

IRIScotland Project

Draft IRIScotland metadata agreement: standards and guidelines for institutional repositories

Alan Dawson and Gordon Dunsire

Centre for Digital Library Research, University of Strathclyde

June 2007

Revised as part of the IRIScotland project extension: October 2008

The original version of June 2007 is available at:

<http://cdlr.strath.ac.uk/pubs/dawsona/IRISMetadataDraftv1.pdf>

Executive summary and recommendations	p.2
Background	4
Impact on IRIScotland cross-repository service	5
Methodology	5
Scope	6
Developments since the original version of this report	6
Format of proposals and recommendations	9
Specific recommendations for each Dublin Core element	10
Example metadata record in OAI_DC format	29
Other cataloguing and data entry issues	30
Model guidelines for creating metadata in institutional repositories	31
Draft IRIScotland metadata agreement – what is it good for?	32

Executive summary

The IRIScotland pilot cross-repository search service is intended to enhance exposure of the research output of Scotland.

The quality of the service is dependent on the consistency, coherency and completeness of the metadata it provides.

Quality can be improved if there is a cross-repository agreement on how metadata created in local institutional repositories is made available to the service.

General recommendations

All repositories using the relevant technical standards must make local metadata available for cross-repository aggregation in the prescribed OAI_DC format. The IRIScotland project has identified several areas where the quality of the metadata in this format can be improved by local repository services. In most cases, these improvements can be made quickly and at low cost, and will not impact on the maintenance of metadata for local purposes.

- As a short-term solution, repositories included in the IRIScotland cross-repository search service should implement the OAI_DC recommendations given in this document.

Although this will be of significant benefit to the cross-repository service, it will not in itself be sufficient to address all of the issues identified as causing detriment to the functional requirements determined for the service.

In particular, the short-term solution will not allow the service to offer effective browsable indexes of any type, or retrieval by subject; both have been identified as high-priority requirements.

To effect improvements in these areas, a harvestable metadata format which is more complex than OAI_DC is necessary. The standard way of specifying such a format is a Dublin Core application profile. There has been significant recent development activity in this area, including: the formalisation of the structure of an application profile; the formalisation of an application profile for scholarly works (the Scholarly Works application profile, formerly known as the Eprints application profile); and the formalisation of a general bibliographic application profile based on the RDA, the successor to the Anglo-American Cataloguing Rules (not due until mid 2009).

It is possible that the results of this activity will deliver an application profile suitable for IRIScotland.

- Developments in relevant Dublin Core application profiles should be monitored, and appropriate representation of IRIScotland requirements should continue to be made.

If development of the IRIScotland cross-repository service is necessary before the final outcomes of this activity, or as a fall-back should the outcomes not be suitable, and to inform the development of national and international metadata agreements, a specific IRIScotland application profile (IRIS_DC) will be required.

- As a longer-term solution, repositories included in the IRIScotland cross-repository search service should develop and implement the IRIS_DC recommendations given in this document.

Recommendations to JISC

Conflicts between various metadata schemas funded or promoted by JISC are causing problems for institutional repositories and aggregation services.

- JISC should carry out research into interoperability between overlapping metadata schemas for scholarly communication and research resources.

Jump-off pages created in institutional repositories often duplicate metadata made available for harvesting. Harvested metadata are often linked to local metadata rather than the

resource being described. Aggregation services are forced to present end-users with redundant metadata and a chain of hyperlinks to the resource being described. The functionality of jump-off pages required by repositories is likely to be better met using item- and collection-level metadata, but this is not currently accommodated by repository application software.

- JISC should develop recommendations for the appropriate application of item- and collection-level metadata to meet the requirements of institutional repositories.

Background

One of the main objectives of the IRIScotland project is to establish “a harvester-based pilot cross-repository search service to enhance exposure of the Scottish research output as a whole”¹. This was a deliverable for Work Package 5, carried out by the Centre for Digital Library Research (CDLR). In order for this cross-repository service² to work effectively, agreement is needed between participating institutions so that the metadata made available for harvesting from their repositories is as interoperable as possible and meets the functional requirements of the cross-institutional service. This draft agreement was also a deliverable for Work Package 5, and is a deliverable for Work Package 3 of the project extension.

Harvesting involves the automatic copying of local repository metadata using the Open Access Initiative Protocol for Metadata Harvesting or OAI-PMH³. The metadata is then aggregated into a single database to allow information retrieval across records from multiple repositories.

Different institutions use different software packages for managing repositories. Aberdeen, Edinburgh, St Andrews and Stirling universities use DSpace, Queen Margaret and Strathclyde universities use EPrints, and Glasgow University uses both DSpace and EPrints. All these repositories are mediated by the local university library (though none of them is currently cross-searchable with the library's catalogue). The National Library of Scotland is setting up a pilot hosted repository service using Fedora software. There is no common coverage of research output between the repositories; some are confined to particular types of resource such as theses or published journal articles, while others have a broader scope.

All of these repositories can export metadata which conforms to OAI-PMH.

OAI-PMH accommodates multiple metadata formats, but mandates that the repository must offer every record in OAI_DC format⁴ regardless of what other formats may be made available. OAI_DC is essentially restricted to the 15 basic metadata elements of Dublin Core which must not be qualified in any way.

Some repositories also offer records in the UKETD_DC application profile developed by the EThOS project⁵ for UK electronic theses and dissertations, which uses qualifications to Dublin Core elements to enhance the metadata structure.

Repositories usually employ other variant formats or local qualifications of and additions to Dublin Core elements for local purposes; many of these formats are determined by the software package being used. Mappings from local formats to standard formats for harvesting may be lossy, so what is clear in the local repository may become ambiguous in a cross-repository service.

Other standard formats for local repositories and aggregation services are under development. One of particular significance for IRIScotland was the recent EPrints application profile^{6 7}, which is based on Functional Requirements of Bibliographic Records (FRBR)⁸ and

¹ IRIScotland project. Available at: <http://www.iriscotland.lib.ed.ac.uk/>

² IRIScotland project pilot cross-repository service. Available at: <http://cdlr.strath.ac.uk/iriscotland/>

³ Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0. Available at: <http://www.openarchives.org/OAI/openarchivesprotocol.html>

⁴ Johnston, P. XML schema for OAI_DC. Available at: http://www.openarchives.org/OAI/2.0/oai_dc.xsd

⁵ UKETD_DC application profile. Available at: http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=UKETD_DC+Application+Profile

⁶ Allinson, J., Johnston P. & Powell, A. Dublin Core application profile for scholarly works, Ariadne 50, January 2007. Available at: <http://www.ariadne.ac.uk/issue50/allinson-et-al/>

splits the metadata between the four levels of Work (e.g. a concept), Expression (e.g. an edition), Manifestation (e.g. a PDF file), and Item (a single copy). It is of interest to IRIScotland because its scope is similar, although not identical (it does not cover datasets or documentation generated during the research process), and it is compatible with recent developments in Dublin Core such as the Dublin Core abstract model⁹.

