DSpace on Fedora Data Model Sketch

This page represents work in process on one possible approach to modeling a DSpace repository in Fedora. It will evolve, and may or may not succeed in that goal It exists for the purpose of conversation and collaboration and does not indicate a formal direction.

Goals:

- Outline the technical approach
- Outline the community approach

People:

Brad, Chris, TIm (Aug 16th); Also Bill and Andrew (Oct 5th)

Working from: https://wiki.duraspace.org/display/GSOC/Google+Summer+of+Code+2008+Fedora+Integration

- Bitstream persistent ID refers to where the file is on the filesystem (an internal ID)
- Missing elements:
    - Groups (containing Groups, or EPeople)
    - EPerson (related to communities and collections) - via resource policy table.
    - Metadata values associated with items, fields
    - Metadata schema

Working from: https://wiki.duraspace.org/display/DSDOC/Storage+Layer

- Question: Does this need to be completely reversible?
    - Answers: perhaps not, although it is desirable as it forms a well defined test/proof.
- Categorizing the tables in that diagram:

Tables that should be in fedora data model for DSpace:

- BITSTREAM (Mapping choice: Fedora DataStream(time series of object) using RELS-INT for missing fields - preferred, or Fedora Object) Note that more than hundreds of datastreams in a fedora object is a bad thing.
- ITEM (associates all the datastreams; Needs to be a Fedora Object, note there will be a "template for collection" variant distinguished by it's relationship type)
- BUNDLE (sets of bitstreams related to a DSpace item: one primary bundle with the original files in it; other bundles can contain derived objects; Needs to be a Fedora Object (entity, referenced by ITEM), or possibly squashed into the entities and datastreams for Fedora - challenge is whether to build fedora model / disseminators to reproduce the activities of DSpace's biz logic - do we model the semantic endpoints in a later phase?)
- EPERSON (as a Fedora Object; Identity, and authorization for activities on items, collections, communities; also used for notifications on subscriptions, and for workflows; Fedora treats this as external to the data model and handled via mechanisms like xacml; Need to translate EPERSONS into fedora entities, have mechanisms for associating them with fedora's AuthN and AuthZ systems, and potentially tease out the workflow by creating workflowstate objects or better, datastreams on the item)
- COLLECTION (as a Fedora Object; pull out the workflow steps, possibly pull out the license and copyright to their own Fedora Object or Objects)
-

Tables that should NOT be in fedora data model (for DSpace):

- CHECKSUM_RESULTS (Considered to be administrative state; ephemeral in relationship to the actual repository contents)
- CHECKSUM_HISTORY (admin state)
- MOST_RECENT_CHECKSUM   (admin state)
- ITEM2BUNDLE (db_linkage)
- BUNDLE2BITSTREAM (db linkage)

- BITSTREAMFORMATREGISTRY (Map to Fedora format URI, possibly preserve the original DSpace idea of the format in the BITSTREAM)
    - Note that DSpace sometimes uses this to represent what the support level is for a given format. We might want to model this in fedora by providing entities.
- FILEEXTENSION (Map to Fedora format URI)
- BI_* and VIEW* can be ignored as they can be readily reconstructed and are really implementation conveniences
- COMMUNITY2COMMUNITY, COMMUNITY2COLLECTION, COMMUNITY_ITEM_COUNT, COLLECTION2ITEM, GROUP2GROUPCACHE, GROUP2GROUP (db linkage)
- HANDLE (for items, collections, or communities, but not individual files; we should make the handle a RELS-EXT or alternate id, but not the pid in Fedora)
- WORKSPACEITEM (The workflow itself to be defined as an entity in a workflow content model; the state to be stored either as a property in RELS-EXT, or as a relationship to an entitiy in the workflow content model)
- WORKFLOWITEM (The workflow itself to be defined as an entity in a workflow content model; the state to be stored either as a property in RELS-EXT, or as a relationship to an entitiy in the workflow content model)
- EPERSONGROUP (This should be an external concern, expressed and handled by the xacml / fesl)
- TASKLISTITEM (Needs to be translated into the workflow content model)
- HARVESTED_ITEM (OAI_ID becomes an alt id or a relationship - can we have multiple alt ids in fedora)
- HARVESTED_COLLECTION (preferably a data stream, potentially versioned, on the COLLECTION object)
- SUBSCRIPTION (Either attribute(s) on EPERSON of interest in notification, or attributes(s) / relationships on COLLECTION, or datastream on COLLECTION - investigation needed)
- COMMUNITY (Eliminate these, and allow collections to contain other collections)

Tables needing further consideration and modeling:

- METADATAVALUE, METADATA_FIELD_ID, METADATASCHEMAREGISTRY (Map the schemas to Fedora; can a single metadata format encapsulate all the fields - an RDF blob or an XML blob in invented format) Need an extensible set of global metadata schema; can we model this in RDF with RDF schema.
- RESOURCEPOLICY (Options: in XACML, place inside the objects or collection they apply to; otherwise place them external and have the application interpret and apply - need to investigate and decide as part of an overall auth discussion)

Notes

- Both Community and collection FK links to bitstream are just for a single logo, and should be modelled separately.
- Probably want to have a workflow content model, and have each item carry a relationship to the workflow entity, and either store its state as a property in RELS-EXT or as another relationship. For now, workflows are associated with the collection, but eventually might want to allow per item relationships.
- Should we move anything presentation related out into presentation objects in a presentation content model?

Steps:

- Validate that the ER diagram is reasonably accurate as of DSpace 1.8 - (loosely checked)
- Move from the ER diagram to Fedora objects using "standard principles" - (categorization completed above)
- Using the categorizations above, propose a Fedora content model for DSpace data - TBD
- Propose a Fedora content model for DSpace workflow (separate from the content model for the repository contents) - TBD
- Build a sample conversion / crosswalk into those content models.