Collecting metadata from institutional repositories

Gordon Dunsire, ***Centre for Digital Library Research, University of Strathclyde, Glasgow, UK***

Abstract

The purpose of this article is to review metadata issues identified in recent research carried out in Scotland on services based on metadata aggregation via OAI-PMH, and to examine the role of collection-level description in managing ingest to harvested repositories, subsequent harvesting by secondary aggregators, and the contextualisation of institutional and aggregated repositories in the wider information retrieval environment.

The paper reviews the output of several projects involving institutional repositories and collection-level description in Scotland.

Collection-level description is a useful tool for aggregator services, but further work is required to accommodate information about the manipulation of metadata sets. Communities need to consider how best to incorporate structured collection information within the OAI-PMH for their specific purposes.

The paper shows the importance of recent developments in collection description metadata for implementors of OAI-PMH services, building on the simple placeholders for such metadata allowed by the protocol.

Keyword(s):

Information retrieval; Data collection; Information facilities.

Introduction

This paper focuses on issues encountered by services which aggregate metadata from multiple institutional repositories using the Open Archives Initiative protocol for metadata harvesting (OAI-PMH). The paper uses the term repository to refer to local and aggregated sets of metadata records, rather than collections of the resources described by those metadata. Only aspects of the metadata affecting information retrieval are considered.

Background

The CATRIONA II project which ran from 1996 to 1999 investigated the existence and management of quality, locally-created electronic teaching and research materials in Scottish universities, and examined issues associated with the management of wider access to such resources from within and without the institution. The project confirmed that significant quantities of materials were being created in all types of university, and that access to it was

severely restricted even though it was perceived to be of use to members of the university and others (Nicholson and Gold, 1998).

The project concluded that there was a strong case for individual institutions to develop services to make their teaching and research output more accessible, and that local efforts would require national co-ordination to address interoperability issues affecting access across multiple institutions. The positive role of the library in managing metadata and other resource access services was emphasised (Nicholson *et al.*, 1999). Although CATRIONA II did not research specific metadata issues, several subsequent projects carried out in Scotland have investigated aspects of metadata in institutional repositories and aggregation services.

The Harvesting Institutional Resources in Scotland Testbed (HaIRST) project researched the design, implementation and deployment of a pilot service for access to resources created autonomously by Scottish tertiary education institutions, including colleges and universities. An important aim of HaIRST was to identify issues of metadata interoperability arising from the requirements of local institutional repositories and their impact on services based on the aggregation of harvested metadata. The project ran from 2002 to 2005.

The Managing Digital Assets in Tertiary Education (Mandate) project developed a toolkit to support the creation and implementation of digital asset management and preservation in the further education environment, and demonstrate its application in the context of John Wheatley College in Glasgow. A part of the overall approach was to develop workflow models and templates to support the effective creation of metadata suitable for storage and retrieval processes and supporting managed information lifecycles. A specific application for the College was an OAI-compliant server for sharing resource information with other institutions, building on the service developed by the HaIRST project. Mandate ran from 2004 to 2005.

The aim of the STARGATE project was to lower technical barriers to the implementation of OAI-compliant repositories by exploring the use of static repositories to expose publisher metadata to OAI-based disclosure, discovery and alerting services. The project also built on the infrastructure created during the HaIRST project, and ran from 2005 to 2006.

The Institutional Repository Infrastructure for Scotland (IRIScotland) project is ongoing, and aims to develop a cross-repository infrastructure to promote the research output of Scottish education institutions as a whole, including agreements on design and metadata standards and a fully working service implementation. The project started in 2005 and is due for completion in 2007.

Findings

The efficiency and effectiveness of any information retrieval service requires coherency and consistency in its metadata. Aggregator services potentially face two distinct but related categories of variation in harvested metadata: structure and content.