Impact on IRIScotland cross-repository service

The pilot service currently only harvests metadata in OAI_DC format.

This is the only format which is common to the extant Scottish repositories, because it is mandated by the protocol.

OAI_DC metadata is confined to only 15 elements, which restricts the ability of the service to meet identified functional requirements. For example, users have indicated that searching and browsing the names of departments and institutions to which the authors of an item belong is a priority; but it is not possible to offer this functionality because there is no distinct metadata element to accommodate the information.

There are no standards for the content of the elements. This again restricts the functionality that can be offered by a cross-repository service. It is not possible to offer any effective browsable indexes because of wide variations in content. For example, author names may be entered in given name-family name order, or vice-versa. A particularly significant problem is in subject retrieval; the variation in terms used for subject topics is so great that even a keyword search is ineffective, leading to many false-drops and missed hits.

Some immediate improvements to the cross-repository service can be made by examining the mapping of richer metadata structure (usually qualified Dublin Core elements) to simple OAI_DC, and encouraging the use of controlled content.

Further improvement requires the development of a common agreement between partner repositories on enriching the structure of the metadata they provide for harvesting, and on controlled content standards such as subject vocabularies and item types. An effective and sustainable approach to achieving this is the development of a community-wide Dublin Core application profile (IRIS_DC) which specifies qualifications to the basic metadata elements and relevant content standards. However, there is a great deal of current activity which may inform, or even deliver, an appropriate application profile for IRIScotland. It was not possible to assess this during the timescale of the original IRIScotland project, not least because the formal specification of a DC application profile¹⁰ was not finalized until the end of 2007.

Methodology

To assist the process of reaching such an agreement, CDLR organized a general seminar to discuss metadata issues for Scottish repositories in collaboration with the Cataloguing and Indexing Group in Scotland and the National Library of Scotland.

This was followed by an IRIScotland project workshop, also held at the National Library of Scotland, on 31 January 2007. Most of the recommendations included in this document emerged from discussions held at that workshop. Participants were: Gordon Dunsire (CIGS/CDLR, Chair), Theo Andrew (Edinburgh University), Alan Dawson (CDLR), Flora Lee (NLS), Les McMorran (Aberdeen University), Neil Nicholson (NLS), William Nixon (Glasgow University) and Alan Stevin (Strathclyde University).

⁷ Eprints application profile. Available at:
http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

⁸ IFLA FRBR Review Group. Functional requirements for bibliographic records - final report. Available at: <http://www.ifla.org/VII/s13/frbr/frbr.htm>

⁹ DCMI abstract model. Available at: <http://dublincore.org/documents/abstract-model/>

¹⁰ The Singapore framework for Dublin Core application profiles. Available at:
<http://dublincore.org/documents/singapore-framework/>

The original version of the draft agreement was published by CDLR in June 2007 for general comment¹¹, and referenced in the report from Work Package 5 of the IRIScotland project¹². It was also submitted to SCURL (Scottish Confederation of University and Research Libraries) for further comment and approval, but SCURL declined to take any action on it at that stage.

Work Package 3 of the project extension has continued to monitor developments affecting research repository metadata.

Work Package 3 has also produced a short paper for the IRIScotland Project Board, to be used as a basis for advocating a metadata agreement. This is given at the end of this document (Draft IRIScotland metadata agreement – what is it good for?).

Scope

The metadata is intended to describe a range of content types for local research output held in institutional repositories in Scotland, including articles, books, book chapters, conference papers, reports, theses, and other materials defined as a scientific or scholarly research text by the Budapest Open Access Initiative¹³, as well as datasets and other resources generated during research. This does not imply that the items described must be available on open access, and the metadata is intended to cover digital items with restricted access and non-digital items which are stored in other repositories within the institution, such as the library, theses collection, etc.

In reaching this agreement, workshop contributors took note of the recent EPrints application profile, which uses a relatively complex data model. As this model will take time to implement in repository management software, it is unlikely to be supported in institutional repositories in the near future. It was therefore agreed that, while the EPrints application profile should inform the IRIScotland metadata agreement, it should not determine it.

The relevance of RDA (Resource description and access)¹⁴ to the metadata agreement was also noted. RDA is designed for all types of material, all formats and all communities, but will not be published until mid 2009.

Developments since the original version of this report (June 2007)

The EPrints application profile has been renamed the Scholarly works application profile (SWAP)¹⁵. The Dublin Core Metadata Initiative has established a Scholarly Communications Community¹⁶ to exchange information and discuss issues related to the use of Dublin Core in describing research papers, texts and other resources created and used in scholarly communication. This group has a specific remit to discuss the use of SWAP. SWAP has been

¹¹ Draft IRIScotland metadata agreement: standards and guidelines for institutional repositories. June 2007. Available at:
<http://cdlr.strath.ac.uk/pubs/dawsona/irismetadatadraft.pdf>

¹² IRIScotland project. Work package 5: report, evaluation and recommendations for the pilot service. September 2007. Available at:
<http://cdlr.strath.ac.uk/pubs/dunsireg/iriswp5recommendations.pdf>

¹³ Budapest open access initiative. Available at: <http://www.soros.org/openaccess/>

¹⁴ Joint Steering Committee for Development of RDA: RDA Resource description and access. Available at: <http://www.collectionscanada.ca/jsc/rda.html>

¹⁵ Scholarly works application profile. Available at:
http://www.ukoln.ac.uk/repositories/digirep/index/Scholarly_Works_Application_Profile

¹⁶ DCMI Scholarly Communications Community. Available at:
<http://dublincore.org/groups/scholar/>

used to test the DCMI description set profile¹⁷. Compliance with the DCMI abstract model, which has been published as a DCMI recommendation, has been confirmed.

Although SWAP is regarded as being complete and stable, the impact of recent developments in FRBR and RDA were presented and discussed at the Community's workshop during the Dublin Core annual conference 2008¹⁸.

The workshop also discussed issues arising from the IRIScotland project, specifically on the topics of compatibility and interoperability between SWAP and similar metadata schemas, and the use of so-called "jump-off" pages in institutional repositories.

