Multiple Metadata / Best Metadata Return

Advocates: Hussein Suleman, Michael Nelson
Status: Open
Last Updated: 19 October 2001

Description

The OAI protocol currently supports a simple mapping of metadata names to metadata formats, whereby a metadata record can be requested for exactly one record in exactly one format in a single GetRecord request. In the case of ListRecords, all records within a set and/or date range may be requested but there is still the restriction of a single metadata format. This is usually sufficient for simple harvesting with the intention of transferring a stream of metadata records from the source archive to a service provider.

However, in some cases, it may be desirable to obtain the most complete metadata format or a set of metadata formats for an identifier. In order to accomplish this it is currently necessary to submit multiple requests with different parameters and this is not most efficient.

Scenarios

1. A union archive is being created to collect together all the metadata for a community in order to support multiple local services at a site. This union archive contains metadata in different formats and there is no way for services built upon it to determine which metadata records to request for given identifiers without first issuing ListMetadataFormats - this results in two requests per record dissemination. Could this be accomplished with a single request ?
2. A union archive is being created to mirror the contents of an archive. In order to mirror all the metadata formats available at the source archive, the union archive needs to issue multiple ListRecords requests with different metadataFormat parameters. Can this be made more efficient ?
3. For ListRecords and GetRecord, could we specify a sequence of metadata formats (instead of just one) with the semantics that the first matching format is returned, defaulting to DC ? The advantage is that with a single request it may be possible to get the "best metadata" corresponding to the most comprehensive representation for an identifier.
4. If no single metadata format can represent the essence of a digital object, it may be necessary to use multiple metadata formats in lieu of devising a new scheme simply for aggregation. In this case, can we specify multiple formats in the request ?

Issues

1. ListIdentifiers considers each record to be an atomic entity - thus deletions are signalled on a per-record basis. ListRecords, however, is specific by metadata format and to maintain orthogonality, should possibly have an equivalent atomic version (i.e. where complete records containing all metadata formats are disseminated at once).

2. Who should choose the best metadata format(s) ? The service provider may have preferences based on what it can process but the data provider may have better mappings in some formats. If both are allowed preferences, resolution may be complex.
3. Adding more metadata formats to the request/response increases complexity - how many sites require such additions to the protocol ?

Possible Solutions

1. Do not change the protocol to accommodate multiple metadata formats for each record. Union archives can harvest multiple times and establish a best-practice for mirroring algorithms. This can include such recommendations as deleting all instances of a record if any metadata rendition has a "deleted" attribute.
2. Devise a representation expression scheme for metadata formats that is computationally complete. This would include "or" and "and" operators, supporting metadataFormat parameters such as "oai_etdms or (oai_rfc1807 and oai_dc)". This is almost definitely too complicated.
3. A two-part solution:
    1. Use a list of metadata formats to indicate preferences in descending order. This could take the form of a comma-separated list such as "oai_etdms,oai_dc" or a list in whatever scheme is chosen for representing requests (e.g. XML, SOAP), and the data provider would respond by sending back a record in the first format it can support.
    2. If no metadata format is supplied, use the preference of the data provider. This data provider preference can be adopted independently of the list and has a low cost to data providers.
4. Use a list format like above with "and" semantics rather than "or" semantics. Encode multiple <metadata> tags in each record to accommodate this.