Although the provision of an unqualified Dublin Core (DC) metadata structure, oai-dc, is mandatory for repository compliance with the OAI-PMH, the protocol allows for other metadata

structures to be harvested. The reduction of variation in metadata structure within a community of institutions can be achieved by a community-wide agreement either to use the same structure in every local repository or one of a set of structures which can be mapped to a common structure within the aggregator service. The limitations of oai-dc as a metadata structure to meet functional requirements even in a simple environment for learning and administrative resources were exposed in the model metadata schema and mappings in the Mandate toolkit (Robertson *et al.*, 2006). The limitations of unqualified DC applied to eprints and related research outputs have also been discussed in relation to UK initiatives including IRIScotland (Eprints Application Profile Working Group, 2006). Aggregator services based on oai-dc are therefore not likely to meet the information retrieval functions required by many communities; the IRIScotland service requirements for retrieval by institution, department and Research Assessment Exercise (RAE) unit identified from a survey of academic authors (Dunsire, 2006a) cannot be supported by unqualified DC.

Community agreement on a single metadata structure richer than unqualified DC is likely to be hampered because there is wide variation in the scope of resources to be described within a local repository, leading to divergent functional requirements between the institution and the community. HaIRST identified institutions using MARC21 records in the library catalogue to describe teaching, learning, research and administrative resources, while John Wheatley College's implementation of the Mandate toolkit uses an in-house metadata structure. Variation also exists where the coverage of a repository is restricted to a specific class of resources; some members of IRIScotland offer only theses and dissertations, while others include working papers, pre-prints, conference presentations, and other materials associated with research. Efforts by any single institution to reconcile local requirements with those of the community are also likely to be stymied by participation in multiple communities with differing goals. For example, if full operational services are developed from the IRIScotland and Electronic Theses Online Service (EThOS) projects, Scottish university repositories may have to offer metadata compatible with both (Dunsire, 2006b).

It is therefore likely, for the foreseeable future, that aggregator services will have to harvest a variety of metadata formats and map them to a common structure, as confirmed by the experience of the National Science Digital Library (NSDL) in the USA (Hillmann *et al.*, 2004). The same observation can be made about individual institutions and aggregator services which ingest multiple formats to a single repository with the aim of exposing the metadata to harvesters using the OAI-PMH. An example of an individual institution engaged in this activity is the Los Alamos National Laboratory (Goldsmith and Knudson, 2006), while a potential aggregator service is represented by STARGATE which recasts existing metadata to oai-dc format for static repositories (Robertson, 2006).

Variation in the content of institutional repository metadata can be caused by the same lack of clarity of the scope of the repository and its functional requirements. There are additional factors, including variation in the skills and training of those creating the records, absence of support tools such as data-entry guidelines or authority files for names and subjects, and the legacy effects of changing guidelines and practice through time. IRIScotland has detected widespread errors in the subject metadata it has harvested (Dawson, 2006), with significant impairment to the functionality for subject retrieval identified by the academic author survey (Dunsire, 2006a).

Communities can reduce variation by adopting an application profile or common set of content guidelines associated with an agreed metadata structure as recommended by the HaIRST project (Dunsire, 2005), but there is little consistency between different communities, even if they have similar functional requirements. HaIRST identified contradictory guidelines in different application profiles; for example, a personal name in the metadata elements for author and contributor is entered "as it appears on the title page" in the Networked Digital Library of Theses and Dissertations in the USA (Atkins *et al.*, 2006), while the draft UK Eprints application profile, which also covers theses, advises entry in surname-comma-forename format (Eprints Application Profile Working Group, 2006). An institutional repository is unlikely to develop consistent guidelines which satisfy both approaches, and aggregator services may need to develop tools to reconcile structural differences in content.

Aggregator services therefore need to be aware of local metadata structure and content policies for each repository they harvest, if they are to meet functional requirements beyond those supported by unqualified DC. Such information can be used to support the automated parsing and recasting of local structure and content, where consistency allows, into that required by the service. The STARGATE and IRIScotland pilot services have shown that this information can usefully include the syntax and semantics of structure attributes to allow content to be mapped correctly, and which attributes extraneous to the service can be dropped. It is also possible to add missing content during ingestion, for example the resource format when it is implicit in a local repository scoped only for that format. If the aggregator service itself acts as a repository for harvesting, information about the original set of metadata and its transformation may be useful to secondary aggregator services, and so on. Again, these findings are confirmed elsewhere (Hillmann *et al.*, 2004; Lagoze *et al.*, 2006).

