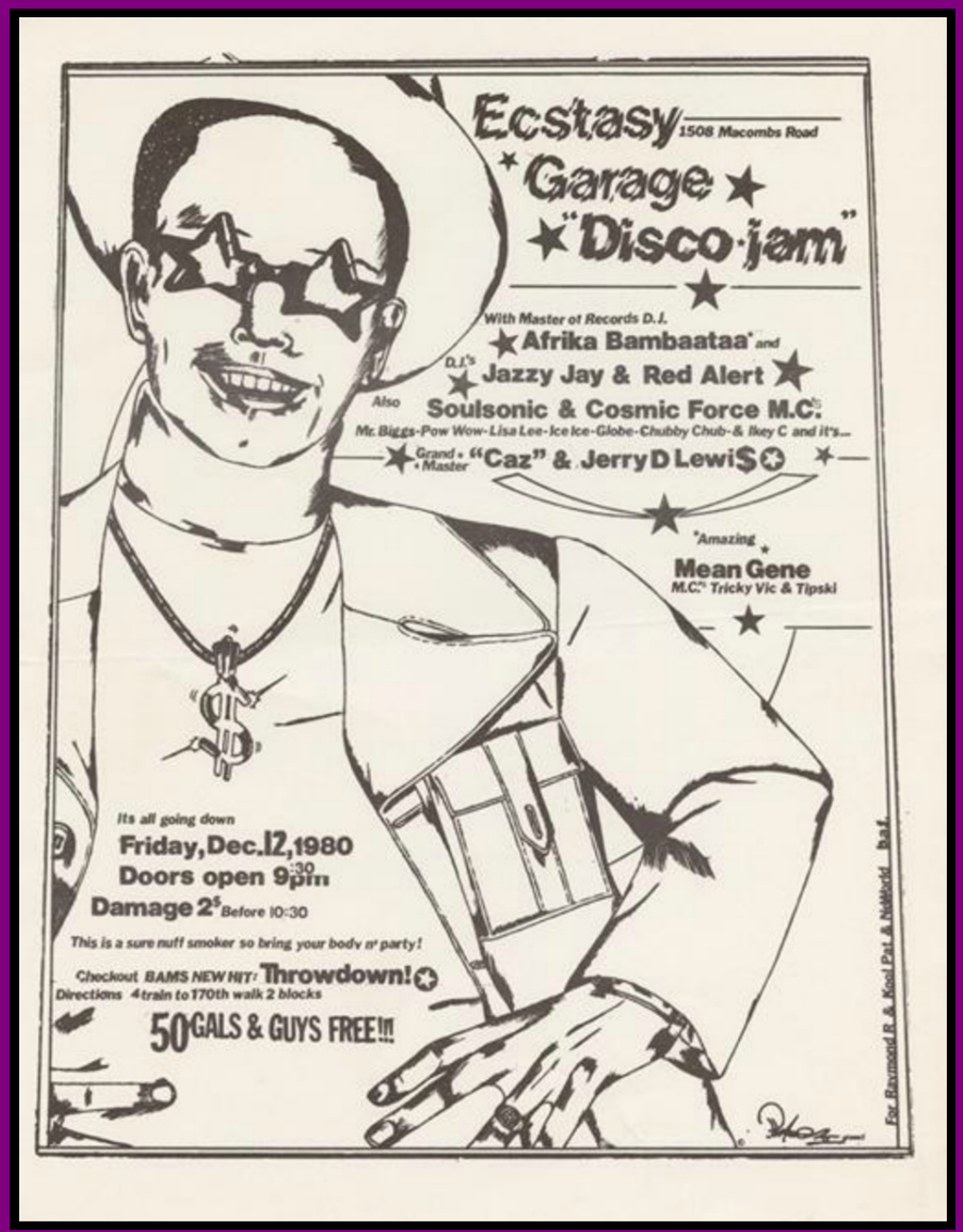
A STATEMENT so common as to have almost acquired the standing of an axiom runs like this: "Every properly appointed kitchen should have as an adjunct a well-planted and thoroughly cared for fruit and vegetable garden." The writer or speaker who promulgates this venerable and apparently unimpeachable platitude rarely thinks it necessary to defend the position, but immediately presses on to tell us what we should plant.

FIG. 29. *The farm strawberry-bed.*

1. The kitchen-garden is an adjunct of the kitchen.

The kitchen-garden belongs to the domain of the housewife. Why should she plant it at all? Surely she has work enough within doors.

