An OAI Approach to Sharing Subject Gateway Content

Andy Powell
UKOLN
University of Bath
Bath, UK
+44 1225 323933
a.powell@ukoln.ac.uk

ABSTRACT
The Resource Discovery Network (RDN) has taken a collaborative approach to the development of a network of subject gateways, each of which offers a variety of services to its subject-focused community. This paper considers the use of the Open Archives Initiative Protocol for Metadata Harvesting as a mechanism for sharing metadata records between those gateways in order to build cross-subject resource discovery services.

Keywords
Resource discovery, subject gateways, content sharing, metadata, Dublin Core, XML, Open Archives Initiative, RDN.

1. RESOURCE DISCOVERY NETWORK
The Resource Discovery Network (RDN) [1] is a national service funded by JISC to provide access to high quality Internet resources for the UK Higher and Further Education communities. The RDN is a cooperative network of subject gateways, including BIOME [2] (health, medicine and life sciences), EEVL [3] (engineering, mathematics and computing), HUMBUL [4] (humanities), PSIgate [5] (physical sciences) and SOSIG [6] (social science, business and law).

A subject approach allows gateways to develop strong links with subject-focused communities within the education sector. By working with subject specialists, who identify and describe Internet resources in their subject area, rich databases of high quality resource descriptions can be created. By targeting their services within a subject area, communities of end-users can also be developed.

Individual gateways provide search and browse access to their databases of resource descriptions. Whilst this forms the core of their services, other functionality such as Web indexes, full-text content provision and community-building collaborative services are also provided.

Although each gateway presents its own Web-based user-interface, they all work within a shared policy, business and technical framework that ensures a consistent set of services. This framework is defined by a set of policy documents [7] covering collection development, cataloguing guidelines, technical standards, interoperability and IPR.

Metadata records created within the RDN are based on the Dublin Core Metadata Element Set (DC) [8]. There are agreed guidelines for use of the DC, although there is some divergence across the gateways for particular elements. For example, each gateway uses a subject specific classification scheme. Although most usage falls within the 15 Dublin Core elements, there are some extensions to provide the required level of detail.

Gateways are free to implement services using software that is appropriate for their needs and experience (for example, governed by local computing service policy) provided they conform to the policies listed above.

2. SEARCHING VS. SHARING
In order to provide services across the RDN, for example to provide resource discovery services to end-users who are interested in multiple subject areas or in the intersections between RDN subject areas, we have provided centralized mechanisms for searching and browsing across all the RDN subject gateways. Although this cross-searching service is currently delivered using the Whois++ protocol, as implemented in the ROADS suite of subject gateway tools [9], we have envisaged migrating to the use of Z39.50 at some point in the future. We have also planned on implementing cross-browsing using some sort of saved-search mechanism.

However, there are problems with this approach. There are inevitably some delays associated with cross-searching. The response time for searches based on sending the same query to multiple search targets tends to be limited by the worst performing target or intervening network delays. While this might be acceptable where the user has directly initiated a search, it is less likely to be so for browsing based on cross-searching, where the underlying search is hidden from the end-user. Furthermore, it is very difficult to build flexible browse interfaces, targeted at multiple audiences, and other value added services (such a annotation and reading list services) based on a distributed set of gateway databases.

For these reasons, we have begun to investigate the possibility of basing our cross-RDN services on record sharing rather than cross-searching.

3. OPEN ARCHIVES INITIATIVE
The Open Archives Initiative (OAI) has recently developed a harvesting framework that provides a mechanism for sharing metadata records between cooperating services based on HTTP and XML. Despite its background in the e-prints community (the term ‘archive’ refers to repository of e-print articles), the Open Archives Initiative Protocol for Metadata Harvesting [10] has been specifically designed with a wide range of application areas in mind. It is potentially useful in any scenario where metadata needs to be shared between networked service components.

The OAI Protocol allows metadata records to be shared between *data providers* (repositories of metadata records) and *service providers* (services that harvest metadata from the data providers). Each record comprises three parts:

- *header* - including a unique record identifier and a datestamp,
- *metadata* – metadata about a resource in a single format.
- *about* – metadata about the record.

