UK Metadata Guidelines for Open Access Repositories

Phase 1: Core Metadata, March 2013

Version 0.9 2013-03-13

Summary

- These UK Metadata Guidelines for open access repositories and the associated Application Profile have been developed by UKLON [University of Bath] at the request of JISC and RCUK. They comply as closely as possible with the OpenAIRE Guidelines and EThOS.

- Both JISC and RCUK strongly endorse these Guidelines and expect their usage to be widely adopted and supported by higher education and research institutions in the UK. To this effect a letter from RCUK to Vice Chancellors has been sent.

- The development of these Guidelines was based on a central use case: the ability to track research outputs across systems. More specifically, RCUK requires the means to monitor compliance with its open access policies.

- There is a particular focus on two new metadata elements: project ID (or grant/award number) and funder name. This information is not routinely exposed in institutional repositories at present. Collecting and exposing this information is a key requirement of the new Guidelines.

- As well as introducing new metadata elements, it is hoped that the introduction of these Guidelines will help standardise peoples' interpretation of common metadata elements. Analysis of RepUK, one of the UK's metadata aggregations, indicates that consistency is lacking with respect to how different institutions interpret metadata requirements.

- JISC and RCUK anticipate that UK institutions will begin the process of adopting and complying with the Guidelines as soon as possible. A statement from the sponsors outlining their expectations with respect to the speed of adoption when the Guidelines are launched in mid-April.

- To help with the compliance process, a plugin for EPrints repositories (versions 3.3.x) and a patch for DSpace repositories (version 1.8.2; version 3.x at a later stage) will be freely available.

- Questions about these Guidelines may be addressed to: admin@rioxx.net Further information can be found on the following website: rioxx.net

1. Introduction

The successful development of open access repositories in very many of the UK's higher education and research institutions is testament to the efforts of repository managers and their information management colleagues. The result of these efforts is a growing body of research information that can be freely discovered and re-used by people around the world. The foundations of the UK's repository infrastructure are firmly established but there remain opportunities for the community to build and improve services that provide additional value to a variety of stakeholders in the research communication chain.

Those involved in the collection and management of information understand the central role played by metadata in the success of their institutional information management systems. Accurate, rich, high quality metadata enhances not only discoverability and re-usability but also the extent to which different stakeholders can use outputs for different purposes.

Analysis of the UK's aggregations of metadata collected from open access repositories indicate that, at present, there are inconsistencies in the ways in which metadata is managed. For example, a recent snapshot of the content of UK open access repositories in the tertiary sector clearly showed a significant disparity between the number of full-text pdfs indicated by metadata and the actual number of actionable pdfs in those repositories. The development of national guidelines for the management of metadata specifically for open access repositories aims to reduce ambiguity regarding the implementation of metadata standards and improve the overall quality and consistency of metadata. These guidelines target publications metadata specifically.

The key impetus for the development of these national guidelines is the government-driven need for Research Councils to be able to identify the research outputs from projects they have funded. At present there is no straightforward or systematic way for these funders to identify when relevant articles appear in open access repositories. The introduction of two new core metadata fields is designed to address this particular problem, namely a field describing a project's identity – such as a grant number - and a field describing the identity of the funder. This information is not routinely collected in open access repositories at present.

Updates

While this first iteration of national metadata guidelines for open repositories focuses on these two new fields in addition to the familiar bibliographic metadata, work is under way to achieve consensus on a common vocabulary to describe the open access status of different digital objects, their licensed status and any applicable embargo conditions. As the number of open access full text items grows, the need for common metadata standards assumes a new level of importance. It is important to note, therefore, that these guidelines will be augmented in a second phase in coming months.

Helping with compliance

The widespread adoption and implementation of these metadata guidelines in the UK is being strongly advocated and supported by the Research Councils and the JISC. A software plugin for EPrints (version 3.3.x) and a patch for DSpace (version 1.8.2 with a patch for version 3.x to follow later) will help automate the compliance process but the important work collecting the additional metadata will ultimately depend upon the goodwill and expertise of information and repository managers around the nation.