Group 4: Digitized Flyers



Assessing Existing Metadata

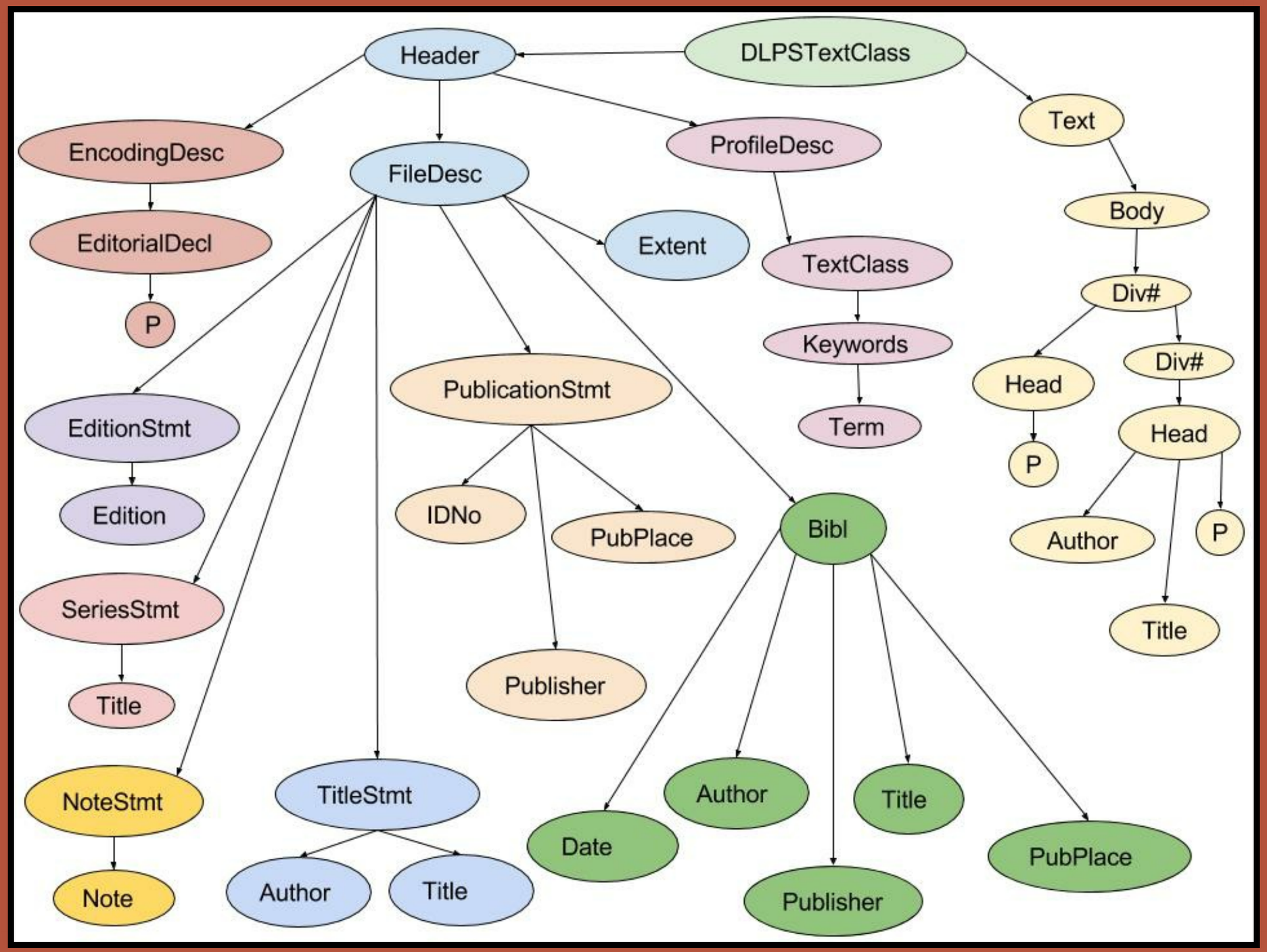
analyze what you've got as part of planning for what you want

DLXS

- Digital Library eXtension Service
- XML-focused digital library toolset/platform
- Meant to be extensible...
- ... but means modeling can be convoluted
- No longer maintained/supported

http://www.dlxs.org/

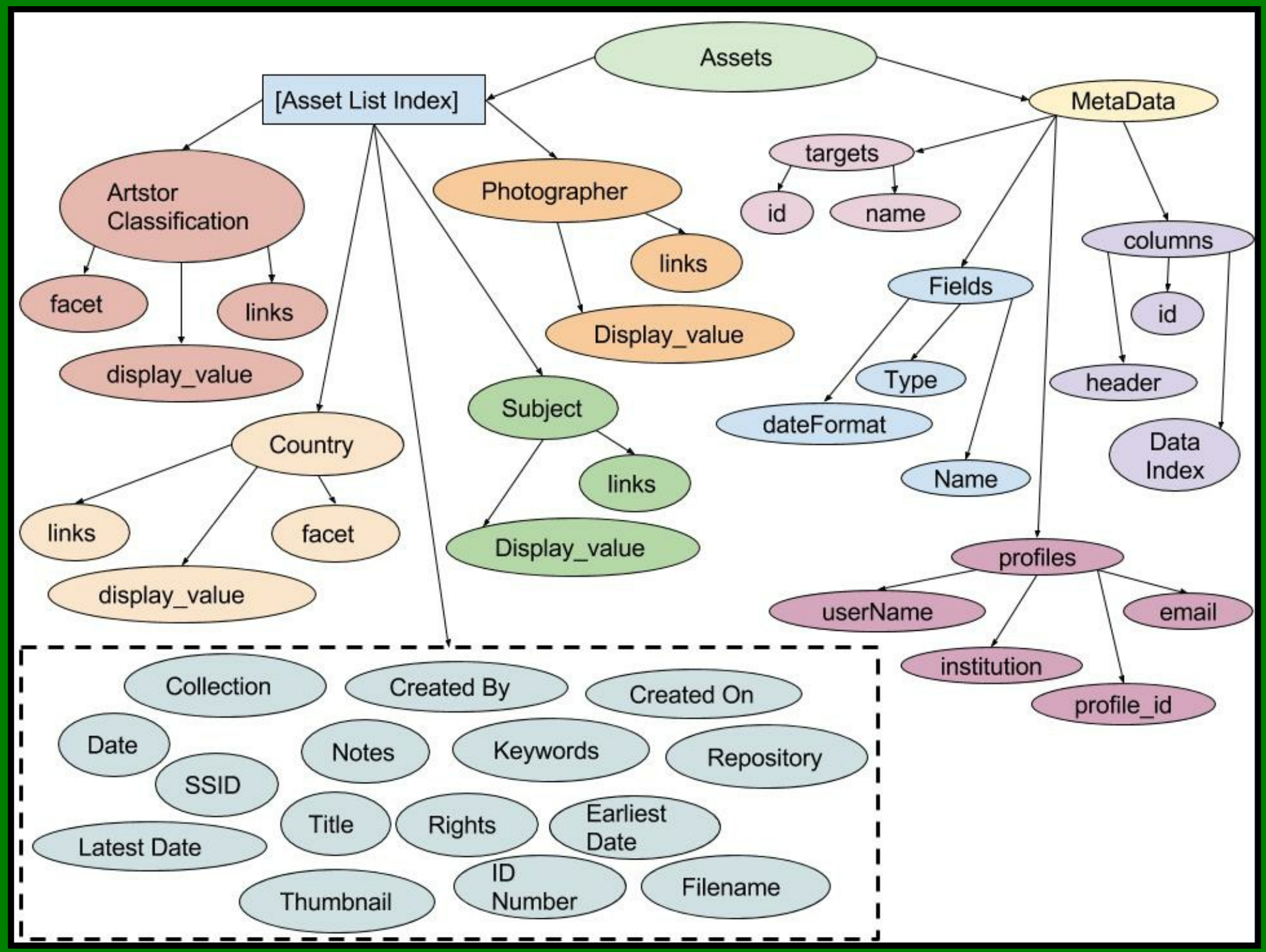
DLXS Cornell "Model"