Compatibility and interoperability of metadata schemas for research output

At least one of the repositories whose metadata is being harvested for the IRIScotland cross-repository service has experienced points of conflict in attempting to make the metadata compliant with difference schemas.

In particular, this repository and probably others, if not all, within the scope of IRIScotland are committed to membership of the EThOS service¹⁹, which uses the UKETD_DC²⁰. This metadata application profile is intended to describe electronic theses and dissertations, and is therefore narrower in scope than SWAP, or any other metadata schema that might be suitable for IRIScotland. Whether the UKETD_DC application profile is suitable for extension to scholarly works other than theses and dissertations remains unclear.

The UKETD_DC application profile itself identifies a point of conflict in the treatment of citations, commented as 'Citations to previously published sections of this thesis. Applies particularly to "thesis by publication". Where possible, citation information entered should conform to a recognised citation standard.' The profile maps this attribute to two different qualified Dublin Core elements: relation.hasVersion; and identifier.citation. The latter is to be used "for DSpace (for historical reasons)". These qualified elements will map to two distinct simple Dublin Core elements, dc.relation and dc.identifier respectively, unless specific action is taken by the repository system. That is, DSpace systems would have to map dc.identifier.citation to dc.relation to be compliant with Dublin Core. If this is not done, any aggregation service harvesting metadata in OAI_DC format would have to be forewarned that the repository was using DSpace so that it could apply the mapping itself.

But an earlier version of UKETD_DC did not distinguish between citations to other resources (i.e. previously published sections) and the thesis itself. Unless retrospective conversion of metadata created using the earlier version is carried out, aggregation services will not be able to apply the mapping safely.

The UKETD_DC application profile and SWAP are disjunct; that is, each has metadata elements that are not present in the other. They are not therefore directly interoperable. Aggregation services can ensure interoperability by harvesting only dumbed-down OAI_DC (simple DC) records; if they wish to utilize the richer formats, they will have to develop and maintain costly bespoke mappings.

Further problems arise between SWAP and other metadata schema because it is based on FRBR and has a hierarchical structure. Other schema, including UKETD_DC and the native

¹⁷ Description set profiles: a constraint language for Dublin Core application profiles. Available at: <http://dublincore.org/documents/2008/03/31/dc-dsp/>

¹⁸ Dunsire, G. FRBR and SWAP: recent developments and implications. Available at: <http://dc2008.de/wp-content/uploads/2008/09/scholarly-workshop.pdf> (pp16-22)

¹⁹ EThOS: Electronic theses online system. Available at: <http://www.ethos.ac.uk/>

²⁰ UKETD_DC: the metadata core set recommended by EThOS. Available at: http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=UKETD_DC%3A+The+metadata+core+set+recommended+by+EThOS

formats used by DSpace and E-prints, are flat. Some of the issues encountered are described in a blog posting from Warwick University²¹.

The proposals for the original IRIScotland metadata agreement regarding the use of OAI_DC, SWAP, and a separate IRIS_DC, therefore remain valid.

JISC should investigate these issues further, as it has made significant investment in the development of these various metadata schemas.

Jump-off pages

The problem of jump-off pages is described in the IRIScotland metadata proposal for dc.identifier.

Two main functional requirements for these pages have been identified: a single location bringing together different manifestations and copies (in SWAP terms) of a resource – the “jump” function; and institutional branding – the “cover sheet” function.

The first requirement is what SWAP is designed to meet. FRBR and the SWAP metadata structure collate different manifestations and copies hierarchically under a single expression. SWAP is a better way of satisfying this function than the jump-off page because it avoids hard-wiring the resource itself (and thus avoids awkward “What is the resource?” questions), places the maintenance workflow in the proper area (metadata, not data), and allows the requirement to be met across multiple repositories.

The second requirement is more appropriately met by using collection-level metadata. The JISC Information Environment Services Registry application profile²² has a metadata attribute for the URL of the logo of a collection (for example, an institutional repository). This attribute was also added to the Scottish Collections Network (SCONE) metadata schema as a result of the cc-interop project²³, and SCONE metadata records are already used in the IRIScotland cross-repository service. It should be noted that the Dublin Core collections application profile (DCCAP)²⁴ does not include a similar attribute.

Other information commonly found on jump-off pages is also better accommodated in existing item-level or collection-level metadata schemas. For example, there is an attribute for access rights (defining who is allowed to access, what restrictions there are, fees, etc.) in SWAP and DCCAP.

The information given in jump-off pages can be ambiguous. At least one Scottish university repository uses the phrase “Full text not currently available from this archive” to mean that the resource is not stored in the repository itself, rather than indicating whether access is open or not. In many instances the phrase is followed by an active hyperlink leading to the resource. While this is a matter for the repository rather than any aggregation service, it further brings into question the utility of the jump-off page.

It is recommended that further research be conducted into the functionality of jump-off pages. The user experience at local and cross-institutional levels is likely to be improved if they can be avoided.

²¹ SWAP and E-prints structures don't match. Available at: http://blogs.warwick.ac.uk/wrap/entry/swap_and_e-prints/

²² IESR: application profile. Available at: <http://iesr.ac.uk/profile/>

²³ Dunsire, G. 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/CCIEExtendSCONE.pdf>

²⁴ Dublin Core collections application profile. Available at: <http://dublincore.org/groups/collections/collection-application-profile/>

Format of proposals and recommendations

This document proposes the recommended usage of each one of the 15 fields that comprise the Dublin Core metadata standard.

For each field, any relevant issues are summarised, and recommendations are provided for both syntax and content for creating records for harvesting in OAI_DC format, and for developing a specific IRIScotland format (IRIS_DC).

Fields are listed in alphabetical order of element name.

This document is not intended to be an application profile itself, although components may be reusable in a formal application profile.

An example of an OAI_DC record meeting the recommendations is given towards the end of this document.

dc.contributor

Issues

Different types of contributor may need to be distinguished, requiring the use of qualified tags.

Unqualified tags may cause problems for aggregation services which want to omit certain types of contributor, but qualified tags need to be used consistently to ensure correct handling.

dc.contributor is included with ***dc.creator*** in the "author" retrieval focus. The semantic of this focus may become too diffuse or incoherent if it includes supervisors, institutions where research is carried out, etc.

If a local name authority file is in use, the authority number must be replaced by the authorised name for output.

Personal names can be recorded in "display" format of given name followed by family name, or "browse" format of family name followed by given name.

Corporate names can be recorded in inconsistent ways.

Punctuation to separate parts of names may be used inconsistently.

The distinction between contributors and authors (creators) can be blurred.