Timelines

JISC and the RCUK anticipate that higher education and research institutions will adopt these Guidelines as soon as possible. There is currently no expectation that legacy metadata will be made compliant with the Guidelines, but that new records being ingested by repositories will be compliant. When the Application Profile and these Guidelines are officially launched in mid-April there will be a statement about the sponsors' expectations for adoption; the timeline has not been finalised at the time of writing this draft.

2. Phase 1 Metadata Muidelines

2.1 Standards

A foundation of well-established and widely-adopted technical standards is essential for the interoperation of systems. In the sphere of open access repositories in the UK the two most relevant standards for the time being are Dublin Core and OAI-PMH. A brief overview of these standards is presented below for readers who are not very familiar with them.

2.1.1 Dublin Core

Devised in Dublin (Ohio, USA) in the mid-1990s, the Simple Dublin Core Metadata Element Set includes fifteen core elements for the purpose of describing electronic resources in straightforward terms. Under the direction of the Dublin Core Metadata Initiative, an organisation set up to further the development of metadata standards, refinements to the original set of elements have been introduced to enhance specificity. In addition, Qualified Dublin Core helps with the interpretation of elements primarily through the use of encoding schemes which include globally recognized unique identifiers such as ISSNs and controlled vocabularies like MeSH.

2.1.2 OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting was launched in 2001 and underpins many of the services that harvest metadata from open access repositories. To facilitate harvesting repositories are recommended to support oai_dc, the simple Dublin Core record format as defined by the OAI-PMH DC XML schema[1] (though repositories may also provide metadata using a range of other formats). It is expected that open access repositories in the UK will be compliant with OAI-PMH 2.0.

2.1.3 CERIF

The CERIF standard, managed through EuroCRIS[2], is becoming an increasingly important standard for the exchange of research management information in the UK. The extent to which the UK metadata guidelines will need to be adapted to embrace the CERIF standard in the future is currently under investigation. OpenAIRE has recently announced that it plans to support the use of CERIF-XML to facilitate interoperability.

2.2 Metadata elements summary

The table below summarises the set of metadata elements that the Research Councils and the JISC would like organisations to collect, particularly those in receipt of Research Council funding. At present the scope is limited to metadata that pertain to publications. In future other research outputs including datasets may be included in the metadata guidelines. For the time being the Dublin Core (DC) metadata standard, familiar to all in the repository world, is being used. Qualified DC is used in two instances (dcterms:issued and dcterms.relation). The two new elements central to the development of these guidelines – project ID (a

[1] http://www.openarchives.org/OAI/2.0/oai_dc.xsd
[2] http://www.eurocris.org/Index.php?page=homepage&t=1

unique identifier normally provided by the funder) and funder name – have been conceived as an extension to the well-established bibliographic metadata elements using the *rioxxterms* namespace.

Key for inclusion: **M**: mandatory; **R**: recommended; **O**: optional

Element	Inclusion
dc:title	M
dc:creator	M
dc:identifier	M
dc:source	M
dc:language	M
rioxxterms.projectid	M
rioxxterms.funder	M
dcterms:issued	M
dc:format	R
dc:publisher	R
dc:description	R
dc:subject	R
dc:rights	R
dc:coverage	O
dc:audience	O
dc:type	O
dc:contributor	O
dc:relation	O
dcterms:references	O

2.2 Detailed description of the metadata elements

These UK-specific Guidelines have been developed with reference to the Driver and OpenAIRE Guidelines (which are related to the OpenAIRE project[3]) and UKETD_DC, the metadata core set recommended by the British Library's Electronic Theses Online Service (EThOS)[4]. As far as possible this phase of the UK Guidelines and the associated technical documentation deviate from both these resources as little as possible to limit problems with interoperability while at the same time achieving the key goals of UK stakeholders. In fact these UK guidelines differ from the OpenAIRE guidelines in only two metadata elements, namely rioxxterms.projectid and rioxxterms.funder. More detail about each element is provided below.

This guidance also draws on the wealth of information provided by the Dublin Core Metadata Initiative (http://dublincore.org/) largely because, in the quest for accurate, appropriate and consistent use of metadata, it is important that the RIOXX application profile and guidelines are rooted in standards that have been

[3] http://www.openaire.eu
[4] http://ethos.bl.uk/Home.do

developed over many years and which have been widely adopted around the world.