SharedShelf

Artstor's digital assets platform, with API we pipe to Solr.

Cornell's SharedShelf "Model"



Metadata Assessment

Hop topic right now:

- Europeana Metadata QA Efforts/Work
- Various DPLA Hub Metadata QA Efforts
- DLF AIG Metadata Assessment Working Group
- Other scripts, projects, efforts...?

Assessment Scripts

Metadata Assessment with Python & Unix Pipes, Filters:

- find field usage frequencies
- get unique lists of field values
- generate other field usage stats

Built off of Mark Phillips' Metadata Breakers.

Metadata Assess Python Scripts Preparation

1. Open your shell
2. Change (cd) into where you downloaded the GitHub Repository/Materials
3. run:

```
$ pip install -r scripts/requirements.txt
```

4. let that process finish

Metadata Assess Python Scripts Overview

These scripts (dlxsexport_analysis.py and artstor_analysis.py) work by:

1. Loading metadata dump into memory
2. For each record in metadata:
    - finding record identifier
    - if non-empty, storing field for analysis

Metadata Assess Python Scripts Overview

With analysis for each record (previous slide)...

- **-x**/**-e**: runs XPath/ObjectPath on record & return values
- **-i**: returns record identifier for value returned
- **-p**: returns 'True/None' if XPath or ObjectPath does/does not return value
- No flags, returns field analysis for all records

Running Metadata Assess Python Example

$ python [location of scripts] [location of data dump]

```
$ python scripts/dlxsexport_analysis.py data/hunt_books.xml
```

Running Assessment Scripts: Overview

```
$ python scripts/dlxsexport_analysis.py data/hunt_books.xml

       /record/ENCODINGDESC/EDITORIALDECL/P: 124/124 | 100%
                  /record/FILEDESC/EXTENT: 124/124 | 100%
     /record/FILEDESC/PUBLICATIONSTMT/IDNO: 124/124 | 100%
 /record/FILEDESC/PUBLICATIONSTMT/PUBLISHER: 124/124 | 100%
  /record/FILEDESC/PUBLICATIONSTMT/PUBPLACE: 124/124 | 100%
    /record/FILEDESC/SOURCEDESC/BIBL/AUTHOR: 124/124 | 100%
      /record/FILEDESC/SOURCEDESC/BIBL/DATE: 124/124 | 100%
      /record/FILEDESC/SOURCEDESC/BIBL/NOTE: 124/124 | 100%
 /record/FILEDESC/SOURCEDESC/BIBL/PUBLISHER: 124/124 | 100%
  /record/FILEDESC/SOURCEDESC/BIBL/PUBPLACE: 124/124 | 100%
     /record/FILEDESC/SOURCEDESC/BIBL/TITLE: 124/124 | 100%
          /record/FILEDESC/TITLESTMT/AUTHOR: 124/124 | 100%
           /record/FILEDESC/TITLESTMT/TITLE: 124/124 | 100%
/record/PROFILEDESC/TEXTCLASS/KEYWORDS/TERM: 124/124 | 100%
               /record/TEXT/BODY/DIV1/HEAD: 124/124 | 100%
```

All Values at XPath w/Record ID

```
$ python [location of scripts] [location of data dump] [id flag] [xpath value]
```

```
$ python scripts/dlxsexport_analysis.py data/chla_journals.xml -i
  -x 'FILEDESC/SOURCEDESC/BIBL/AUTHOR'
```

All Values at XPath w/Record ID

```
$ python scripts/dlxsexport_analysis.py data/chla_journals.xml -i
  -x 'FILEDESC/SOURCEDESC/BIBL/AUTHOR'

  5075626_4287_001      Rural Sociological Society.
  5075626_4287_002      Rural Sociological Society.
  5075626_4287_003      Rural Sociological Society.
  5075626_4287_004      Rural Sociological Society.
  5075626_4288_001      Rural Sociological Society.
  5075626_4288_002      Rural Sociological Society.
  5075626_4288_003      Rural Sociological Society.
  5075626_4288_004      Rural Sociological Society.
  5075626_4289_001      Rural Sociological Society.
  5075626_4289_002      Rural Sociological Society.
  5075626_4289_003      Rural Sociological Society.
  5075626_4289_004      Rural Sociological Society.
  5075626_4290_001      Rural Sociological Society.
  5075626_4290_002      Rural Sociological Society.
...
```

