REPORT

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

EQUELLA data repository pilot

Author(s):	Mark Berry & Dr Thomas Parsons
Audience:	Project stakeholders & JISC MRD
Published:	30/11/2012

Contents

1. Introduction

The following report details the analysis carried out to determine whether the EQUELLA digital repository[1] could be utilised for research data at The University of Nottingham (Nottingham). They are intended to be a position paper on our understanding at present, and will contribute towards the development of a research data archive system within the project lifetime. It is assumed that readers of this report will have a familiarity with both EQUELLA and research data management concepts and terminology.

2. General Workflow and Key Issues

The screenshots in this report go through a wizard for the entry of metadata in EQUELLA. They help to visualise the questions that need to be answered for each page and a discussion and a series of questions follows each page. The last section shows the metadata definition used for the demo, and highlights some issues that emerged regarding the metadata definition.

There are additional workflow requirements at the end of the wizard process which have not been defined. The key question is: At what stage should a DOI be obtained, and what kind of validation of the data is required before submitting the details to DataCite[2]?

Only Open datasets can obtain a DOI, and a link to a landing page must be provided and maintained by the publishing institution. Non-open datasets - if they too are to be stored in this repository - may need another (internal) unique identifier.

The workflow requirement is probably that somebody (a librarian?) needs to validate the metadata entered by the researcher before obtaining a DOI (either via DataCite or in some other workflow software solution).

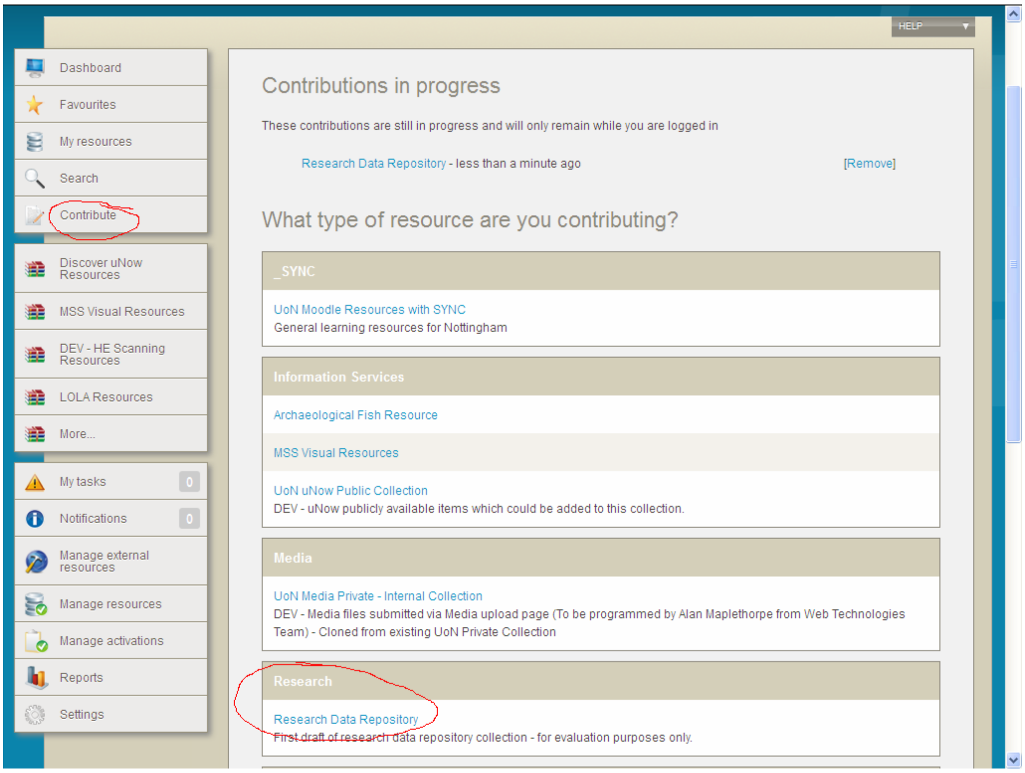
Note that this implementation in EQUELLA does not include a 'landing page' for public display of the metadata: it is assumed that a public website for browsing and accessing datasets would be implemented outside of EQUELLA in a separate web application.

[1] http://www.equella.com/

[2] http://www.datacite.org/

2.1. EQUELLA Contribute Page

In order to add a metadata record to EQUELLA, the user first clicks the 'Contribute' button on the menu on the left hand side of the screen, bringing up the following page (see Figure 1) which lists all the collections in EQUELLA. They should then choose 'Research Data Repository' to start adding metadata for a dataset.



Contributions in progress

These contributions are still in progress and will only remain while you are logged in

Research Data Repository - less than a minute ago [Remove]

What type of resource are you contributing?

_SYNC

UoN Moodle Resources with SYNC
General learning resources for Nottingham

Information Services

Archaeological Fish Resource

MSS Visual Resources

UoN uNow Public Collection
DEV - uNow publicly available items which could be added to this collection.

Media

UoN Media Private - Internal Collection
DEV - Media files submitted via Media upload page (To be programmed by Alan Maplethorpe from Web Technologies Team) - Cloned from existing UoN Private Collection

Research

Research Data Repository
First draft of research data repository collection - for evaluation purposes only.