Role of collection-level description

A local repository can be treated as a collection of metadata records, so any information about the repository as a whole can be regarded as an attribute of a collection-level description (cld). A repository is equivalent to an analytic finding-aid or catalogue, which can be described using a subset of the general attributes identified for cld (Heaney, 2000). These attributes include the electronic location of a repository, a description of its scope and purpose, metadata format, and information about the institution which acts as owner, creator and collector of the metadata records and as administrator of the repository service. More specific attributes deemed useful to metadata aggregator services could extend the subset to identify application profile, missing elements, and other local information, as well as service-required parsing information such as what elements and values are added during ingest. This is similar to the approach taken by the Collection Registration Service of NSDL (Lagoze *et al.*, 2006)

It is worthwhile taking a consistent, structured approach to cld because the data can be used for several purposes by aggregator services. As well as driving ingest processes, the information can support the user interface, provide "explain" facilities to secondary aggregator services, and relate the service to wider contexts and information environments. These functions have been researched and demonstrated during the HaIRST, STARGATE, and IRIScotland projects using the Scottish Collections Network (SCONE) collection descriptions service.

SCONE uses a cld schema based directly on Heaney's (2000) model and subsequently refined and extended in a number of research projects (Dunsire, 2002, 2004a). SCONE has been integrated with the Co-operative Information Retrieval Network for Scotland (CAIRNS), a virtual metadata aggregator service based on the Z39.50 protocol, to create a pilot Scottish information environment (Dunsire, 2004b). In particular, the SCONE metadata controls the catalogue selection function of the CAIRNS user interface by greying-out and removing target metadata sets which do not support the user-selected search option. This benefits the user by shortening the time taken to search the aggregated repository, and confirming that failed searches are the results of the user's query and not local metadata policies. The interface also provides descriptions of the collections described by the target catalogues. The extension of this functionality to harvested metadata aggregations is being tested with the IRIScotland cross-repository service.

All operational institutional repositories in Scotland identified or created during the HaIRST, STARGATE, and IRIScotland projects have been registered in SCONE following standard service guidelines for analytical finding-aids. The SCONE interface does not expose finding-aids directly; rather, it is metadata about their corresponding resource collections which are searched and displayed. SCONE uses the concept of functional granularity (Heaney, 2000) to infer the existence of a collection of resources, even if it is physically distributed, if there is an analytic or hierarchical finding-aid for it. So for each institutional metadata repository, it is assumed there is an institutional resource repository. Furthermore, a metadata aggregation is treated as a separate analytic finding-aid, with the functionally equivalent resource collection being the aggregation of the resources described, albeit distributed. Aggregations and their constituent repositories are related hierarchically, as super- and sub-collection. This allows the institutional repository to be contextualised with other resource collections owned or made available by the institution, such as the library, and any aggregator services which harvest the repository. The internal logical structure of the repository can also be represented by appropriate sub-collection descriptions. An example is the University of Stirling Digital Research Repository, which is logically divided into departmental “communities”, and is harvested by at least two aggregator services, including IRI-Scotland (Figure 1).

The OAI-PMH itself provides three distinct ways of accommodating cld information, in the “about”, “setDescription” and “description” containers. “About” is a record- or item-level attribute which is recommended for provenance information to track harvesting history and changes for records which have been harvested and are subsequently exposed by an aggregator service for re-harvesting by secondary aggregators (Lagoze *et al.*, 2002a). The XML schema provided in the OAI-PMH (Lagoze *et al.*, 2002b) has been extended by NDSL to provide specific information about changes applied at ingest, rather than just a flag to indicate that the record has been altered (Hillmann *et al.*, 2004). It is necessary to accommodate the information at the item level because the OAI-PMH allows harvesting of a single record from a repository and the protocol must be able to handle the situation where transformation has been carried out on some, but not all, records in a particular ingest.

