Metadata and EPrints Customisation for the UWE Data Repository: Objectives, requirements and standards for a data repository

Purpose of this document

This document records and explains the rationale behind the choice of metadata for the UWE data repository. It explains the principles which guided metadata development during the pilot project for the benefit of the wider project team. For the future it will inform work on and developments of the data repository and will enable any future repository team to determine if the same rationale and principles apply or if adaptations are necessary for future developments.

Starting point

Information from the British Library DataCite metadata workshop on 6th July 2012 (British Library and DataCite, 2012b) was used to analyse metadata requirements for the UWE data repository. The metadata should provide sufficient information on the associated datasets to enable anticipated practices such as citation, discovery and re-use.

Main objectives of metadata:

Ensure data is citable, by including required citation information.

Promote data discovery.

Enable re-use.

Ensure a recognised metadata schema is used but data entry provision is not burdensome for researchers.

Additional requirements

Include links to papers which make use of data. We should expect to include links to outputs relating to the original research including papers in the UWE research outputs repository and in academic journals. Where re-use of the data can be identified, through citation, it would be useful to include links to any outputs from such research to demonstrate impact and re-use potential.

Uniquely identify subsets of the data as well as the whole dataset.

Provide a mechanism for access if electronic (or enough information to enable access if not online).

Metadata should be usable by software and search tools (and humans) to allow others services to be built.

Meeting the main objectives

1. **Data Citation**

Data citation requires the name of a responsible agent and a globally unique and persistent identifier enabling location and access. The responsible agent for UWE data citation would be the UWE Data Repository. The EPrints software provides a unique, persistent identifier for each dataset.

No agreed standard citation style for data exists and styles vary between journal publishers and repositories, so required information is variable. The DCC (Ball & Duke, 2011), DataCite (DataCite, 2011) and UWE (University of the West of England, date unknown) identify the following fields or properties as required for data citation:

Table 1: DCC, DataCite and UWE metadata fields required for Data Citation

DCC required field	DataCite required property	UWE required fields
Author	Creator	Author
Publication date	Publication Year	Publication year
Title	Title	Title of data
Location	Identifier	Location
	Publisher	
		Title of database
		Version (if more than 1 version)

For consistency within UWE we should ensure the data repository supports the UWE citation style for datasets:

Author surname, initials. (Year of publication) Title of data, *Title of Database* (version) [online]. Available from: URL [Accessed DD Month YYYY].

Example:

NASA Goddard Spaceflight Center (2008) Geometric and seismological data, *Global Change Master Directory* (2008 version) [online]. Available from: http://gcmd.nasa.gov [Accessed 15 October 2008].

After discussion with the chair of the UWE Library Referencing Task & Finish Working Group the UWE required fields are agreed to represent:

Author=Creator(s) of the dataset

Year of publication = year of publication in the UWE data repository

Title of data=title of research project

Title of database= UWE data repository

Version= publication year (optional, include if more than 1 version)

Location=unique persistent URL.

Mandatory elements for citation

The minimum metadata elements required for citation are author, publication date, title and location (table 1).

The data repository software (EPrints) provides a globally unique, persistent identifier: the URL of the research project dataset. Therefore, for the UWE data repository, the location becomes the persistent identifier.

Publisher

Alex Ball, presenting at the second British Library DataCite workshop on metadata, recommended including the online publisher in the metadata for citation as it provides a remedy if links break (Ball, 2012). As long term retention and preservation of datasets is envisaged the publisher will be included in the UWE mandatory fields to provide an identifiable responsible body for long term support.

The publisher identifies the entity archiving and publishing the data; normally UWE but if the data is published by a leading institution as part of a collaborative project or archived in a national or subject repository they would be the publisher. In this case the repository would link to the data in either the leading institution repository or national or subject archive. If the data is owned by a funder, an agreement with the funder/owner is required to allow UWE to publish the data and should specify if the funder needs to be cited as a co-publisher.