Whether you are creating metadata through a manual process or setting up the automatic conversion of existing records to new ones, these Guidelines exists to help with the organisation and management of those metadata. Care should be taken to attribute the most appropriate element to information. There may be occasions where the choice of element is not clear-cut so you will need to make a judgment. The key is to make these judgments on a consistent basis for your repository.

Please note that for the moment these Guidelines are designed primarily with publications in mind. The comments below often refer to a "resource" which for now should be taken to mean "publications". This semantic constraint may be amended in future versions of the Guidelines as other types of research outputs are considered for inclusion.

dc element	dc:title
Inclusion status	Mandatory
Format and comments	This refers to the resource's title and any sub-titles. Title should be entered using free text. Title is the form of words by which a resource will be formally known and should be represented using the original spelling and wording. Since these Guidelines are focused primarily on publications, journal and book titles are likely examples. The *recommended* format for subtitles is Title:Subtitle

dc element	dc:creator
Inclusion status	Mandatory
Format and comments	The creator of a resource may be a person, organisation or service. Where there is more than one creator, use a separate dc:creator element for each one. Enter as many creators as required. The dc:creator element should take an optional attribute called "id". This will hold a machine-readable unique identifier, where available, for the creator. Ideally the element will include a machine-readable id and a text string in the body of the element.

	For example, <dc:creator id=http://"identifier-for-this-creator-entity">name-of-this-creator-entity</dc:creator> Where the creator is a person, the *recommended* format is Last Name, First Name(s) and to include an ORCID ID, if known, in its HTTP URI form, such as: <dc:creator id=http://orcid.org/0000-0002-1395-3092>Lawson, Gerald</dc:creator> Note that if the creator is a person and you wish to record that person's affiliation, the affiliation should be recorded using the dc:contributor element.

dc element	dc:identifier
Inclusion status	Mandatory
Format and comments	This element must contain a globally unique and persistent identifier for the resource being described. A commonly-used example is a publisher's DOI. The identifier should be an HTTP URI that can be de-referenced (and is, thus, actionable). The aim of this element is to allow access to the resource so it is *recommended* that the identifier points to the actual resource being described by the RIOXX record – such as a pdf, normally held in the local repository – rather than to an intermediary resource such as a repository web page.

dc element	dc:source
Inclusion status	Mandatory
Format and comments	The source label describes a resource from which the current resource is derived (in whole or in part). This may be a working paper, a collection of works or a book for example. It is *recommended* that the source is

	referenced using a unique identifier from a recognised system e.g. the unique 8-digit International Standard Serial Numbers (ISSN) assigned to print and electronic periodicals or the International Standard Book Number (ISBN).

dc element	dc:language
Inclusion status	Mandatory
Format and comments	This refers to the primary language in which the content of the resource is presented. The element may be repeated if the resource contains multiple languages. A coded value or text string may be used. The values used for this element must conform to ISO 639-3 which offers two and three letter tags: “en” or “eng” for English and “en-GB” for English used in the UK.

dc element	rioxxterms.projectid
Inclusion status	Mandatory
Format and comments	This is an addition to Dublin Core’s fifteen generic elements and is designed to collect the grant numbers issued by funders to Principal Investigators that directly relate to the current resource. Note that different funders may use different language to describe their grant or awards. It is *mandatory* to use the full alphanumeric identifier provided by the funder in its original format e.g. ST/K001234/1 (denoting an STFC award). Sometimes publications have more than one funder associated with them; these must be recorded using separate instances of rioxxterms.projectid. In cases where projects have been funded internally, use whichever internal code is appropriate.

dc element	rioxxterms.funder
Inclusion status	Mandatory
Format and comments	This is an addition to Dublin Core's fifteen generic elements and is designed to collect the canonical name of the entity responsible for funding the resource. The funder name must be recorded here as text. It is very important that funders can identify outputs they have funded so you should use a controlled list of funder names. A list has been provided for this purpose and is available through the RIOXX website (http://docs.rioxx.net.funders). Where more than one funder has contributed to the resource, each must be entered as a separate instance of rioxxterms.funder.