If There Is Value for OPath

```
$ python [location of scripts] [location of data dump] [present flag] [opath value]
```

```
$ python scripts/artstor_analysis.py data/hiphop_flyers.json
  -p -e 'Performers'
```

If There Is Value for OPath

```
$ python scripts/artstor_analysis.py data/flyers/hiphop_flyers.json -p -

167_1334196 True
167_1334194 True
167_455428 True
167_1333885 True
167_455357 True
167_1335132 True
167_1334244 True
167_1335130 True
167_1335136 True
167_1335134 True
167_1334327 True
167_1335138 True
167_1334323 True
167_455297 True
167_1334095 True
167_1334116 True
...
```

Use Pipes & Filters to Target Analysis

```
$ python scripts/dlxsexport_analysis.py data/hunt_books.xml
    -x 'FILEDESC/EXTENT' | sort | uniq -c
# All the unique values in 'FILEDESC/EXTENT', organized by count
$ python scripts/dlxsexport_analysis.py data/hunt_books.xml -p
    -x 'FILEDESC/EXTENT' | grep None
# All the record identifiers for the Promoter field is missing
$ python scripts/artstor_analysis.py data/hiphop_flyers.json
  -i -p -e 'Promoter' | grep None
```

These can help you quickly assess a fields value for review, enhancement or mapping.

And Now You…

1. Review your dataset's values;
2. Note objects, fields, values;
3. Confirm preliminary mappings;
4. What fields apply to:
    - the physical resource?
    - the digital resource?
    - the files?
    - other related entities?

PCDM & Data Modeling

Let's start modeling our digital repository objects

RDF (Resource Description Framework)

Briefly...

- Standard model for data exchange on Web
- RDF is made up of triples, i.e.

```
resource_uri predicate_uri object_uri
```

- Can be serialized/stored in number of ways

PCDM

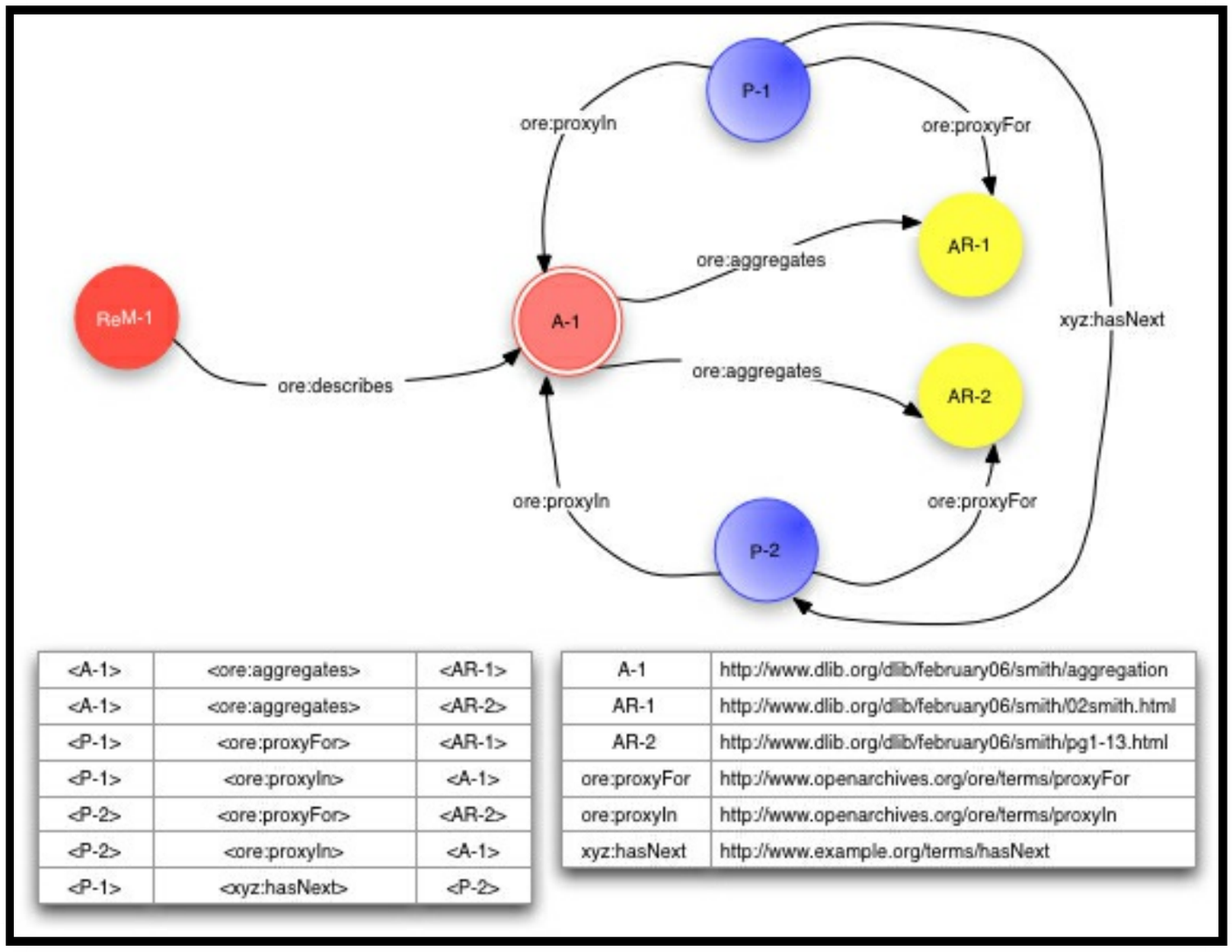
Portland Common Data Model

- Started Officially in 2015
- Community Effort to Make Digital Repository Objects More Interoperable
- Ongoing work & Discussion at:
  - pcdm.org
  - github.com/duraspace/pcdm
  - groups.google.com/forum/#!forum/pcdm