Optional elements for citation

Common data citation styles (in publications and other repositories) which require additional fields should be accommodated through optional elements. Publisher, resource type and version are the most common additional fields.

In the UWE data repository the resource type field will be automatically populated as dataset.

Version will only apply if multiple versions exist in the UWE repository, it is anticipated most datasets will only have one version. The British Library DataCite workshop on metadata (British Library and DataCite, 2012b) recommended using different DOIs (persistent identifiers) for different versions. In the UWE data repository each version will have its own unique dataset and persistent identifier (URL); the version number is needed for clarity and completeness of the data record. The UWE outputs repository uses 'year of publication' to identify relevant versions. Consistency of approach across the UWE research repositories would be simpler for researchers and repository staff. Therefore, in the UWE data repository the version identification is the year of publication and, if multiple versions exist in a single year, an appended alphabetical character (a, b, c etc.) should be added to the year to uniquely identify each version.

Table 2: Recommended Metadata for citation

Recommended Metadata for citation	DCMI element	Notes or vocabulary terms
Required		
Author/creator	Creator	Multiple authors possible
Title of project	Title	
Publication date	Date	publication
Location	Identifier	Location=identifier
Publisher	Publisher	UWE unless collaborative or

		archived outside UWE data repository
Optional		
Resource Type	Type	DCMI Type Vocabulary = dataset
Version		Use publication date if more than one version exists

2. Promote data discovery

Whilst it is known that search engines, and Google in particular, are used extensively by researchers to discover information, focussing exclusively on full text indexing and searching as an alternative to a metadata schema would exclude discovery of any non-text items within the repository. The data for the seven pilot research projects of the UWE JISC RDM project include video, audio and other file formats for which only metadata can be interrogated. The UWE Research Output Repository has recognised the diverse research profile of UWE and the challenge of creating metadata for diverse material (Clarke & Lawson, 2012).

Initially, the 15 elements recommended for data discovery by Alex Ball at the second British Library DataCite workshop on metadata (Ball, 2012) would provide basic discovery metadata for the research projects in the UWE data repository pilot (table 3). If in the future additional terms are required, a wider range of DCMI elements or terms (DCMI Usage Board, 2012) or application profiles (Coyle & Baker, 2009) could be investigated for their suitability.

Two of the fifteen elements are included as optional elements in the UWE metadata for citation (resource type and version).

Table 3: Recommended metadata for data discovery and DCMI elements

Recommendations: metadata for discovery	DCMI element	Notes or vocabulary terms
Contributors	contributor	
Abstract/Summary/ Description	description	
Subject/Keywords	subject	Library of Congress Classification (LCC) standard in EPrints
Rights/Restrictions	rights	Licence and access restrictions
Spatial Coverage	coverage - spatial	
Temporal Coverage	coverage - temporal	
Derived publications	relation	
Related Datasets	relation	
Resource Type	resource type	Included in citation

		metadata
File Format(s)	format	mime_type (at data file level)
Important Dates	date	
Language	language	
Version		Included in Citation metadata
Size (of data files)	Class=size or duration	
Metadata Record Date		Use publication date

3. Enable re-use

Information enabling re-use relies on a contextual understanding of the research, the methodology used and, where appropriate, details of experiments undertaken for example instrument settings used. Including data entry fields for project methodology (at the project level) and data description (at the data file level) will facilitate this.

Best practice recommends the research methodology needs to be captured at an early stage and throughout the data life cycle. In addition the information required varies by and within disciplines. Only researchers can provide this level of detailed information (metadata) on their research data. Such metadata can be provided as additional file(s) deposited with the research project data set, for example an XML file. It should be explained to researchers that this additional information is optional and any such additional metadata files should conform to established subject and disciplinary standards. It would be advisable to verify metadata quality prior to publication.

How support and advice for researchers on metadata standards and verification of metadata will be provided needs to be determined; if this will be librarians then training will be required.