OAI_DC recommendations

- Only editors and creators of partial contents such as prefaces and illustrations should be mapped to ***dc.contributor***.

IRIS_DC syntax recommendations

- ***dc.contributor*** may be used unqualified.
- The following qualified forms **may** be used if required:
 - dc.contributor.editor***
 - dc.contributor.funder***
 - dc.contributor.illustrator***
 - dc.contributor.institution***
 - dc.contributor.supervisor***
- ***dc.contributor.advisor*** should be mapped to ***dc.contributor.supervisor***; this makes the qualifier label compatible with the Scholarly works application profile .
- ***dc.contributor.author*** should be mapped to ***dc.creator***.
- ***dc.contributor.other*** should be mapped to ***dc.contributor***.

IRIS_DC usage recommendations

- ***dc.contributor*** (and qualified forms) is repeatable.

IRIS_DC content recommendations

- Any guidance offered by the Scholarly works application profile should be adopted.

Organisation name

- Each institution should choose one form of its name and use it consistently in the ***dc.contributor.institution*** field, avoiding variations.

e.g. *University of Strathclyde*

(**not** *The University of Strathclyde* or *Strathclyde University*)

- If department names are included, this should be done in a consistent manner by including the department name after the institution name.

e.g. *University of Strathclyde. Department of Law*

(**not** *Department of Law, Strathclyde University*)

dc.coverage

Issues

The relationship between ***dc.coverage*** and ***dc.subject*** can be confused. The temporal or spatial topic of the resource should be recorded in ***dc.coverage***, not ***dc.subject***.

It is not possible to link the content of ***dc.coverage*** to ***dc.subject***, when there is more than one topic and corresponding instance. If the topic of a resource is a subject qualified by a place, e.g. *Birds of Orkney*, then ***dc.coverage=Orkney*** and ***dc.subject=Birds***; if this is the only topic, then there is an implicit link and the correct complete topic can be inferred. With two such topics, e.g. *Birds of Orkney and North Sea fish*, then ***dc.coverage=Orkney***, ***dc.coverage=North Sea***, ***dc.subject=Birds***, and ***dc.subject=Fish***; the complete topic could be erroneously taken to be *Fish of Orkney and North Sea birds*, or *Birds and fish of Orkney and the North Sea*, etc.

It is not possible to provide separate searchable or browsable indexes for spatial and temporal subjects unless the content vocabularies are chosen to allow machine parsing; e.g. all spatial content in spelled-out textual form and all temporal content in numerical form.

The content of a ***dc.coverage*** field may be ignored by some metadata aggregation services, so its general value is questionable.

The provision of a specific focus for spatial topics is likely to be useful in an information retrieval service which is geographically scoped.

The provision of categories of subject such as general topic, spatial topic and temporal topic enhances retrieval precision.

OAI_DC recommendations

- ***dc.coverage*** should be used for record display and general keyword searches only.

IRIS_DC syntax recommendations

- ***dc.coverage*** should always be qualified, with one of the two standard qualifiers from the original Dublin Core specification, i.e. ***dc.coverage.spatial*** or ***dc.coverage.temporal***.

IRIS_DC usage recommendations

- ***dc.coverage.spatial*** and ***dc.coverage.temporal*** are repeatable.

IRIS_DC content recommendations

- Only places or time periods should be used, to modify the general subject coverage of the resource.

e.g. *France*

18th century

dc.creator

Issues

Identifying the citation order of authors can be difficult if the item contains variations.

If a local name authority file is in use, the authority number must be replaced by the authorised name for output.

Personal names can be recorded in "display" format of given name followed by family name, or "browse" format of family name followed by given name.

Corporate names can be recorded in inconsistent ways.

Punctuation to separate parts of names may be used inconsistently.

The distinction between authors (creators) and other contributors can be blurred.

OAI_DC recommendations

- ***dc.contributor.author*** must be mapped to ***dc.creator***.
- Unqualified or other qualified content of ***dc.contributor*** should not be mapped to ***dc.creator***.

IRIS_DC syntax recommendations

- ***dc.creator*** should be used unqualified for the main author or authors or other creators of the resource.
- ***dc.contributor.author*** should be avoided. However, in view of previous usage, IRIScotland will aim to handle ***dc.contributor.author*** as ***dc.creator***.

IRIS_DC usage recommendations

- ***dc.creator*** is repeatable.

IRIS_DC content recommendations

- Any guidance offered by the Scholarly works application profile should be adopted.

Ordering

- Names should be given in "browse" format, i.e. surname first.
e.g. *Law, Derek*
(not *Derek Law*)

Multiple authors

- A separate field should be used for each creator.
e.g. *Jones, Richard*
Andrew, Theo
(not *Jones, Richard and Andrew, Theo*)

Authority files

The issue of authority files for personal names is out of scope for this agreement.

dc.date

Issues

This field is useful for retrieval and results ordering if content format is consistent.

Different types of date cannot be distinguished.

Certain types of resource may have multiple associated dates; e.g. date of creation, date of submission, date of publication, date of last amendment.

Multiple instances of some types of qualified date can enhance resource discovery. For example, repeats of ***dc.date.revised*** can be cited in chronological order of the content to give a logical sequence, but multiple instances of ***dc.date.embargoed*** will be confusing to users.

There can be significant variation in the way that a date is recorded; e.g. *5 Jan 2007*; *5th January 2007*; *January 5th 2007*; *5/1/07*; *1/5/2007*; etc. The last two forms will be misinterpreted by users in different countries.

OAI_DC recommendations

- Only the publication date, or nearest equivalent for the type of resource, should be mapped to ***dc.date***.

IRIS_DC syntax recommendations

- ***dc.date*** may be used unqualified if the content refers to the date of publication of the resource, which is the most useful date for users (i.e. the same meaning as ***dc.date.issued***).
- Qualified forms of ***dc.date*** should only be used if their local meaning is clearly defined and if they are required for a specific purpose.

IRIS_DC usage recommendations

- ***dc.date*** is not repeatable.
- Qualified forms of ***dc.date*** may be repeatable, depending on local requirements, but the semantic of the qualifier must clearly distinguish the relationship between the repeated dates.

IRIS_DC content recommendations

- For most purposes the year of publication is sufficient.
e.g. *2007*
- If a more precise date is preferred, then the ISO 8601²⁵ standard format yyyy-mm-dd should be used.
e.g. *2007-03-10*
- Date ranges should not be used.
- Database administration information should be stored elsewhere in the database, not in ***dc.date*** fields.

