

Registry Servers and Harvesting

Todd King
Aaron Roberts
Carl Cornwell
Raymond Walker
Jan Merka
Tom Narock

Registry Servers

- Provide the ability to extract resource descriptions from a provider.
 - Agencies
 - Virtual Observatories
 - Projects
 - Researchers
 - anyone
- Accessed in self-declared networks
 - Location A harvests from B, C, G, and H
 - Location B harvest from G, H, I

Harvesting

- The collection of data from one or more sources.
- Can involve selection
 - Only some part (the fruit) of available information is retained.

Basic Harvesting Requirements

1. All resources.
2. All resources of a certain type.
3. All resources "published" within a time frame.
 - Between two dates.
 - Prior to a date.
 - After a date.
4. A specific resource
 - Any combination of the 1, 2, and 3

Advanced Harvesting Requirements

- Resources matching description content criteria (Google-like)
 - Scored according to relevance
 - Organized by facets
- Resources matching data content criteria
 - Specific statistical attributes.

Possible Solutions (1 of 3)

- Adhoc Push-Pull files
 - XML is uploaded to a location with FTP, HTTP or e-mail.
 - or–
 - Location pulls XML files from FTP or HTTP servers.

Note: Won't address all Basic Harvesting Requirements

Possible Solutions (2 of 3)

- SPASE registry server.
 - Simple REST interface
 - ResourceID
 - ReleaseDate
 - SPASE XML response.

Note: Can support both basic and advanced requirements

Possible Solutions (3 of 3)

- **OAI-PMH**

(Open Archives Initiative - Protocol for Metadata Harvesting)

- Well documented.
- Supported by other tools.
- Response is OAI + embedded metadata (DC, SPASE).
- Supported actions ("verbs"):
 - GetRecord: A resource description
 - Identify: Registry information
 - Related action: ListMetaDataFormats
 - ListRecords: Multiple resource descriptions
 - Options "from" and "until"
 - Related action: ListIdentifiers ("headers" only)
 - ListSets: Available groupings (facets) of the data.

Note: Can support both basic requirements.

Discussion

- What requirements do we need/want
 - basic
 - advanced
- Should basic and advanced harvesting be provided by the same service?
 - Yes - Option 2 (SPASE specific)
 - No - Any option.