Collection Architecture – Notes

These notes pull together some of the informal and email discussions we have had about the collection architecture that is required to support faculty submissions and ETDs.

Top Level Nodes in RUcore

Referring to the figure below, there is one single node (collection object) to which all of our collections are attached. This node is set up in the repository now with the collection ID of “DLR” for digital library repository. By indexing at this level, we can provide search and browse capability for all of the content in the repository.

At the second level, we have represented three major collections, each mapping to a CNRI prefix – 1782 for Rutgers University, 1782.1 for Rutgers University Libraries, and 1782.3 for NJDH. Note that we do have one other prefix, 1782.2, which is unused at this point. Note, also, that NJDH is a dynamic collection and, to minimize clutter in the figure, I have not shown the RUL sub-collections. For example, Special Collections (SPCOL) is a sub-collection of the RUL collection.

ETDs

For ETDs at the next level, I am proposing a single collection object (RUETD) which will aggregate all of the ETD collections for all graduate schools. Under this single collection object, we would have sub-collections for each graduate school, e.g. GSNB, GSAPP, SCILS, etc. All of the dissertations/theses for a particular graduate school would be aggregated under the respective collection object. The collection ID, object architecture and CNRI prefix are brought together in the persistent ID. The PID has a general format of prefix/[collection ID].[object architecture].[unique integer identifier]. For ETD collection IDs, we have decided to use the following syntax: “etd-[id for grad school]. For example, the PID for a dissertation from the GSNB would therefore look like the following: 1782/etd-gsnb.etd.2345. For a dissertation from SCILS, the PID would look like the following: 1782/etd-scils.etd.2346

We have discussed the possibility of using dynamic collections to allow a specific department to include their dissertations in a department collection. This approach is illustrated by showing dissertations D1 and D2 being pulled into a department collection (the dashed line indicates dynamic membership).

Department and Faculty Collections

At the same level as the RUETD collection, I have inserted a collection object in the figure to aggregate all of the collections for academic departments. So, as shown in the figure, a department might have many personal faculty collections (pre-prints, post-prints, technical reports, etc) and also have a collection object under which all of their ETDs would be pulled together dynamically. Note, that the metadata must include the department name, otherwise we have no way to create a dynamic collection for the department.

Collection Hierarchy in the User Interface

In our most recent steering committee meeting, we agreed that showing three levels of collection hierarchy in the user interface should be sufficient. This premise suggests that the collection picklist on RUcore would appear as shown in the following examples:

For faculty collections (3 levels):

		→ Mary Smith
	→ Anthropology Department	→ Jane Doe Coll
		→ John Smith
Department Collections	→ History Department	→ John Doe Coll
	→ Sociology Department	→ Pete Doe Coll

For ETDs (2 levels):