Table 4: Metadata requirements summary

Required repository field (project level)	**DCMI**	**Type, vocabulary and notes**	**Recommended for**
Author/creator	Creator	Multiple authors possible	Citation
Publication date	Date	Also identifies version if more than one version exists, automatic	Citation
Title of project	Title		Citation
Location	identifier	automatic	Citation
Publisher	Publisher	Automatic if UWE	Citation
Optional repository field (project level)			

Resource type	Type	DCMI Type Vocabulary = dataset, automatic	Citation, discovery
Version			Citation, discovery
Contributor(s)	Contributor		Discovery
Abstract/Summary/ Description	Description		Discovery
Subject/Keywords	Subject	LCC	Discovery
Rights/Restrictions	rights	Licence and access restrictions	Discovery
Spatial Coverage	coverage - spatial		Discovery
Temporal Coverage	coverage - temporal		Discovery
Derived publications	relation		Discovery
Related Datasets	relation		Discovery
Important dates	date	automatic	Discovery
Language	Language		Discovery
Metadata Record Date		Publication date, automatic	Discovery
Methodology			Re-use
Required repository field (data level)			
Data description	description		Re-use
File format	format	mime_type, automatic	Discovery
Size	Class=size or duration	automatic	Discovery

4. Using a recognised metadata schema and reducing data entry workload for researchers

4.1. A metadata schema for the UWE data repository

The current UWE outputs repository uses the Dublin Core schema which is well understood, widely used and interoperable. Using the same schema for the data repository would be of some advantage for the repository administrators in simplifying workflows.

The alternative DataCite metadata schema mandates the use of a DOI as the unique identifier (DataCite, 2011). To enable this entails UWE becoming a member of the British Library DataCite service and entering into a contractual relationship with the British Library which requires members to give a 'clear and public indication to preserve data' (British Library and DataCite, 2012a) through, for example, formal plans, service agreements and statements. Although no final decision has been reached, the current costs of membership (£1500 per annum) and the resources required to establish a contractual relationship

suggest this is an unlikely step during the current pilot stage in the development of a research data management service at UWE.

These two schemas are the most relevant to the UWE project; investigation of other schemas was not undertaken due to restrictions on UWE capacity to accommodate a variety of schema. Therefore Dublin Core is the recommended metadata schema for the data repository.

4.2 Reducing data entry workload for researchers

At present UWE has a limited number of IT systems supporting research projects and no interoperability requirements for the data repository with those that currently exist. Therefore we expect most of our metadata to be entered manually by researchers on deposition of research data. Whilst this may allow a richer set of metadata than possible with harvested metadata it places the burden for metadata acquisition on the researcher. If the burden was considered too great the metadata capture would fail. Therefore, we have limited mandatory fields to five, those required for citation and location. An additional seventeen are optional. Of these twenty-two, eight are standard or entered automatically by EPrints. A key element of our researcher testing will be to assess the acceptance and usability of the metadata fields by researchers. If metadata entry is found to inflict a burden we will review these requirements.

Automatic validation of fields will be employed wherever possible to improve data quality and reduce data entry workload for researchers.

5. Analytics

To allow for some basic analytics by research administration the following additional information is required:

Project cost code

PASS (Project Approval Support System) reference number

Source of funding or governance (funder, sponsor)

Additional organisations (collaborator, partner)

UWE research centre or institute

UWE faculty and department

Additional project team members

6. Data Repository Customisation

6.1. Deposit workflow

Deposit agreement -> Project details -> Data file upload -> Deposit research data (submit complete).

6.1.1. Deposit agreement

Agreement to terms and conditions is mandatory

6.1.2. Project details

Data entry and metadata requirements for project data listed in Appendix 1.

6.1.3. Data File upload

Data entry and metadata requirements for data files listed in Appendix 2.

6.1.4. Research project data submission complete (all data files loaded).

6.2. Landing page

All projects require a landing page which should provide descriptive metadata, a sample citation, a link to accompanying paper(s), instructions on how to access data and the licence under which the data is released.

6.3. Researcher testing

During testing all the researchers found the amount of information and the data entry requirements acceptable. The information required was relevant and achievable with one exception; the PASS reference number. They requested the information requirements for data entry should be available to enable advance preparation for deposit, which should address this issue. These requirements will be included in the repository web page guidance.