Where an automated transformation is applied to every record in an ingested set, however, it seems redundant to provide the same information in every record subsequently represented for secondary harvesting. Information about a collection-level transformation, or a pointer to it,

could be accommodated in the repository-level “description” container, along with the rest of a cld for the repository as a whole, including scope and additional information about the institution. Similarly, the set-level “setDescription” container can be used to store or point to a cld for the sub-collection represented by the set, and functional granularity can be invoked to establish an equivalence between set and collection in every repository where the set container is used. It should be noted that sets are excluded from the static repository specification, so general cld can only be used at the repository level, suggesting that sub-collections are best treated as separate collections each with its own repository.

All three containers available in the OAI-PMH are optional, and the protocol expects communities to develop guidelines on their use and suitable XML schemas for expressing the content. Two simple cld schemas currently in development, the Dublin Core collection description application profile and NISO Collection description specification, are independent of any traditional information management domain, but neither accommodates attributes for collection-level processing appropriate to aggregator services.

Conclusions

Collection-level description can be a useful tool for aggregator services, but further work is required to accommodate structured information about the manipulation of metadata sets, both to assist with automatic processing at ingest and to expand provenance data for secondary harvesting. Institutions and communities need to consider how best to integrate structured collection information with the OAI-PMH for their specific purposes, and they should be aware of the use of collection-level description services to landscape or simplify user access to complex information environments.