²⁵ International Organization for Standardization. ISO 8601: Numeric representation of dates and time. Available at: <http://www.iso.org/iso/en/prods-services/popstds/datesandtime.html>

dc.description

Issues

This is a free-text field and little content standardisation can be imposed.

Some content may be inserted using cut-and-paste methods from the resource being described, resulting in the inclusion of non-standard characters such as smart quotes and long dashes which can cause XML syntax errors or unwanted display artefacts.

Internal formatting such as bullet points and multiple paragraphs cannot be properly harvested.

OAI_DC recommendations

Mapping of qualified elements can cause severe problems, especially if the semantic of the qualifier is not reflected in the content. E.g. ***dc.description.sponsorship=British Telecom*** is mapped to ***dc.description=British Telecom***, which is ambiguous; better would be ***dc.description.sponsorship=Sponsored by British Telecom***.

- The semantic of a qualifier should be mapped to a standard term or phrase preceding the content of the unqualified ***dc.description***.
- If this is not possible, qualified elements should **not** be mapped.

IRIS_DC syntax recommendations

dc.description should be used unqualified, and only for the abstract or summary of the resource.

The following qualified forms suggested by DSpace should **not** be mapped:

dc.description.abstract (use ***dc.description*** instead)
dc.description.provenance
dc.description.sponsorship (use ***dc.contributor.funder*** instead)
dc.description.statementofresponsibility
dc.description.uri

IRIS_DC usage recommendations

- ***dc.description*** is not repeatable.

IRIS_DC content recommendations

- A description should be included wherever possible. If no description is available, the ***dc.description*** field should be **omitted**, in preference to “No abstract provided” or similar text.
- The description should not begin with the word “Abstract” or similar introductory text.
- Any further guidance offered by the Scholarly works application profile should be adopted.

Paragraphs

- Descriptions should usually be limited to a single paragraph, of any length. No formatting should be included, so any bullet points or long dashes, e.g. in structured abstracts, should be changed to dashes or numbered lists.

e.g. 1) *First thing*, 2) *Second thing*, 3) *Third thing*

(not

First thing

Second thing

Third thing)

- If two different descriptions are available, the longer one should be used.

Non-standard characters

- Where possible, only characters that can be typed on a standard keyboard should be used in the description, as in the **dc.title** field. Particular care should be taken if copying text from other sources, such as Word documents or PDF files, as these often include smart quotes, long dashes, bullet points and paragraph marks, which all need to be converted to plain text or removed. In PDF documents, “ff” and “fi” are often represented by non-standard characters, which should be changed to ordinary keyboard letters.

dc.format

Issues

This field is used for display only, not retrieval. Its value for users is unclear.

There is potential confusion with ***dc.type***.

OAI_DC recommendations

- ***dc.format*** should be output from qualified and unqualified fields only if it refers to the physical characteristics of the resource, such as file encoding format (e.g. *PDF*), and the information is not intrinsic to another field (e.g. the URI in ***dc.identifier***).

There will be general user expectation that all items described will be in a digital format. It is important that information about non-digital formats is made available.

- ***dc.format*** should be output if the item is not in a digital format.

IRIS_DC syntax recommendations

- ***dc.format*** may be used, in qualified or unqualified form, if the field has some clear value. It is more likely to be useful for non-text resources, and may be omitted for text resources.

IRIS_DC usage recommendations

- ***dc.format*** is repeatable.

IRIS_DC content recommendations

- Any guidance offered by the Scholarly works application profile should be adopted.

dc.identifier

Issues

There is some confusion about the meaning and purpose of this field, which is intended to be machine readable, and differing practices are in use in different repositories.

DSpace proposes several possible qualifiers. At Strathclyde University a further field has been added to the EPrints database to record the departmental database identifier (where relevant), to assist with de-duplication and item tracking.

DSpace and other repository management systems automatically populate *dc.identifier* with the local repository "jump-off" page or "cover sheet" for the resource. These are essential local metadata records repeating much of the content already seen by the user of the cross-repository search service for harvested records. The user then has to use the URI from the local metadata to get access to the item. This may confuse and frustrate users of the aggregated metadata who cannot get direct access to the item and see the same metadata displayed twice during the retrieval process.

Multiple resource identifiers can be confusing for the user.

It can be very difficult for an aggregation service to determine which identifier can be used to link to the resource from its metadata.

OAI_DC recommendations

- Only the URI of the resource being described should be included in ***dc.identifier*** which should be used unqualified.

IRIS_DC syntax recommendations

- ***dc.identifier*** should be used unqualified.

IRIS_DC usage recommendations

- ***dc.identifier*** is not repeatable.

IRIS_DC content recommendations

- This field should be used to record the URI of the content of resource, and for no other purpose, to ensure reliable linking between the metadata record and the item itself.
- An ISSN is an identifier but is associated with a publication rather than an article, and therefore if it is necessary to record it then it should be included in the ***dc.relation*** field and associated with the journal title. The same principle applies to the ISBN for book chapters, where the ISBN applies to the entire book and should therefore be associated with the book title in the ***dc.relation*** field.

dc.language

Issues

This field is used for display only, not retrieval.

Multiple languages cannot be linked to other metadata elements. If a resource is written in English, but with an introduction in French, then ***dc.language=fre*** is ambiguous because it is not possible to determine whether the whole resource, or only part, is in French.

OAI_DC recommendations

- ***dc.language*** should record the language of the main part of the resource if it is not English.

IRIS_DC syntax recommendations

- ***dc.language*** should record the language of the main part of the resource if it is not English.

IRIS_DC usage recommendations

- ***dc.language*** may be used if it has some value, but may be omitted for English language resources.
- ***dc.language*** is not repeatable.

IRIS_DC content recommendations

- The 3-character language code should be used as defined by ISO 639-2²⁶.
e.g. *eng*
(not *English*)

²⁶ Library of Congress Network Development and MARC Standards Office: ISO 639-2 codes for the representation of names of languages. Available at:
<http://www.loc.gov/standards/iso639-2/>

dc.publisher

Issues

For many items in repositories it is not always clear who the publisher is. While paper journals have a clear publisher, digital versions of articles in repositories do not.

Digital versions of the resource may differ slightly from printed versions, and may have been published when the author was employed at a different institution.

OAI_DC recommendations

- ***dc.publisher*** may be used unqualified or may be omitted.

IRIS_DC syntax recommendations

- ***dc.publisher*** may be used unqualified or may be omitted.

IRIS_DC usage recommendations

- ***dc.publisher*** is repeatable.