PCDM Continued

- http://pcdm.org/models#
- github.com/duraspace/pcdm/blob/master/models.rdf
- Builds off Object Reuse & Exchange (ORE) Data Model
- Interacts with other specifications (LDP), platforms (Fedora 4), but is meant to be neutral

ORE Abstract Model

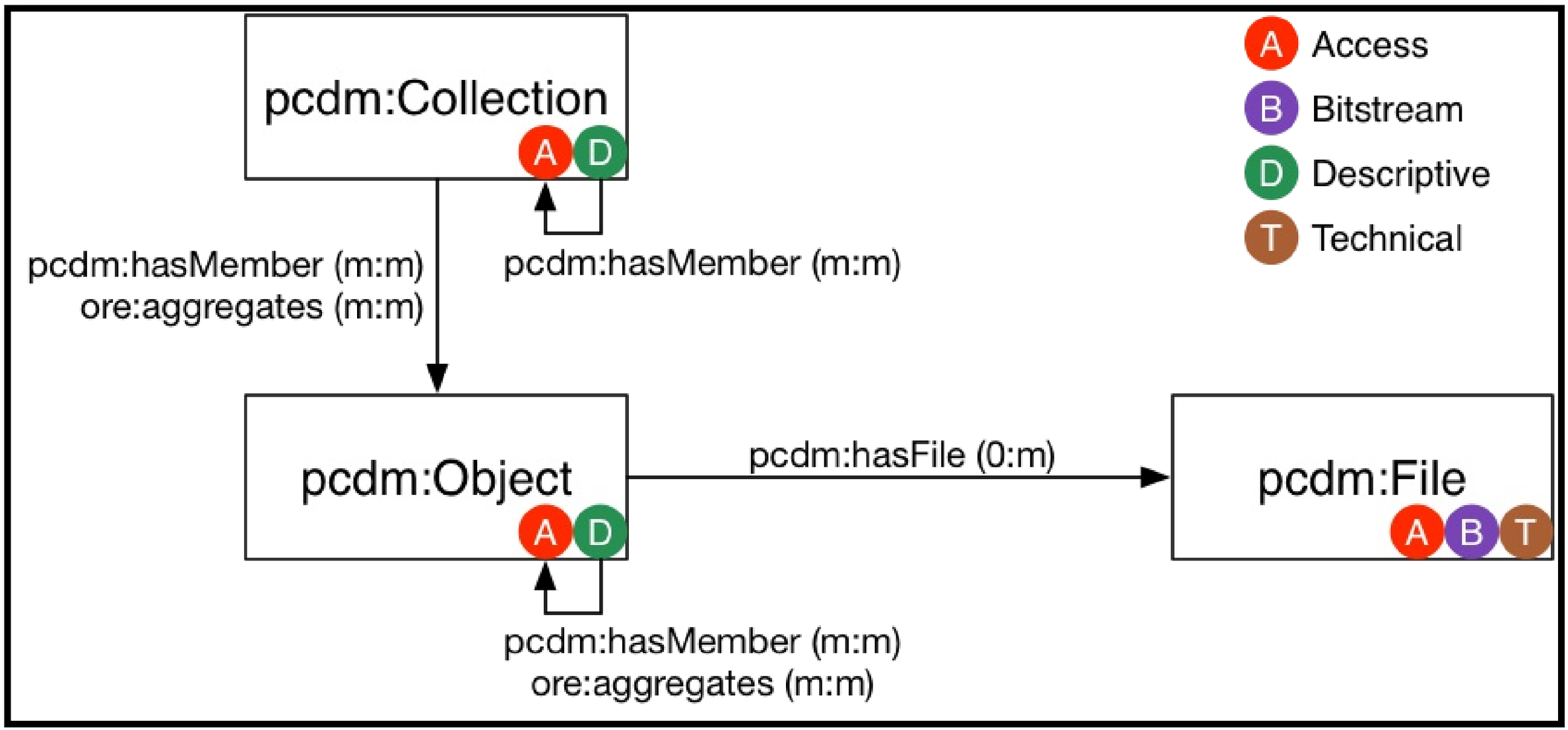


<A-1>	<ore:aggregates>	<AR-1>
<A-1>	<ore:aggregates>	<AR-2>
<P-1>	<ore:proxyFor>	<AR-1>
<P-1>	<ore:proxyIn>	<A-1>
<P-2>	<ore:proxyFor>	<AR-2>
<P-2>	<ore:proxyIn>	<A-1>
<P-1>	<xyz:hasNext>	<P-2>

A-1	http://www.dlib.org/dlib/february06/smith/aggregation
AR-1	http://www.dlib.org/dlib/february06/smith/02smith.html
AR-2	http://www.dlib.org/dlib/february06/smith/pg1-13.html
ore:proxyFor	http://www.openarchives.org/ore/terms/proxyFor
ore:proxyIn	http://www.openarchives.org/ore/terms/proxyIn
xyz:hasNext	http://www.example.org/terms/hasNext