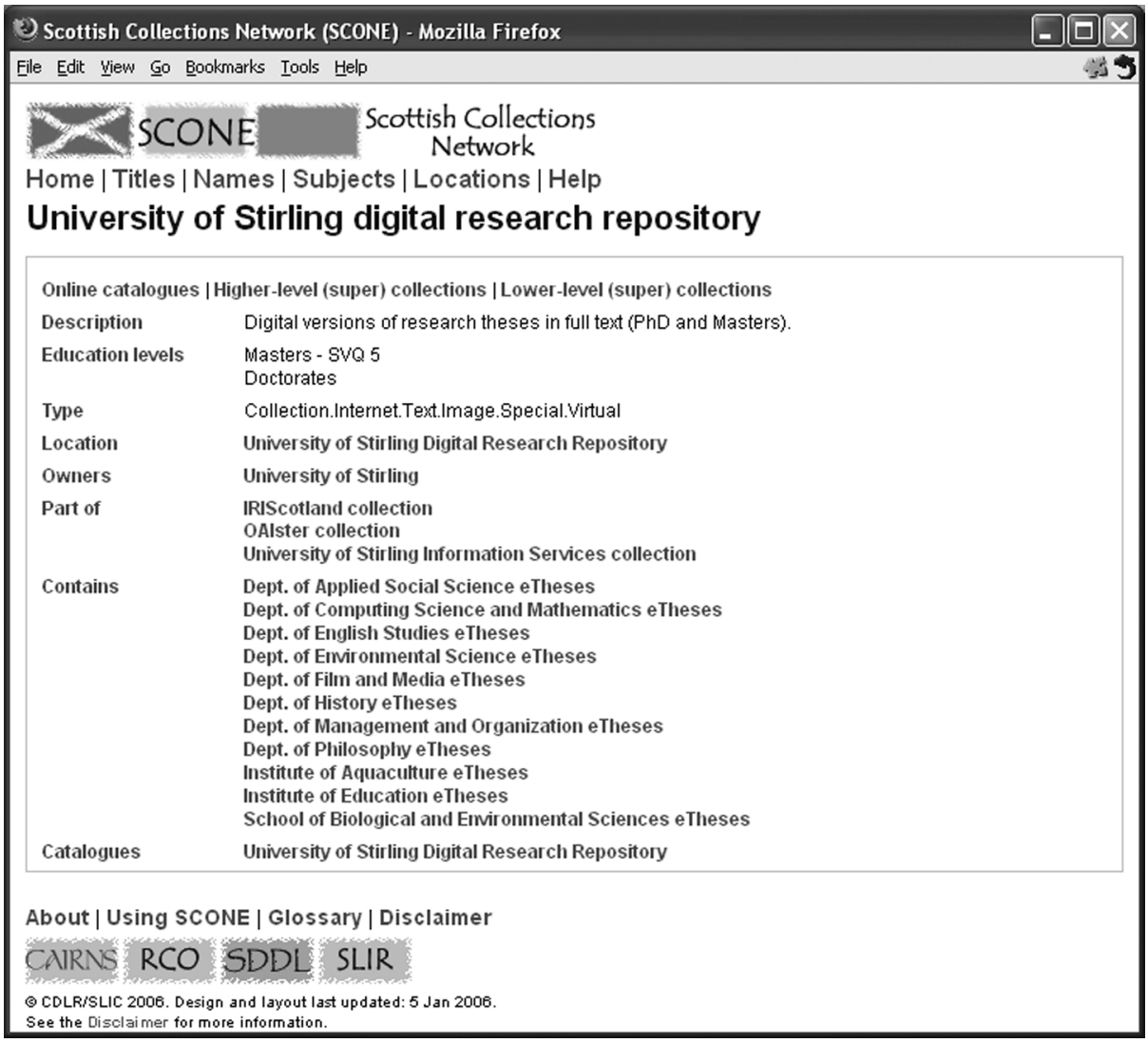

Figure 1*Contextualisation of resource repository collections*

References

Atkins, A., Fox, E., France, R., Suleman, H. (2006), "ETD-MS: an interoperability metadata standard for electronic theses and dissertations", version 1.00, revision 2, available at: www.ndltd.org/standards/metadata/current.html, .

Dawson, A. (2006), "Thirty problems for subject interoperability (and a few possible solutions)", available at: http://iriscotland.nls.uk/wiki/index.php/WP5MS0608, .

Dunsire, G. (2002), "Technical and functional description of the SCONE demonstrator service: final report of the SCONE RSLP project", annexe B.1, available at: http://cdlr.strath.ac.uk/pubs/dunsireg/SCONEFPNXB1.pdf, .

Dunsire, G. (2004a), “Extending the SCONE collection descriptions database for cc-interop: report for work package B of the cc-interop JISC project”, available at: http://cdlr.strath.ac.uk/pubs/dunsireg/CCIExtendSCONE.pdf, .

Dunsire, G. (2004b), “Collection landscaping in the common information environment: a case study using the Scottish Collections Network (SCONE): Report for Work Package B of the JISC CC-interop project”, available at: http://cdlr.strath.ac.uk/pubs/dunsireg/CCICLDLandscaping.pdf, .

Dunsire, G. (2005), “Harvesting institutional resources in Scotland testbed project: final report”, available at: http://cdlr.strath.ac.uk/pubs/dunsireg/HaIRSTFinal.pdf, .

Dunsire, G. (2006a), “Analysis of the cross-repository service retrieval question of the academic author survey”, available at: http://iriscotland.nls.uk/wiki/index.php/Analysis%5Fof%5Fthe%5Fcross-repository%5Fservice%5Fretrieval%5Fquestion%5Fof%5Fthe%5Facademic%5Fauthor%5Fsurvey, .

Dunsire, G. (2006b), “Implications of the EThOS project”, available at: http://iriscotland.nls.uk/wiki/index.php/Implications%5Fof%5Fthe%5FEThOS%5Fproject, .

Eprints Application Profile Working Group (2006), “Eprints application profile”, available at: www.ukoln.ac.uk/repositories/digirep/index/Eprints%5FApplication%5FProfile, .

Goldsmith, B., Knudson, F. (2006), "Repository librarian and the next crusade: the search for a common standard for digital repository metadata", *D-lib Magazine*, available at: http://dlib.ukoln.ac.uk/dlib/september06/goldsmith/09goldsmith.html, Vol. 12 No.9, .