The current subject list was uniformly disliked by all the testers. It was felt to be too general for UWE research, for example it does not include UWE research subjects (occupational therapy, health care, long term conditions or health psychology), the only subject term feasible (medical) would be inadequate for reliable, relevant discovery. Ideally subject options should match or give a reasonable approximation to UWE research areas. A review of subject classification usage in UK research found some interesting and relevant work: blog posts from the Engage project (Unknown, 2012), and CERIF for Datasets (Garfield, 2012) which suggested no classification scheme is widely accepted for research data or has the required level of granularity.

As the value of including a subject field will be far outweighed by the inaccuracies in selection and use and the frustration and irritation to depositors and users alike, the subject field was removed from our metadata and instead we have opted for uncontrolled keywords. We recognise this is not ideal but aim to produce better quality descriptive terms and facilitate more accurate discovery.

Subsequent emails to the JISC MRD list (Ensom, 2012) suggested the HESA JACS3 classification or RCUK and LoC in SKOS (Boyd, 2012) as alternatives. These can be explored later when investigation of research in other faculties following this pilot project can inform decisions on the selection of an appropriate subject schema for the data repository.

7. Policy/strategy

Maintaining data as citable is the responsibility of the data publisher/distributor. UWE and the data repository need a formal digital preservation and versioning policy/strategy to:

7.1. Determine under which Creative Commons licence the data in the repository will be made available to users and what uses will be permitted.

7.2. Ensure published data and metadata remains accessible (static) over time, persistent identifiers are assigned to data and remain associated, published data remains accessible, depositors are provided with a citation for use with published papers and a link to the paper once published. The EPrints software will manage this but a policy statement is needed to cover and state responsibilities which belong to the repository.

7.3. Determine if the repository accepts open access datasets only. If restrictions are acceptable, the circumstances under which restricted access data is accepted and if restrictions apply to the whole datasets (recommended) or individual files within a dataset. Metadata is always open access; these restrictions apply to the full text files only.

7.4. Versioning control e.g. the use of different identifiers (& citations) for different versions of the same dataset. A definition describing what is a new version e.g. the circumstances or frequencies which constitute a new version and requirements or limits for storing and maintaining earlier versions. Determine the method for dealing with frequently revised datasets or frequently expanding datasets (series) and whether these types of datasets should be live.

7.5. Explain responsibilities of researchers for example: using standard file formats for data for ease of publication and sharing, obtain citations from data publishers to include with papers, including citations for prior or re-used datasets.

7.6. Decide how to discover, list and include links to known citing papers and maintain citation lists on the re-use of datasets in the repository. It may be necessary to wait until data publication and citation is more advanced and embedded in practices before this can be easily achieved.

7.7. Determine what, if any, metadata support will be provided for researchers who supply additional metadata files using subject or disciplinary standards?

7.8. Determine what, if any, data quality assurance support will be required for deposited data.

7.9. Consider dataset item types within the outputs repository; amend or delete existing types once pilot data repository live. Decide if datasets in the outputs repository should move to the data repository.

7.10. Consider if data will expire. If so, decide on the criteria for expiration.

Responsibility for determining these strategies and policies has yet to be decided.

8. Future Development

A useful development would be using data in the UWE PASS (Project Approval Support System) to load data repository fields. The potential for interoperability between PASS and the data repository will need investigating to support this work. A first stage would be to include the PASS

reference in the data repository fields. Benefits lie in the reduced data entry workload for researchers and improved data quality and consistency.

Where re-use of the data can be identified, through citation, it would be useful to include links to any outputs from such research to demonstrate impact and re-use potential. How information on citation is discovered and maintained requires further investigation but should be included as part of the development of policy and strategy for the data repository.

Appendix 1

Project data entry fields (EPrints) - use drop down lists for data validation wherever possible.