IRIS_DC content recommendations

- It is acceptable to include the name of the publisher of the relevant printed journal or book in ***dc.publisher***, although this has limited value.
- The title of a journal, book or conference proceedings in which an item is published should be included in ***dc.relation*** not ***dc.publisher***.
- The name of the repository provider, e.g. *University of Glasgow*, should be included in ***dc.contributor.institution*** and not in ***dc.publisher***, except for theses or other items where the local institution is clearly the sole publisher.

dc.relation

Issues

The purpose and value of this field is often unclear. DSpace proposes a large and confusing number of qualifiers.

OAI_DC recommendations

- Qualified ***dc.relation*** fields should only be mapped when there is a direct relationship between the content and the resource being described.

IRIS_DC syntax recommendations

- ***dc.relation*** may be used unqualified, but then the content of the field should make the nature of the relationship clear by using a standard introductory phrase, e.g. "*Published in ...*".
- The only qualified form of relation to be recommended is ***dc.relation.ispartof*** to record publication details for a journal article, paper or book chapter.

IRIS_DC usage recommendations

- ***dc.relation*** is repeatable.

IRIS_DC content recommendations

- The primary use of ***dc.relation*** should be for recording the publication title, volume and issue number for journal articles and conference papers.
E.g. *Published in Library Review, 54(1)* (if unqualified ***dc.relation*** used)
Library Review, 54(1) (if ***dc.relation.ispartof*** used)
- If for any reason ***dc.relation*** can not be used to record this information then it may be added to the end of the ***dc.description*** field.
- If a full article citation is required, it should be included in an additional ***dc.relation*** field, **not** in ***dc.identifier***.

dc.rights

Issues

This field is used for display only, not retrieval.

At Glasgow University a further field has been added to record additional rights information, and a cover sheet has been added to each item to ensure that rights information is prominently displayed to repository users.

OAI_DC recommendations

- ***dc.rights*** should contain only information about access and usage conditions for the item.

IRIS_DC syntax recommendations

- ***dc.rights*** should be used unqualified, if at all.

IRIS_DC usage recommendations

- ***dc.rights*** is repeatable.

IRIS_DC content recommendations

This is a free text field that may be used to include a copyright statement and may also include statement of permissions.

- The word 'copyright' or (c) should be used, **not** the copyright symbol.

dc.source

Issues

This field is used for display only not retrieval. Its use and value is unclear.

OAI_DC recommendations

- *dc.source* should be **not** be used.

IRIS_DC syntax recommendations

- *dc.source* should be **not** be used.

IRIS_DC usage recommendations

- *dc.source* should be **not** be used.

IRIS_DC content recommendations

- If information about the source of an item is important then it should be included in the *dc.description* field.

dc.subject

Issues

There is no single subject scheme suitable for common use between local repositories because of local interoperability, legacy and workflow issues.

Mapping between subject schemes is a difficult and time-consuming process.

Most repositories use either uncontrolled keywords or a broad but shallow scheme such as the top two or three levels of Library of Congress Classification.

Harvested records will contain a mixture of controlled and uncontrolled subject terms and classification numbers, which is not suitable for consistent or coherent browsing or searching.

Information about the controlled subject vocabulary or classification scheme used is required if future terminology services and tools are to be employed to improve browsing and searching.

OAI_DC recommendations

- ***dc.subject*** should be used for record display and general keyword searches only.

IRIS_DC syntax recommendations

- ***dc.subject*** should be qualified with the standard qualifier for the subject vocabulary or classification scheme used.
e.g. ***dc.subject.lcsh=armor***, ***dc.subject.ddc=025.04***
- ***dc.subject*** should be used unqualified if there is no control over the terms used, or if the subject vocabulary or classification is local.

IRIS_DC usage recommendations

- ***dc.subject*** (and qualifications) is repeatable.

IRIS_DC content recommendations

A comprehensive and satisfactory solution to subject interoperability is out of scope for this agreement. However, any controlled vocabulary is better than uncontrolled keywords, and should be used if possible, even if a formal subject scheme is not in use. Internal consistency is important and potentially useful, even if interoperability is limited.

- The names of people and corporate bodies or titles of projects and documents are not usually accommodated in standard subject topic and classification schemes, and should not be inserted into a standard qualification of ***dc.subject***.
E.g. do not use ***dc.subject.lcsh=IRIScotland***
Instead, special treatment is required, such as the use of a name authority file which is also applied to ***dc.creator*** and ***dc.contributor***.
- Time periods and geographic scope (places) should be included in ***dc.coverage*** fields, not appended to subject terms in ***dc.subject***.

dc.title

Issues

Identifying the title of an item can be difficult if there are variations.

It is not possible to distinguish the "proper" title from variants if more than one ***dc.title*** field is used.

The inclusion of leading definite and indefinite articles can be inconsistent.

Where a leading article in a title affects the ordering of items for retrieval, it should be up to the harvesting and information retrieval services to handle.

The inclusion of subtitles can be inconsistent.

Capitalisation of words in the title can be inconsistent.

The inclusion or transcription of non-standard characters may violate XML standards.

OAI_DC recommendations

- Only a single ***dc.title*** should be mapped, and it should be the main title of the resource.

IRIS_DC syntax recommendations

- ***dc.title*** should be used unqualified for the main title of the resource, including any subtitle.
- ***dc.title.alternative*** may be used sparingly if regarded as essential.

IRIS_DC usage recommendations

- ***dc.title*** is not repeatable.
- ***dc.title.alternative*** is repeatable.

IRIS_DC content recommendations

- Any guidance offered by the Scholarly works application profile should be adopted.

Subtitles

- Append any subtitle to the main title, after a space, colon, space.
e.g. *Too many houses for a home : narrating the house in the Chinese diaspora*

Alternative title

- If considered essential, for example to record an abbreviated version of a title, ***dc.title.alternative*** should be used for alternative titles. The use of ***dc.title*** unqualified should be avoided for alternative titles.

Leading articles

- Keep leading definite and indefinite articles in the title as it appears on the item being described.
e.g. *The Mandibular canal of the edentulous jaw*
(not *Mandibular canal of the edentulous jaw*)

Capitalisation

- Follow the recommendations of AACR2 and RDA; i.e. capitalise only proper names and the first word in the title (other than definite or indefinite articles).
e.g. *Performance monitoring in digital library systems*
(**not** *Performance Monitoring in Digital Library Systems*)

Punctuation

- Titles should **not** be included in quotes and should **not** end with a full stop.