http://www.openarchives.org/ore/1.0/

PCDM Overview



https://github.com/duraspace/pcdm/wiki

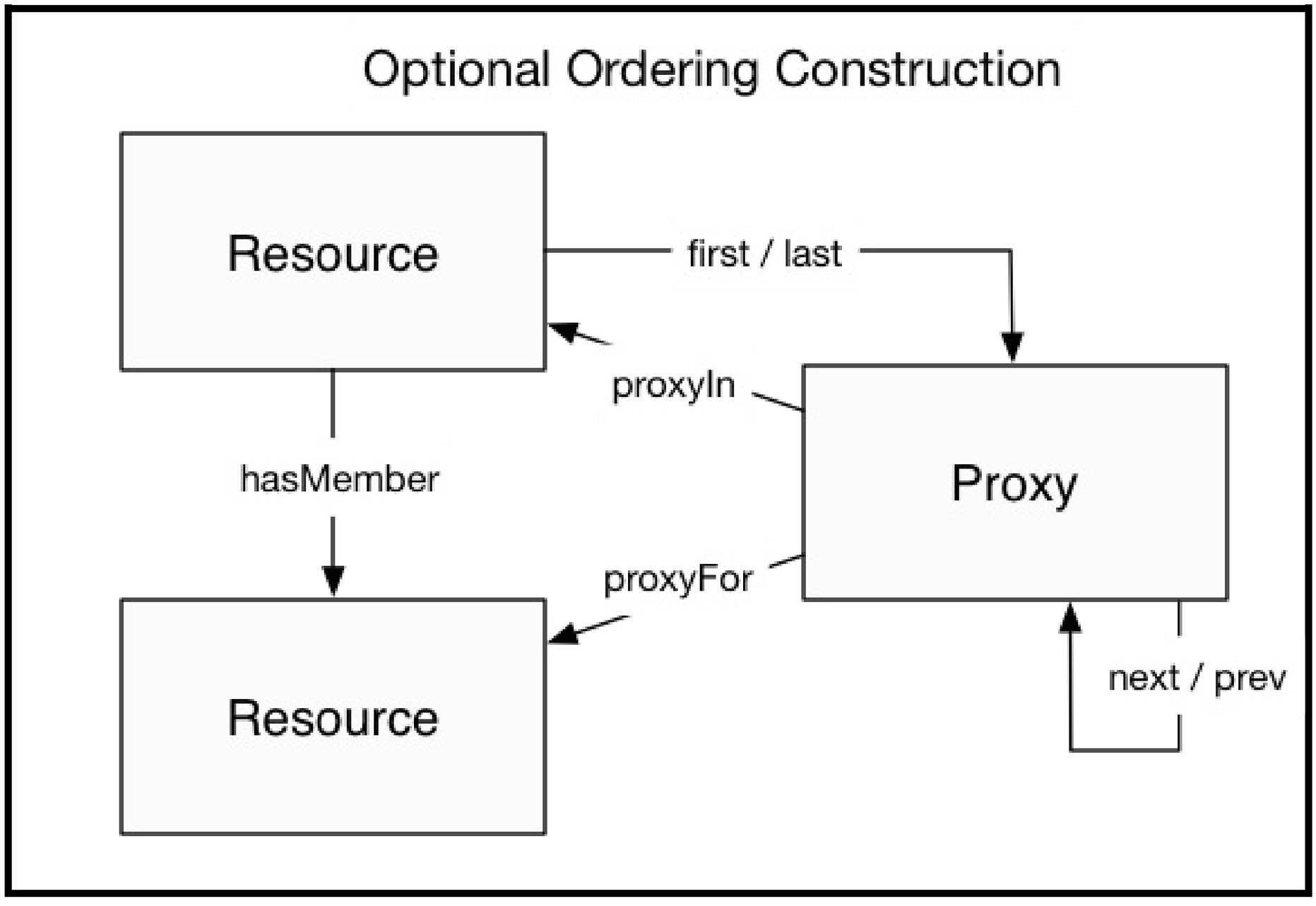
PCDM Classes

- **pcdm:Object**: An Object is an intellectual entity, sometimes called a "work", "digital object", etc...
- **pcdm:Collection**: A Collection is a group of resources...
- **pcdm:File**: A File is a sequence of binary data and is described by some accompanying metadata...
- **pcdm:AlternateOrder**: An AlternateOrder is an alternate ordering of its parent's members. It should only order the parent's members...
- **pcdm:AdministrativeSet**