Field	EPrints field	Type of field	Suggested help text
Title of research project	Title	text entry box	The title of the project. The title should not end with a full stop, but may end with a question mark. There is no way to make italic text, please enter it normally. If you have a subtitle, it should be preceded with a colon [:]. Use capitals only for the first word and for proper nouns. Example: **A review of pain management and patient response following major joint surgery in an English NHS hospital trust** Example: **Cognitive-behavioural approaches in routine care: Rheumatology as a model**
Brief Summary of project	Abstract	text entry box	A summary of the project aims and objectives. If the project has a formal summary or description then that is what should be entered here. No complicated text formatting is possible.
Methodology	uwe-methodology	text entry box	Please include information on the methodology used in your research project which would enable other researchers to understand the research methods used and the data collected.
Creators	Creators	multiple lines; 3 fields: family name, given name, email address	Researchers involved in creating the data: normally list researchers working on this specific project but if research is usually attributed to all members of a research team then please list all team members. UWE email addresses only, non-UWE email addresses not allowed due to data protection.
Additional team members	uwe_add_members	text entry box	Members of the research team who were not involved in creating the data: either research staff who did not create research

			data or non-research staff who assisted with the research e.g. statisticians, technical staff. If research is usually attributed to all members of a research team then please list all team members as creators. UWE email addresses only, non-UWE email addresses not allowed due to data protection.
UWE Faculty/Department	divisions		The UWE faculty and department which undertook the research. To select multiple entries, hold the CTRL key button and select additional entries from the list.
UWE Research Centres/Institutes	uwe_centres	Text entry box or select from list?	If the research was undertaken within a UWE research centre or institute please select the appropriate centre or institute. If no research centre or institute was involved move to the next field.
Associated publication details	related_url	link	Any publication such as a journal article or conference paper which resulted from this research. For online articles the preferred link is to the version in the UWE repository if freely available, otherwise link to the publisher article. If only print is available please link to the record in the research repository if available.
Related datasets	uwe_related_datasets	link	Related datasets are those which were used or assisted in the development of this data or have evolved from this dataset through subsequent research or as later versions of the dataset. Please provide a link to the related dataset or the online location (repository or institution) and title of the dataset. If there is more than 1 related dataset, click on the [More input rows] button.
Publisher	publisher	UWE = default, option to change	UWE is the publisher of all research data archived in the UWE data repository. If the dataset is published in a national or subject archive or by a leading institution in a collaborative project; the publisher is the

			archive or institution responsible for publishing the data. This record should then provide a link to the published dataset wherever it is archived.
Source of Funding or Governance	funders	multiple lines; 1 field: Organisation name	Funding bodies or organisations who contributed funding or influenced governance of this research project e.g. funders or co-sponsors with UWE.
Additional organisations	corp_creators	multiple lines; 1 field: Organisation name	Other organisations which assisted or contributed to the research project and data creation e.g. partners, collaborators.
Project Cost Codes	projects	Text entry box	The finance name or code of the project that created this data.
PASS reference number	uwe_pass_ref		The reference number of this research project in the UWE PASS system.
Language	uwe_language	English default, option to change?	English is the standard language used by research projects and data in the UWE data repository. If this project used a different language please enter the correct language for the project and its datasets.
Spatial Coverage	uwe_spatial	text entry box	If the research has distinct spatial groupings or boundaries please indicate them here e.g. City of Bristol, South-West England, Europe.
Temporal Coverage	uwe_temporal	text entry box	If the research has a limited time span or boundaries please indicate them here. For specific start and end dates the preferred format is yyyy-mm-dd to yyyy-mm-dd. For non-specific ranges November 2007 - March 2008 or spring 2009 - spring 2010 is acceptable.
Contact email address	contact_email	text entry box	The contact email address for this item. If the full-text is not available to the public, then requests to view the full-text will be sent to this email. The email address will not be made public. Example: j.smith@uwe.ac.uk
Uncontrolled keywords	keywords		Words or a phrase which describe the research data
Additional	note	text entry	If you think you can specify some