dc element	dcterms:issued
Inclusion status	Mandatory
Format and comments	This element is designed to record the publication date of the resource. For resources such as books or journal articles the "published date" is normally provided by the publisher. For other resources the published date will normally mark the date at which the resource is first made publicly available - which may be the date it is deposited in an open access repository. The date should be encoded using ISO 8601 (post-2004 versions) that follows the following format: YYYY-MM-DD. Year (YYYY) or year and month (YYYY-MM) may be used if the full date is not known.

dc element	dc:format
Inclusion status	Recommended
Format and comments	This refers to the form of the resource being described in the RIOXX record,

	physical or digital, and can refer to the media-type or dimensions of the resource. Where the resource being described is digital, the MIME type of the object pointed to be this RIOXX record's dc:identifier element must be entered here. If more than one category is needed to describe a single resource, use separate instances of the dc:format element.

dc element	dc:publisher
Inclusion status	Recommended
Format and comments	A free text string giving the name of the entity responsible for making the version of record of a resource available. This could be a person, organisation or service. If the status of the publisher is unclear, it is *recommended* to use dc:creator for people and dc:publisher for organisations.

dc element	dc:description
Inclusion status	Recommended
Format and comments	This field may be indexed and its contents presented to people conducting searches. The goal is to describe the content of the resource using free text. It is *recommended* that an English language abstract be used where available. HTML or other structural tags should not be included in this field.

dc element	dc:subject
Inclusion status	Recommended
Format and comments	Normally keywords, phrases or

	classification codes are used to describe the topic of the resource. If using free text, avoid using general keywords. The *recommendation* is to use a formal classification scheme or controlled vocabulary e.g. Library of Congress Classification Headings or Medical Subject Headings (MeSH). When including terms from multiple vocabularies, use separate element iterations. If multiple vocabulary terms or keywords are used, either separate terms with semi-colons or use separate iterations of the Subject element.

dc element	dc:coverage
Inclusion status	Optional
Format and comments	This refers to the scope or extent of the content of the resource. It may include jurisdictional, temporal or spatial information. It is *recommended* that, where possible, a recognised globally unique identifier is used, such as the Thesaurus of Geographic Names, but free text may be used. For example, the place of publication may be recorded.

dc element	dc:rights
Inclusion status	Optional at present
Format and comments	The use of a URL to an appropriate Creative Commons license is *recommended.* E.g. http://creativecommons.org/licenses/by-sa/2.0/deed.en_GB Work is under way to develop consensus on a controlled vocabulary that describes rights to open access items as well as the associated issues of Creative Commons licenses and embargo periods. Once this work concludes the appropriate use of this element is expected to become mandatory.

dc element	dc:audience
Inclusion status	Optional
Format and comments	This field is designed to contain information about the group for which the resource is intended or is considered to be useful. There is no established vocabulary for this but sometimes creators or publishers indicate the intended audience. Note that the Research Outcomes System (ROS) used by most of the UK's Research Councils track whether a resource is for a "non-academic audience" (with a drop-down list of possible selections) and whether a resource is for an "international audience". In the absence of alternative formal vocabularies, following the ROS lead is a reasonable course of action.

dc element	dc:type
Inclusion status	Optional
Format and comments	Type refers to the nature or genre of the content of the resource and can be entered as free text. The development of a controlled vocabulary is likely to be recommended for Phase 2 of the RIOXX project. For the present, use separate instances of dc:type for resources comprising multiple types. Do not confuse this with dc:format (which has to do with the *form* of a resource).

dc element	dc:contributor
Inclusion status	Optional
Format and comments	This element is designed to describe an entity – for example the name of a person, organisation or service – responsible for making contributions to the content of the resource. As many instances of the dc:contributor elements as required may be entered.