PCDM Properties

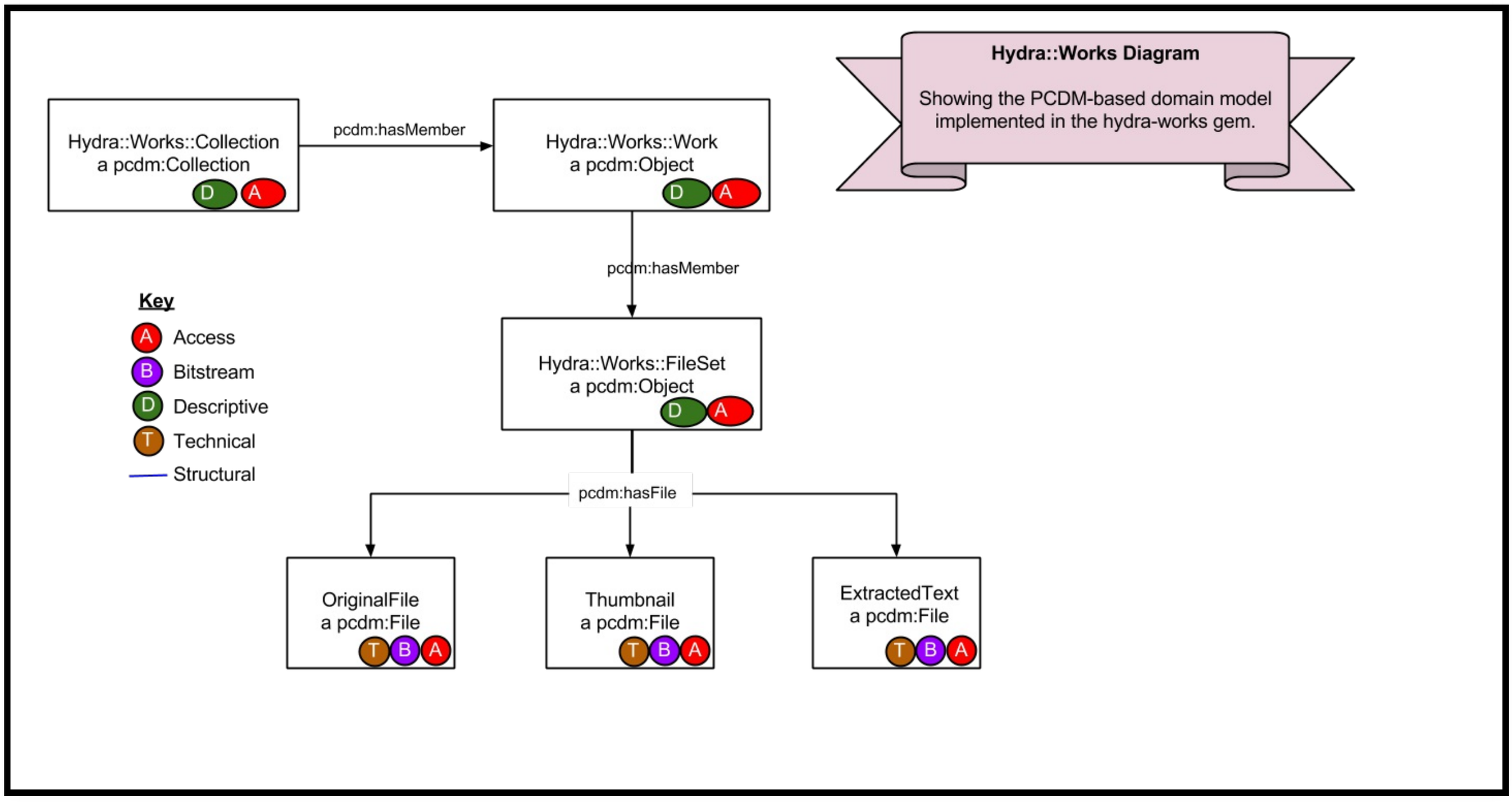
- **pcdm:memberOf**
  Domain: ore:Aggregation | Range: ore:Aggregation
- **pcdm:hasMember**
  Domain: ore:Aggregation | Range: ore:Aggregation
- **pcdm:fileOf**
  Domain: pcdm:File | Range: pcdm:Object
- **pcdm:hasFile**
  Domain: pcdm:Object | Range: pcdm:File
- **pcdm:relatedObjectOf**
  Domain: pcdm:Object | Range: ore:Aggregation
- **pcdm:hasRelatedObject**
  Domain: ore:Aggregation | Range: pcdm:Object