Heaney, M. (2000), “An analytical model of collections and their catalogues, third issue, revised”, available at: www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf, .

Hillmann, D., Dushay, N., Phipps, J. (2004), “Improving metadata quality: augmentation and recombination”, paper presented at DC-2004: International Conference on Dublin Core and Metadata Applications, Shanghai, China available at: http://metamanagement.comm.nsdl.org/Metadata%5FAugmentation–DC2004.html, .

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S. (2002a), “Guidelines for aggregators, caches and proxies, protocol version 2.0 of 2002-06-14”, available at: www.openarchives.org/OAI/2.0/guidelines-aggregator.htm, .

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S. (2002b), “XML schema to hold provenance information in the ‘about’ part of a record, protocol version 2.0 of 2002-06-14”, available at: www.openarchives.org/OAI/2.0/guidelines-repository.htm, .

Lagoze, C., Krafft, D., Cornwel, T.l., Dushay, N., Eckstrom, D., Saylor, J. (2006), “Metadata aggregation and ‘automated digital libraries’: a retrospective on the NSDL experience”, preprint submitted to JCDL2006, available at: http://arxiv.org/ftp/cs/papers/0601/0601125.pdf, .

Nicholson, D., Gold, J. (1998), “Full report and conclusions: electronic research and teaching resource creation at six Scottish universities (random surveys)”, available at: http://cdlr.strath.ac.uk/pubs/nicholsond/catriona2full.pdf, .

Nicholson, D., Dunsire, G., Smith, M., McLeod, I., Fletcher, M., Gold, J. (1999), “Should universities manage services offering institutional or extra-institutional access to locally-created electronic teaching and research resources? Final report of the CATRIONA II project”, available at: http://cdlr.strath.ac.uk/pubs/nicholsond/catriona2final.pdf, .

Robertson, R., Green, C., Kearney, C., Dunsire, G. (2006), “Managing digital assets in tertiary education toolkit version 2.2”, available at: http://mandate.cdlr.strath.ac.uk/docs/mandatetk.pdf, .

Robertson, R. (2006), “Stargate final report”, available at: http://cdlr.strath.ac.uk/pubs/robertsonr/Stargate FinalReport1%5F4.pdf, .

Further Reading

CAIRNS: Co-operative Information Retrieval Network for Scotland available at: http://cairns.lib.strath.ac.uk, .

Dublin Core collection description application profile available at: http://dublincore.org/groups/collections/ collection-application-profile/, .

Harvesting Institutional Resources in Scotland Testbed (HaIRST) available at: http://hairst.cdlr.strath.ac.uk/, .

Managing Digital Assets in Tertiary Education (mandate) available at: www.jwheatley.ac.uk/mandate/, .

SCONE: Scottish Collections Network available at: http://scone.strath.ac.uk/Service/Index.cfm, .

STARGATE: Static Repository Gateway and Toolkit: enabling small publishers to participate in OAI-based services available at: http://cdlr.strath.ac.uk/stargate/, .

Institutional Repository Infrastructure for Scotland (IRIScotland) available at: www.iriscotland.lib.ed.ac.uk/, .

IRIScotland Project Pilot Cross-Repository Service available at: http://cdlr.strath.ac.uk/iriscotland/, .

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S. (2002), “Guidelines for repository implementers, protocol version 2.0 of 2002-06-14”, available at: www.openarchives.org/OAI/2.0/guidelines-repository.htm, .

NISO Z39.91-200x: “Collection description specification” "available at: http://www.niso.org/standards/standard_detail.cfm? std_id=815", .

About the author

Gordon Dunsire has been involved in a number of research projects in the areas of collection-level metadata, distributed information retrieval services, and institutional repositories. He is the principal developer of the Scottish Collections Network and CAIRNS distributed union catalogue. He is a member of several groups developing standards for metadata interoperability at Scottish, United Kingdom, and international levels.