Information		box	useful information about your project or data deposit that can't be entered anywhere else, please enter it here. This information **will** appear on the public summary page for this item.
Comments & Suggestions	suggestions	Text entry box	Any comments to the editor. This information **will not** be displayed to the public.
Publication date	Deposit live date		n/a
Location	Identifier	URI	n/a
Resource type		=dataset (automatic)	n/a
Metadata record date	EPrints time stamp	automatic	n/a

Appendix 2

Data files data entry fields

Field	EPrints equivalent properties	Type of field	Suggested help text
Format	format	Automatic - EPrints needs to recognise data file format.	n/a
Description of data file contents	uwe_content_description	Text entry - mandatory field	To enable others to understand the research data please describe the origin and contents of the data file.
Visible to	security	Drop down list	Please indicate the required security level of this file. Who is allowed to download it?
Licence	license	Drop down list: auto-populated	No licence has been agreed for data in the data repository. Until a licence is agreed please contact the data repository for advice on licences. If an embargo applies immediately, the licence will come into force when the embargo is removed.
Embargo expiry date	embargo	YYYY/MM/DD selection	The date that a funder permitted embargo expires. On or after this date the data files will be made available in accordance with the data repository licence.
Access restrictions	uwe_restrictions	Text entry	List any access restrictions which apply to this data file, reasons and conditions which should be satisfied for access to be granted.
Size	filesize	Size of the file - automatic	n/a

Bibliography

Ball, A. and Duke, M. (2011). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides.* Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides. Accessed: 18 July 2012.

Ball, A. (2012). British Library DataCite Workshop no. 2: Describe, disseminate, discover : metadata for effective citation. *Metadata for Data Citation and Discovery*. [presentation] [online] London/British Library 6 July 2012 Available from: http://www.bl.uk/aboutus/stratpolprog/digi/datasets/workshoparchive/ball-metadata-citation.pdf Accessed: 18 July 2012

Boyd, D. (2012) Email to JISCMRD@JISCMAIL.AC.UK, 16 November.

British Library and DataCite (2012a). British Library DataCite Workshop no. 1: An introduction to Data Citation and DataCite. Available online: http://www.bl.uk/aboutus/stratpolprog/digi/datasets/workshoparchive/archive.html [Accessed: 23 July 2012]

British Library and DataCite (2012b). British Library DataCite Workshop no. 2: Describe, disseminate, discover: metadata for effective citation. Available online: http://www.bl.uk/aboutus/stratpolprog/digi/datasets/workshoparchive/archive.html [Accessed: 23 July 2012]

Clarke, A. and Lawson, A. (2012). Repository metadata for diverse collections. *Catalogue and Index.* 167. pp. 16-19.

Coyle, K. and Baker, T. (2009). *Dublin Core Metadata Initiative.* Available online: http://dublincore.org/documents/profile-guidelines/ [Accessed 23 July 2012].

DataCite (2011). *DataCite Metadata Schema for the Publication and Citation of Research Data.* Available from: http://schema.datacite.org/meta/kernel-2.2/index.html [Accessed 23July 2012].

DCMI Usage Board (2012). *DCMI Metadata Terms.* Available from: http://dublincore.org/documents/dcmi-terms/#elements-date [Accessed 24 July 2012].

Ensom, T. (2012) Email to JISCMRD@JISCMAIL.AC.UK, 13 November.

Garfield, Sheila (2012) Taxonomy Definition. *Cerif for Datasets* [blog]. 11 April. Available from: http://cerif4datasets.wordpress.com/tag/subject-classification-scheme/ [Accessed 20 December 2012]

University of the West of England (date unknown). *How do I reference...*Available from: http://www.uwe.ac.uk/library/resources/general/iskillzone/referencing/referencingcontents/index.html [Accessed 23 July 2012]

Unknown (2012) Choosing a research classification scheme. Research Clusters [blog]. 05 March. Available from: http://researchclusters.wordpress.com/2012/03/05/choosing-a-research-classification-scheme/ [Accessed 20 December 2012]