PCDM Ordering



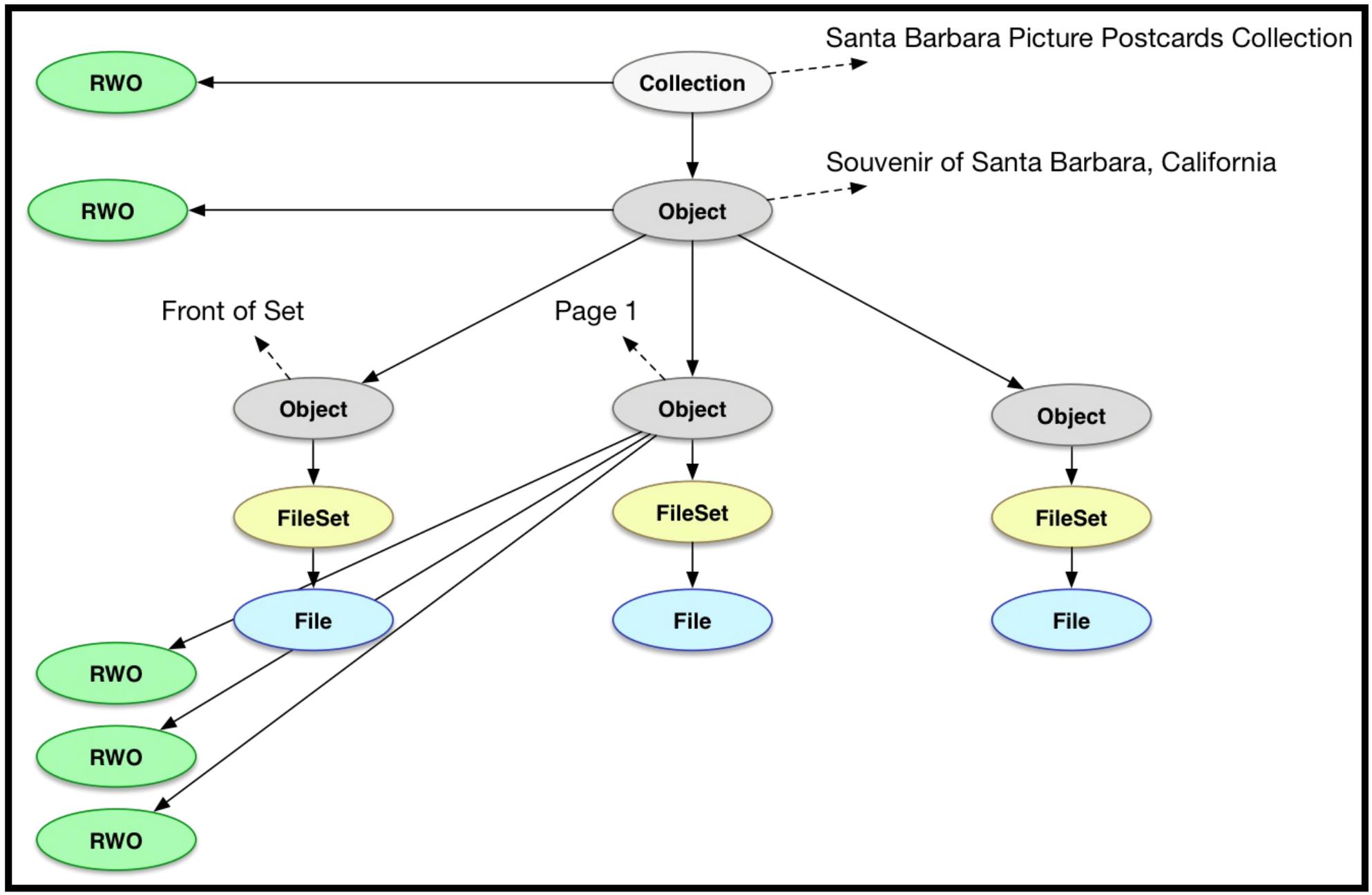
https://github.com/duraspace/pcdm/wiki

PCDM (Works) … TBD?



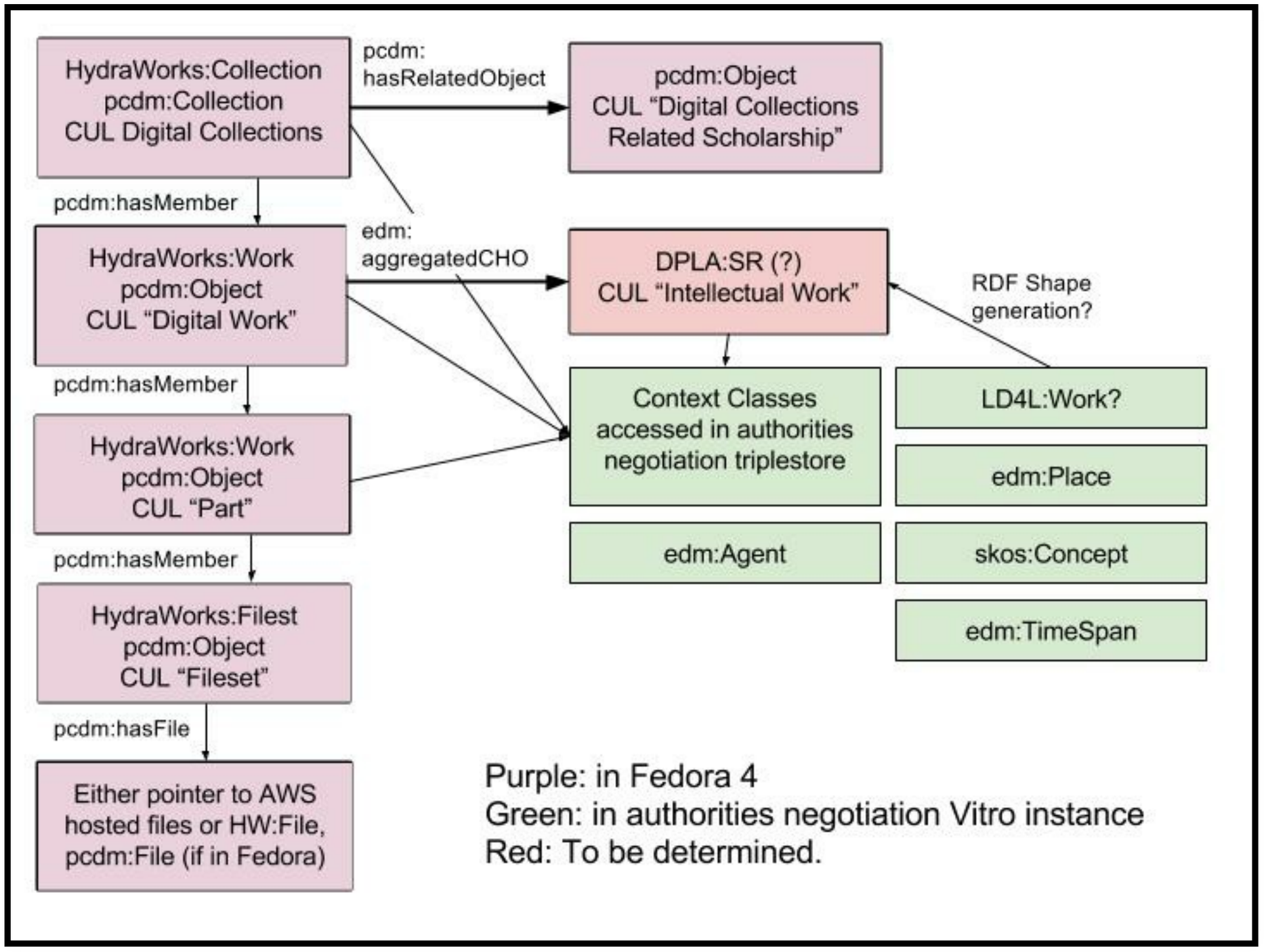
https://github.com/projecthydra/hydra-works

Classic Example: Postcards



https://github.com/hybox/models/blob/master/notes/usecase2.md

Cornell's PCDM...so far



Cornell's PCDM...so far

Bringing it all together...

And Now You...

Building off the metadata review for your objects:

1. Group Your Entities into Possible PCDM Classes;
2. Map fields to relationships between Objects;
3. Try Drawing Model in Google Drawings;
4. What Fits? What Doesn't?
5. Start thinking about properties: Linked Open Vocabs can help! http://lov.okfn.org/