	If the contributor is a person and it is desire to record that person's affiliation, the affiliation must be recorded as a separate dc:contributor element. The dc:contributor element should take an optional attribute called "id", designed to hold a machine-readable and unique identifier, if available, for the contributor. Any ID entered here must be in a form which allows it to be read automatically. The ideal use of this element is to include both a machine-readable ID in the id attribute and a text string in the body of the element. For instance: <dc:contributor id="identifier-for-this-contributor-entity">name-of-this-contributor-entity</dc:contributor> Where the contributor is a person, the recommended format is text in the following form: Last Name, First Name(s) AND to include an ORCID ID, if known, in its HTTP URI form, such as: <dc:contributor id=http://orcid.org/0000-0002-1395-3092>Lawson, Gerald</dc:contributor>

dc element	dc:relation
Inclusion status	Optional
Format and comments	The format of this element should be an HTTP URI that points to a related resource. It is *recommended* that, where available, the publisher's DOI for the main resource being described in the RIOXX record also be entered here in its HTTP form, e.g. http://dx.doi.org/10.1006/jmbi.1995.0238 Each related resource must appear as a separate instance of this element..

dc element	dcterms:references
Inclusion status	Optional
Format and comments	This element should contain an HTTP URI that points to a resource referenced by the resource described in the RIOXX record, e.g. a dataset that underpins an article being described in the record. Each reference must appear as a separate instance of this element.

3. Helping you adopt these guidelines

The sponsors of these guidelines are committed to helping you adopt them. You are likely to be already collecting most of the mandatory metadata but you may need to think about the two additional fields (ProjectID and Funder Name) and where to source the information. If you do not already have this information your institution's Research Office may be able to supply this it. The RIOXX project team is working to agree access for the community to a new directory of unique funder names.

The RIOXX project is working with EPrints and DSpace developers to develop the plugins and patches necessary to facilitate the efficient capture of the required metadata where this is not already done and to expose the captured data, according to the defined RIOXX application profile, through the OAI-PHM protocol. The goal is to make compliance with these metadata Guidelines as simple as possible.

3 Frequently Asked Questions

3.1 Why do we need guidelines

These Guidelines are designed to mitigate the detrimental effects of divergent interpretation of the standards that exist in the open access repository space – OAI-PMH for example – by advocating a common approach. They are not data entry instructions but the guidelines do provides the means to map your data entry processes to the required format. Adopting a common approach through the use of generally-used guidelines has the potential, therefore, to reduce ambiguity, boost the extent to which metadata can be harvested efficiently, enhance the accuracy and value of services built on metadata harvesting and aggregation processes and improve confidence in the veracity of reports based on metadata.

3.2 Do I have to abide by these guidelines?

The development of these Guidelines was instigated and is being strongly supported by RCUK and the JISC. While their use is not compulsory, the benefits for many stakeholders – researchers and other information consumers, funders and institutions – are attractive. Better information discovery, higher quality statistical reporting, higher quality aggregations and the possibility of building

new services will all flow from a consistent approach to collecting and exposing metadata in the UK's open access repositories. Working together, the UK's information management community can continue to promote the importance and usefulness of their open access repositories both within and beyond their own institutions. Aggregators report the quality of metadata improving as a result of the work being done towards reporting for the REF. The Guidelines and associated documentation simply provide the tools to help the research information management community pull in the same direction for the common good.

3.3 Are these guidelines supported by the community?

The RIOXX application profile and guidelines are have been developed in consultation with interested parties in the community and, in particular, with the cooperation of UKCoRR. In addition the project team have ensured as far as possible that there is a high degree of compliance between these and the OpenAIRE guidelines. There is now an opportunity for further feedback from all interested parties. The application profile is unlikely to change before the launch date (15th April) but suggestions to improve these guidelines are welcome until 5th April. Feedback can be sent to admin@rioxx.net. Given the dynamic nature of the sector and the initiative to develop vocabularies and associate metadata elements for open access, these Guidelines will in any case evolve over the course of 2013. There will be further opportunities for people to contribute to the ongoing development of the Guidelines. The basic elements will not change but where there is a need for perhaps greater clarity or additional examples that need will be addressed.

3.4 What is RIOXX and who can I contact if I have questions or need help?

The project to develop the UK Metadata Guidelines is an extension of an earlier project looking at Repository Interoperability Opportunities (RIO), hence the acronym RIOXX.

For questions relating to the appropriate use of metadata elements, the DCMI[5] offers a comprehensive array of relevant information. For questions about the project or the software for EPrints and DSpace, the project is being directed by UKOLN at the University of Bath. The project website provides information about RIOXX and the project team can be contacted using the following email address: admin@rioxx.net.

[5] http://dublincore.org/