Non-standard characters

- Where possible, only ASCII characters that can be typed on a standard keyboard should be used in the title and other fields. For example:

Use	not
ae	æ
... (3 full-stops)	... (ellipsis)
- (hyphen)	— (long hyphen)
' (apostrophe)	' (smart single quote)
" (quote)	" (smart double quote)
cafe (without accent)	café (with accent)
(c)	©

Foreign language texts

- The title should be included as shown on the item.
e.g. *Mein kampf*
- Any translations of a resource should be described with separate metadata records.

dc.type

Issues

This field is potentially useful as a means of limiting searches and filtering results, but requires a small controlled vocabulary in order to be effective.

The standard types mentioned in the original Dublin Core standard are too general to be useful.

EPrints software recognises a standard set of item types, which are in use in some repositories, while Aberdeen University uses a similar but slightly different set of values.

It is not feasible to use ***dc.type*** for limiting and filtering searches because of variation in local terms.

OAI_DC recommendations

- A small controlled vocabulary of item types should be used by each repository for consistency in record displays. If the number of variants is small, the harvesting service may be able to map variants to a standard set of terms.

IRIS_DC syntax recommendations

- ***dc.type*** should always be used unqualified.

IRIS_DC usage recommendations

- ***dc.type*** is repeatable.

IRIS_DC content recommendations

- A small set of distinct item types should be used consistently. Labels from the Eprints type vocabulary encoding scheme²⁷ from the Scholarly works application profile are recommended, as summarised below.

Book
Book Item
Book Review
Conference Item
Conference Paper
Conference Poster
Journal Article
Journal Item
News Item
Patent
Report
Scholarly Text
Submitted Journal Article

²⁷ Eprints type vocabulary encoding scheme. Available at:
http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Type_Vocabulary_Encoding_Scheme

Thesis or Dissertation

Working or Discussion Paper

This set of values has a narrower scope than IRIScotland, and may require augmentation, for example for research data.

As long as values are used consistently in a single repository, it should be possible for aggregating services to map them to these recommended values for item retrieval if necessary.

Example metadata record in OAI_DC format

This example record is shown in OAI_DC format (4), with all the above syntax and content recommendations applied in the <metadata> section indented below. The ordering of fields within the metadata section of the record is not significant, apart from the creator field, where the ordering should be consistent with the ordering of authors of the item itself.

```
<?xml version="1.0" encoding="UTF-8"?>
<record>
<header>
<identifier>oai:eprints.gla.ac.uk:3098</identifier>
<timestamp>2006-10-02</timestamp>
<metadata>
<oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:contributor>University of Glasgow. Faculty of Veterinary
  Medicine</dc:contributor>
  <dc:creator>Babamahmoodi, F.</dc:creator>
  <dc:creator>Aghabarari, F.</dc:creator>
  <dc:creator>Arjmand, A.</dc:creator>
  <dc:creator>Ashrafi, G.H.</dc:creator>
  <dc:date>2006-10-01</dc:date>
  <dc:description>Anthrax is an acute bacterial infection caused by Bacillus
  anthracis. Humans become infected under natural conditions by contact with
  infected animals or contaminated animal products. About 95% of human
  anthrax is cutaneous and 5% respiratory. Gastrointestinal anthrax is very
  rare, and has been reported in less than 1% of all cases. Anthrax
  meningitis is a rare complication of any of the other three forms of
  disease. We report three rare cases of anthrax (gastrointestinal,
  oropharyngeal and meningitis) arising from the same source. The three
  patients were from a single family and were admitted with different
  clinical pictures after the ingestion of half-cooked meat from a sick
  sheep. These cases emphasize the need for awareness of anthrax in the
  differential diagnosis in areas where the disease remains
  endemic.</dc:description>
  <dc:identifier>http://eprints.gla.ac.uk/3098/</dc:identifier>
  <dc:publisher>Elsevier Science</dc:publisher>
  <dc:relation>Published in Journal of Infection 53(4):e175-
  e179</dc:relation>
  <dc:rights>Copyright (c) 2006 Elsevier Science</dc:rights>
  <dc:rights>Reproduced in accordance with the copyright policy of the
  publisher</dc:rights>
  <dc:subject>Microbiology</dc:subject>
  <dc:subject>Veterinary medicine</dc:subject>
  <dc:title>Three rare cases of anthrax arising from the same
  source</dc:title>
  <dc:type>Journal Article</dc:type>
</oai_dc:dc>
</metadata>
</record>
```

In this example, the **dc.coverage**, **dc.format**, **dc.language** and **dc.source** fields are not required.

Other cataloguing and data entry issues

Many of the metadata management issues and interoperability problems identified in Work Package 5 arise from data entry procedures and the copying and pasting of text rather than the absence of agreed metadata standards. For example, the occurrence of leading spaces, smart quotes, long dashes and other non-standard characters in titles, descriptions and other fields can cause problems for searching and retrieval, yet are easily avoided in most cases. These issues have already been summarised and published elsewhere²⁸.

A model set of guidelines which can easily be followed by metadata creators and which will alleviate many of the problems is appended to this document.

²⁸ Dawson, A. Thirty problems for subject interoperability (and a few possible solutions), presentation to Electric Connections 2006. Available at: <http://cdlr.strath.ac.uk/pubs/dawsona/ad200602.htm>

Model guidelines for creating metadata in institutional repositories

These guidelines offer simple, easy-to-use methods for improving the consistency, coherency and usability of metadata created for institutional repositories.

They do not address all of the issues; only those which are easily avoidable.

The guidelines are intended as a model for individual institutional repositories, and can be adapted to suit local requirements:

- ❖ Check that the content has been entered accurately before it is submitted to ensure that spelling errors, accidental keystrokes or other typos are corrected.
- ❖ Check that any punctuation inserted into the content follows standard practice. For example, a title and subtitle are concatenated in the title field using space-colon-space.
- ❖ Check that punctuation is encoded in the basic keyboard character set. Non-"standard" punctuation may be inadvertently pasted into the content field, so any cut-and-paste operations involving punctuation should be checked.
- ❖ Specifically, make sure all hyphens are "short"; do not use the long hyphen that is often employed in North America.
- ❖ Specifically, do not enter "smart" quotes pasted from a word-processed document. The safest practice is to retype quotes and apostrophes, overwriting what has been pasted.
- ❖ Avoid special characters that are not available in the basic keyboard character set. Special characters inadvertently pasted into the content field may subsequently display as gobbledy-gook.
- ❖ Specifically, enter or overwrite the copyright symbol © with (c).
- ❖ Specifically, enter or overwrite diphthongs as separate letters; e.g. ae instead of æ.
- ❖ Do not use quotation marks in titles and abstracts unless absolutely essential to identify an actual quotation. If they must be used it is good practice to use double quotes for quotations, rather than single quotes which can be confused with apostrophes. Single quotes should only be used if embedded with a set of double quotes (i.e. when there is a quotation with a quotation).
- ❖ Ensure that only single spaces are used. Double spaces should be closed-up.
- ❖ Unless otherwise instructed, capitalise only the first word of content which is a sentence or phrase, and any proper nouns. Do not use "title-case" or all upper-case for content.
- ❖ Enter titles as they appear on the item being described, except for case and punctuation as instructed above. Do not remove leading definite ("The") or indefinite ("A", "An") articles. Do not add leading articles if they do not appear on the item.
- ❖ Unless otherwise instructed, enter multiple instances of metadata elements in separate fields. For example, enter only one subject topic or classification in each "Subject" field, repeating as necessary; do not enter more than one topic or classification in the same field.