	→ GSAPP
Dissertations	→ GSNB
	→ SCILS

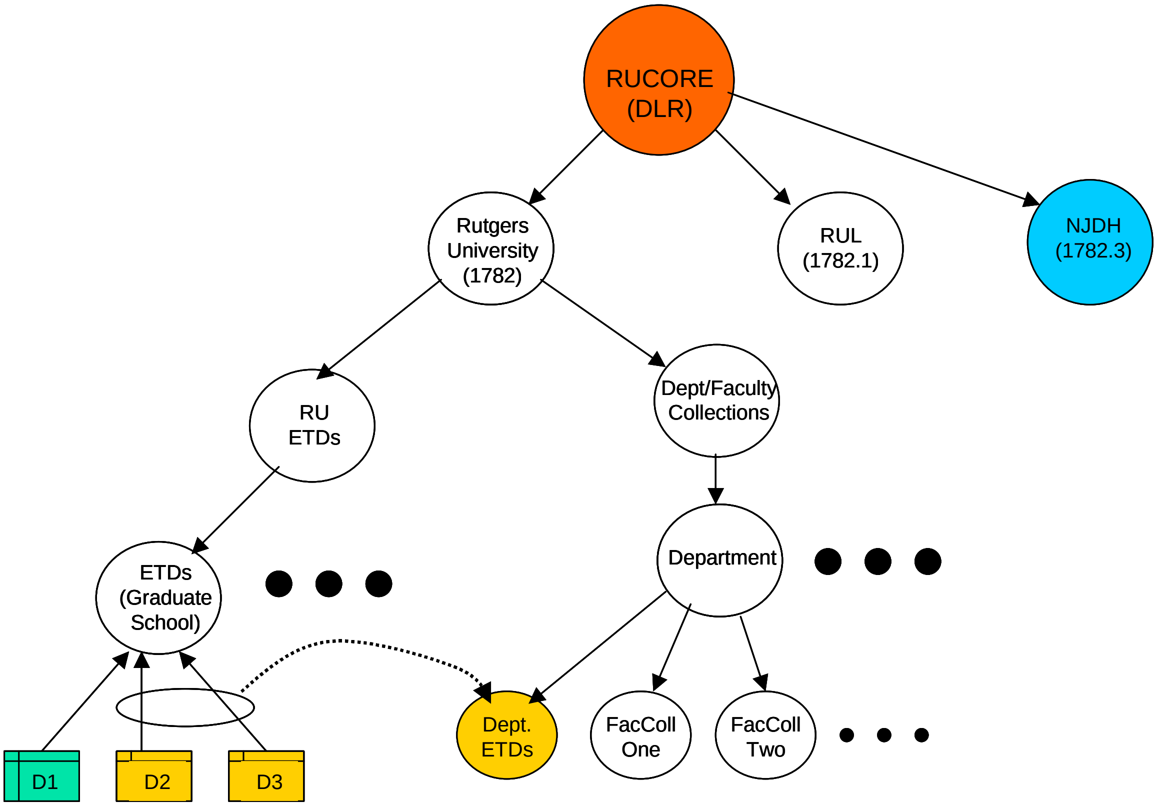


Figure - Collection Architecture

Types of Collections and Indexing

Given the above collection configuration, we actually now have four possible generic types of collections:

1) Simple collection with no sub-collections and with no dynamic aspects. The collection object would use the default structure map. Most of the collections in RUcore are of this type.
2) Dynamic collection in which all collection content is determined by search criteria specified in the structure map of the collection object. An example is our small collection of early English broadsides.
3) Hierarchical collection in which the hierarchy is specified in the structure map (see below for an example).
4) Hybrid of dynamic and hierarchical in which both search criteria and specific sub-collections are represented in the structure map. Although we have not used the term, NJDH is an example of this type.

The object architecture designations for these respectively are: collection, dynamiccollection, hcollection, and hdcollection (for hierarchical and dynamic). As a more flexible and efficient indexing technique, the proposal here is to do separate indexing of each collection as part of the nightly indexing process and to concatenate the selected collections together in real-time based on the user's search criteria.

Hierarchy Backend

In order to implement the collection hierarchy several options surfaced.

A) We could use the SMAP as the record of note and mine data from it to directly generate collection hierarchy. Any editing of the collection structure would have to happen to the SMAP.

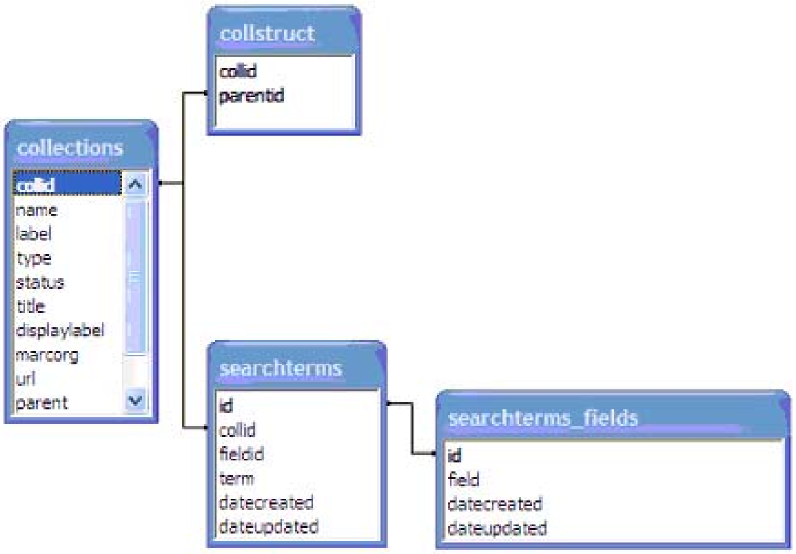
B) Another possibility is to extract the structure information from the SMAP, populate a database and read from that. It would need to be agreed upon that the SMAP is the record of note still and the database would effectively be read-only, except for when the SMAP is edited the database would need to have the appropriate edits as well. While this is an added step to option (A) it does give the flexibility of the database for creating the user-interface desired. This also introduces the possibility of having data get out of sync.

