Set Orthogonality

Advocates: Hussein Suleman, Mohammad Zubair
Status: Open
Last Updated: 19 October 2001

Description

There is no way to determine all the sets that an identifier belongs to. This is typically referred to as set orthogonality because the protocol allows a harvester to find out which identifiers belong to a particular set but not vice versa.

This is not as much of a problem for a flat space of archives, but organizations like NDLTD and NCSTRL have already started to create hierarchical catalogs based on OAI and existing set information is lost at the very first level. Also, the Internet2 Distributed Storage Initiative wants to work on replication of OAs - this will mean harvesting every set and dealing with duplicates. Can we do this in a way that is more efficient without adding to the complexity?

Scenarios

1. A union catalog is an archive that attempts to collect the metadata from multiple archives and republish it in the form of a single collection - this is extremely useful for providing centralized services to distributed communities (e.g. NDLTD, NCSTRL). Current attempts to create union catalogs have been plagued by the problem of set membership not being transmitted by the standard ListRecords and ListIdentifiers requests. In practice this information is either lost or the mirroring algorithms are much more complex than typical harvesting algorithms. If set information was available along with the individual metadata records, this would make replication much simpler and more complete without the added effort.
2. Sets could conceivably be used for browsing through the contents of an archive. However, given a single identifier, it is not possible to find out what sets the identifier belongs to, hence, even if the sets corresponded to browse categories it would not be possible to get from the identifier to the set in which it is contained.

Issues

1. Set orthogonality requires either a new verb or an extension of the response syntax to include this information. Either way it will require more effort on the part of the data providers.
2. Set membership information is not always easily available. In many typical databases, generating all the sets for a particular identifier may require scanning through the whole database while the inverse does not require this. To alleviate this more resources may be required in order to create additional indices. Will data providers be willing to make this optimization? Can we require orthogonality or should it be optional?

Possible Solutions

1. Extend the syntax of "ListSets" to parallel the "ListMetadataFormats" verb. In particular, if an identifier is provided, then the response should indicate the list of sets that contain that identifier.
2. If the "about" container may be repeatable, allow one of these to contain a set listing. For example:

   ```
   ...
   <about>
     <sets xmlns="setnamespace" xsi:schemaLocation="setschema>
      <setSpec>set12</setSpec>
      <setSpec>set42</setSpec>
     </sets>
   </about>
   ...
   ```
3. Add set information to the headers

   ```
   ...
   <header>
     <identifier>oai:anArchive:12345</identifier>
     <datestamp>2000-01-01</datestamp>
     <set>set12</set>
     <set>set42</set>
   </about>
   ...
   ```
4. A combination of (1) and (3)