Draft IRIScotland metadata agreement – what is it good for?

Without an agreement

As outlined in the draft agreement, without an IRIScotland-wide agreement on metadata the quality of resulting services (not confined to the existing cross-repository search service) will continue to be poor. Results from searches will be inconsistent and sometimes incoherent, to the detriment of any kind of "showcase" for Scotland's research output. Metadata harvested by other services will also result in a poor impression of Scotland's activities in these areas.

Budgets

Retroconversion costs

Conversion of existing metadata to meet stated standards compliance at any current or future level incurs significant costs.

- Machine conversion requires pattern analysis of metadata content, mapping, programming, and testing.
- Manual conversion requires documentation, training and scheduling.
 - Manual conversion incurs a per-item cost associated with checking and editing.

It is highly likely that institutions will have to use both machine and manual conversion methods to ensure that the quality of the metadata is maintained.

Conversion costs will therefore rise as the number of metadata records (items) requiring conversion grows.

Early adoption of the IRIS metadata agreement reduces the total cost to the institution.

System costs

Conversion and development of system components which depend on metadata (specifically user interfaces, information retrieval functions, and database) to meet stated standards compliance at any current or future level also incur significant costs.

- Cost may be indirect where third-party and consortium suppliers are involved.

System costs can be reduced by regular maintenance and development of metadata standards to ensure their continuing compatibility with international technical developments.

- Indirect system costs can also be reduced by multiple institutions interacting collectively, rather than individually, with specific suppliers.

The IRIS metadata agreement will seek to influence, as well as monitor, appropriate system developments at international level, on behalf of all institutions using it.

Local service costs

Most institutional repositories expect a wide range of staff and students to interact directly with interfaces for the creation of metadata, as well as subsequent information retrieval.

- This approach has been adopted by the NLS hosted repository service.

Maintaining the stability and familiarity of these interfaces encourages uptake of repository facilities.

Most repositories are implemented with metadata for resources limited to the output of a few departments and one or two types such as theses or conference papers.

- Careful planning is required to ensure that any metadata standards formulated at an early stage are suitable in practice when the repository subsequently grows to true institutional scope.

A well-planned and robust metadata agreement ensures that changes to repository user interfaces are driven by improvements for functional, rather than data, reasons.

Cultures and environments

The promotion of any open-access, Web-based information resource takes place in a highly competitive environment. The more places and contexts from which a resource can be discovered, the more likely it is to be used.

The underlying technology of institutional repositories is designed to maximise the re-use of metadata in diverse contexts through so-called aggregation services.

- The IRIScotland pilot cross-repository search service is an example.
- But all institutional resources in IRIScotland can also be accessed from other services with different contexts such as OAISter (US) and (eventually) DRIVER (Europe).

Metadata which does not comply with the requirements of these services is likely to be excluded at item (record) or institutional level, thus reducing exposure of the institution's output.

- If there is no subject metadata, the item will not be found in the subject index.
- If the subject metadata is of poor quality, it will be excluded from the subject index.
- If the aggregation service is subject-based, the institutional will have no representation in it.

This may result in a de-facto compliance level regime.

- For example, all of an institution's output might be found in a general keyword search only
- and additionally in specific keyword searches such as title or subject
- and additionally in browse searches such as title or subject lists.

The value of an open-access institutional repository is significantly reduced if metadata quality does not support its inclusion in aggregation services operating in contexts of importance to the institution and its members.

Compliance levels for IRIS?

There are two sets of compliance to consider: technical standards compliance, and IRIScotland cross-search service compliance. The former is necessary for the latter.

The draft IRIScotland metadata agreement recommends that metadata complies with the OAI_DC standard. This is mandated for the OAI-PMH standard which every aggregation service uses.

- Without compliance at this level, any development of the IRIScotland service will be bespoke; that is, of no benefit to any other aggregation service and relatively expensive.
- It is really in the best interests of each institution that compliance with the basic technical standards is ensured, outwith any consideration of IRIScotland.

Further compliance levels for IRIScotland might be developed along the lines mentioned, with levels for general and specific keyword searches, and specific browse lists.

- The existing subject keyword search is not effective, and could be improved immediately by excluding some institutions' records from it.

Subject searching is the most difficult area to resolve. It is best tackled through the adoption of a single, recognised scheme such as Library of Congress Subject Headings, but experience shows this is difficult to achieve (but perhaps Scotland has a chance?). Reasons include intellectual disagreement, unfamiliarity at some local institutions, and additional local costs of training and quality assurance.

- If institutions were not willing to adopt a single subject scheme, then compliance with any one of the recognised schemes could still result in an effective service through interaction with the HILT terminologies service which is designed to improve subject searching in a multi-scheme environment (and is Made in Scotland).

Beyond the basic technical compliance given in the draft metadata agreement, the development of compliance levels is closely connected to functional development of the IRIScotland cross-repository search service; they can be mutually dependent.

The draft does not currently provide for anything other than basic technical compliance. The IRIS_DC recommendation is intended as a starting-point for wider discussion which should involve better understanding of user needs and the functional requirements of a full service, and it may not be needed at all if a suitable metadata agreement becomes available from a wider community.

- The IRIScotland extension will provide an update of the draft metadata agreement in these contexts.

The current draft metadata agreement only discusses compliance at a basic level. If agreement cannot be found at that level, then there is little point in pursuing a cross-institutional agreement.

If agreement can be found at that basic level, further levels of compliance can be developed for existing repositories.

New repositories should be able to benefit from the full metadata guidelines updated during the extension.

The Library angle

Most university and research libraries have considerable experience in dealing with these issues, in the context of metadata for print-based resources, and should be able to offer advice to their institutions.

Gordon Dunsire

28 Apr 2008