C) The last option is to have all collection hierarchy described in a database and when there are any edits/additions/etc. we simply write a new SMAP and store that in the collection object and use it for historical/preservation purposes only. Since we were already using a database to partially describe the collections in the user interface this seems the viable option.

Database Implementation

To implement option (C) three new tables would need to be created in the collections database.

1) collstruct – A table that maps parents with children using the collid in the collections tables.
2) searchterms – A table to store all search terms for dynamic collections
3) searchterms_fields – A table to store a controlled vocab list of fields that could be used in the search terms.



Database Diagram

Above is a diagram of the database with these three tables added. To describe a parent/child relationship we store the collections.collid of the parent in collstruct.parentid and the collections.collid of the child in collstruct.collid. Using a JOIN a query can be written to traverse this structure a predefined number of depths. It was agreed that a level of 3 would be sufficient for now, but the script would generate the query dynamically so any depth can be defined at a later date with little to no change in the code. The depth would be a parameter passed to the function that generates the query on the fly. In testing a max depth of 40 levels was found, due to MySQL not allowing for anymore JOINS in the query statement. To create a structure map we wish to display every level down that the collections goes, so no end point is defined and we need to find one. To find an end point we would have to probe the database until NULL results are returned at a level and then stop, this is our max depth for the collection. We can then generate the SMAP for a hierarchal collection. There is nothing interesting going on with the search terms tables, they simple store the term and associate it to the collection using the collections.collid. This would then queried to create the SMAP as well.

User Interface

The user interface will only display collections of type 1 or 3 and that are marked active in the collections table. This has not changed from the last release. Users will be able to choose multiple collections to search, instead of just one, which has been a limitation. Collections are passed in an array in the search objects code in the URI. Updates to this collection label, type and activity should be done in DLR/edit which is where the options are in the collection record edit screen for editing search terms and making a collection a member(child) of another.

Structure Map

The purpose of the structure map will be for historical purposes, in the event the collections history/life needs to be examined. I am proposing that the collection hierarchy and type (hierarchal, dynamic, hybrid or simple) be represented in the structure map of the collection object. This structure map will be easily produced from the collection DB. The following two xml segments show how the structure map would look for the RU ETD collection and for the department/faculty collections.

Department/Faculty Collection Structure Map (3 levels)

```
<METS:structMap TYPE="logical">
<METS:div1 NAME="RUDEPT" ORDER="1" TYPE="hcollection" LABEL="RU Department and
Faculty Collections" >
       <METS:div2 ORDER="2" NAME="Anthopology" TYPE="collection"
LABEL="Anthropology Department Collection" >
              <METS:div3 ORDER="3" NAME="Jane Doe" TYPE="collection" LABEL="Jane
       Doe Collectiion" />
       <METS:div2/>
       <METS:div2 ORDER="2" NAME="History" TYPE="collection" LABEL="History
Department Collection" >
              <METS:div3 ORDER="3" NAME="John Doe" TYPE="collection" LABEL="John
       Doe Collection" />
       <METS:div2/>
       <METS:div2 ORDER="2" NAME="Sociology" TYPE="collection" LABEL="Sociology
Department Collection" >
              <METS:div3 ORDER="3" NAME="Pete Doe" TYPE="collection" LABEL="Pete
       Doe Collection" />
       <METS:div2/>
<METS:div1 />
</METS:structMap>
```

RU ETD Collection Structure Map (2 levels)