Data providers are allowed to support multiple metadata formats (provided they are encoded as XML) but they must be able to generate unqualified DC records conforming to an OAI-defined DC XML schema.

The OAI Protocol is intentionally very simple - it is not intended to provide sophisticated search functionality for example. The protocol defines only six requests (known as *verbs*) - *GetRecord*, *Identify*, *ListIdentifier*, *ListMetadataFormats*, *ListRecords* and *ListSets*. These are not described in detail here. Suffice to say *GetRecord* and *ListRecords* support the retrieval of records from a data provider. *ListSets* supports the grouping of records within a repository into logical groups. Records can be selectively harvested based on these sets or by the service provider asking for records corresponding to a particular range of dates.

The OAI Protocol has recently been finalized. It is hoped that the specification will remain as stable as possible for at least one year to allow an extended experimentation period.

4. IMPLEMENTATION

The RDN took part in the alpha test phase of the OAI Protocol. As part of this we implemented a simple *data provider* consisting of about 400 metadata records describing social science and philosophy Internet resources. Because there is a variety of software in use across the RDN it was decided not to tie our development to a particular subject gateway implementation. Instead, metadata records are exported from a gateway database and imported into the OAI repository. The repository is very simple. Unqualified DC records are stored as XML directly within the server filestore. A top-level directory corresponds to the repository – sub-directories correspond to repository sets. If necessary, records can be placed in multiple sets by using symbolic links. Every record contains an *about* section that indicates the gateway that originally 'published' the record and a simple rights management statement.

A single Perl CGI script supports the OAI protocol. A second Perl script is used to move and convert records from the gateway database to the repository. Clearly, this second script is specific to the gateway software in use. The intention is that the script is run regularly, for example on a nightly basis. For the alpha test we developed a version appropriate for use with the ROADS software.

5. ISSUES

5.1 Record richness

Although the use of unqualified DC in OAI is broadly in line with cataloguing practice within the RDN, the format does not support the full richness of existing RDN metadata records. For example, it is not possible to indicate the subject classification scheme that has been used. We need to determine whether unqualified DC is sufficient for our needs or whether we will have to develop a richer RDN-specific record format.

5.2 Ownership and branding

Significant intellectual effort is involved in creating RDN records - cataloguers and gateways can reasonably expect that their records contain suitable attribution. The OAI record *about* section can be used to indicate the individual that created the record (dc:creator), the gateway that originally made it available (dc:publisher) and a simple rights management statement (dc:rights). It remains to be seen if a simple, unstructured, rights statement is sufficient for our needs.

5.3 How open is open?

The OAI Protocol is 'open' in the sense that it provides a machine interface to the data provider. It does not necessarily mean 'open' in the sense of making information freely available to anyone. In the case of the RDN, our OAI repositories are unlikely to be freely available to all, at least initially. Because the protocol is layered on top of HTTP, we will make use of existing mechanisms within the Web server to restrict access using HTTP Basic Authentication or client IP address validation. In theory there is no reason why OAI service provider and data provider interactions couldn't be secured using SSL, though that is not appropriate or necessary in the case of the RDN.

6. FUTURE WORK

The RDN plan to extend experimentation with the OAI Protocol. A centralized RDN *service provider* that collects records from all the RDN gateways will be developed in order to provide the basis for a central RDN search and browse interface and other value-added services.

7. REFERENCES

[1] Resource Discovery Network http://www.rdn.ac.uk/

[2] BIOME, http://biome.ac.uk/

[3] EEVL, http://www.eevl.ac.uk/

[4] HUMBUL, http://www.humbul.ac.uk/

[5] PSIgate, http://www.psigate.ac.uk/

[6] SOSIG, http://www.sosig.ac.uk/

[7] RDN Policy Documents http://www.rdn.ac.uk/publications/policy.html

[8] Dublin Core Metadata Element Set http://purl.org/dc/documents/rec-dces-19990702.htm

[9] ROADS http://www.ilrt.bris.ac.uk/roads/

[10] The OAI Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.htm