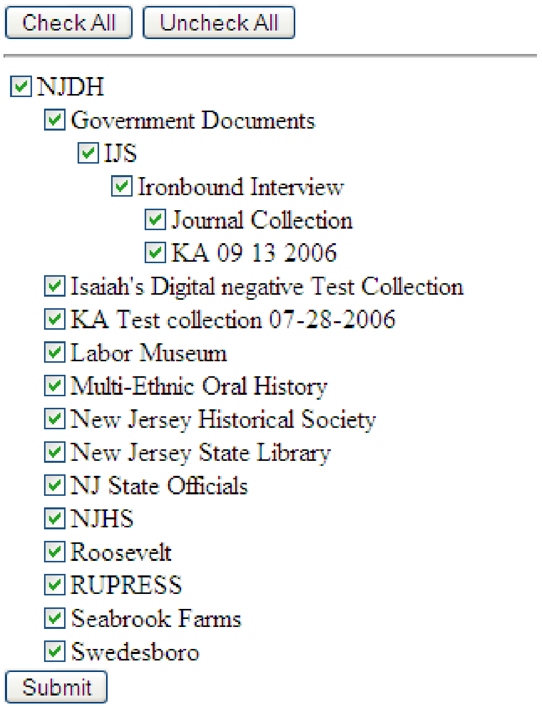
```
<METS:structMap TYPE="logical">
<METS:div1 NAME="RUETD" ORDER="1" TYPE="hcollection" LABEL="RU Electronic
Theses and Dissertations" >
       <METS:div2 ORDER="2" NAME="GSNB" TYPE="collection" LABEL="Graduate School
of New Brunswick" />
       <METS:div2 ORDER="2" NAME="GSAPP" TYPE="collection" LABEL="Graduate
School of Applied Psychology" />
       <METS:div2 ORDER="2" NAME="SCILS" TYPE="collection" LABEL="School of
Communication, Information and Library Studies" />
<METS:div1 />
</METS:structMap>
```

Sample of NJDH on the test system, using the database solution

SMAP

```
<?xml version="1.0" encoding="utf-8"?>
<structMap TYPE="logical" LABEL="default">
 <div1 ID="div1" TYPE="hdcollection" NAME="NJDH" LABEL="NJDH">
  <div2 TYPE="collection" ORDER="2" name="GovDocs" label="Government Documents" />
   <div3 TYPE="collection" ORDER="3" name="IJS" label="IJS" />
    <div4 TYPE="collection" ORDER="4" name="Ironbound" label="Ironbound Interview" />
     <div5 TYPE="collection" ORDER="5" name="JCOLL" label="Journal Collection" />
     <div5 TYPE="collection" ORDER="5" name="KA091306" label="KA 09 13 2006" />
  <div2 TYPE="collection" ORDER="2" name="DNGTEST" label="Isaiah's Digital negative Test
  Collection" />
  <div2 TYPE="collection" ORDER="2" name="KA072406" label="KA Test collection 07-28-2006" />
  <div2 TYPE="collection" ORDER="2" name="ALMBHNL" label="Labor Museum" />
  <div2 TYPE="collection" ORDER="2" name="szhis004" label="Multi-Ethnic Oral History" />
  <div2 TYPE="collection" ORDER="2" name="NJHS" label="New Jersey Historical Society" />
  <div2 TYPE="collection" ORDER="2" name="NJSL" label="New Jersey State Library" />
  <div2 TYPE="collection" ORDER="2" name="NJSO1876" label="NJ State Officials" />
  <div2 TYPE="collection" ORDER="2" name="njhs" label="NJHS" />
  <div2 TYPE="collection" ORDER="2" name="Roosevelt" label="Roosevelt" />
  <div2 TYPE="collection" ORDER="2" name="RUPRESS" label="RUPRESS" />
  <div2 TYPE="collection" ORDER="2" name="SBFarms" label="Seabrook Farms" />
  <div2 TYPE="collection" ORDER="2" name="Swedesboro" label="Swedesboro" />
 <div TYPE="DynamicCollection" ORDER="2">
  <div FORMAT="all" ORDER="3">
   <area OBJECT="Civil War" />
  </div>
  <div FORMAT="book" ORDER="3">
   <area OBJECT="New Jersey" />
  </div>
  <div FORMAT="all" ORDER="3">
   <area OBJECT="New Jersey AND RUL AND Government Documents" />
  </div>
 </div>
</div1>
</structMap>
```

User Interface



- ☑ NJDH
  - ☑ Government Documents
    - ☑ IJS
      - ☑ Ironbound Interview
        - ☑ Journal Collection
        - ☑ KA 09 13 2006
  - ☑ Isaiah's Digital negative Test Collection
  - ☑ KA Test collection 07-28-2006
  - ☑ Labor Museum
  - ☑ Multi-Ethnic Oral History
  - ☑ New Jersey Historical Society
  - ☑ New Jersey State Library
  - ☑ NJ State Officials
  - ☑ NJHS
  - ☑ Roosevelt
  - ☑ RUPRESS
  - ☑ Seabrook Farms
  - ☑ Swedesboro

rcj 10/20/2006
cmm 1/8/2007