Figure 1: Contribute page

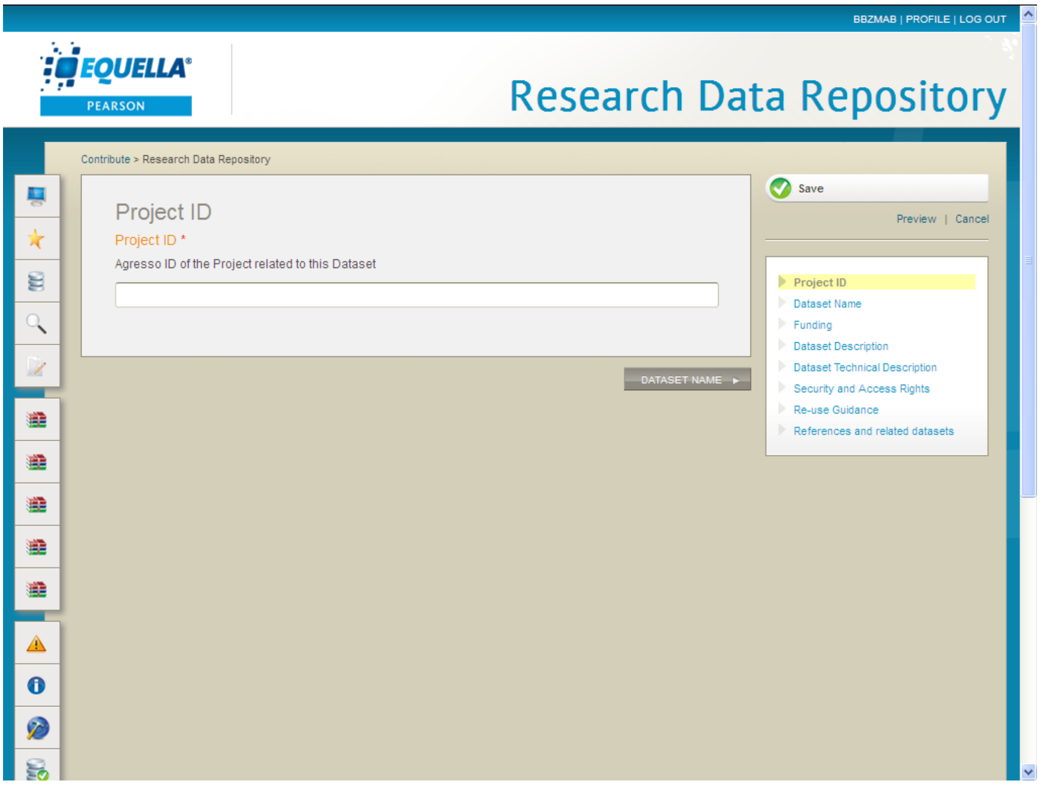
The EQUELLA interface presents a lot of features and options to the user that are not relevant to the specific use case of entering metadata for a research dataset, and this additional functionality might be unnecessarily confusing for end-users. It may be possible to provide a URL to take researchers directly to the "Contribute New Item to Research Data Repository" wizard, but it is not immediately obvious how this can be achieved. Some customisation of EQUELLA's appearance is possible, but it is probably not feasible to radically simplify the whole interface because the other functionality may be needed for other projects using EQUELLA.

2.1.1. Questions

- Is this interface suitable for end-users (researchers) to use, or would a separate data-entry web application need to be written as a front-end to an EQUELLA metadata store?
- What should the Collection be called and what should be the text for its description? It’s currently labelled a "Research Data Repository", but it's not really a repository, more of a metadata data catalogue.

2.2. Project ID

Figure 2 illustrates the data entry process in EQUELLA using this wizard. Note that the menu items down the left hand side remain in place (although they are irrelevant to the particular task at hand), and down the right hand side the pages of the wizard are set out, with Save, Preview and Cancel options. This interface is fixed in EQUELLA, and may or may not be suitable for the use case of a researcher entering metadata for research datasets.


BBZMAB | PROFILE | LOG OUT
EQUELLA
PEARSON
Research Data Repository
Contribute > Research Data Repository
Save
Preview | Cancel
Project ID
Dataset Name
Funding
Dataset Description
Dataset Technical Description
Security and Access Rights
Re-use Guidance
References and related datasets
DATASET NAME ▸

Project ID

Project ID *

Agresso ID of the Project related to this Dataset

Figure 2: Project ID entry page

The wizard begins with the entry of a Project ID. The assumption here is that metadata will then be pulled in from various Nottingham business systems to populate the rest of the pages in the wizard, as far as possible. However, note that what is being entered here is metadata for a **dataset**, not for a project. The metadata pulled in using a Project ID might not be accurate for the dataset, and if it's pulled in automatically, researchers might not notice that it needs to be modified for the dataset. Some datasets might not have associated projects with IDs. There might also be other IDs that are relevant - for example, possibly a Dataset ID from DMP Online[3].

2.2.1. Questions

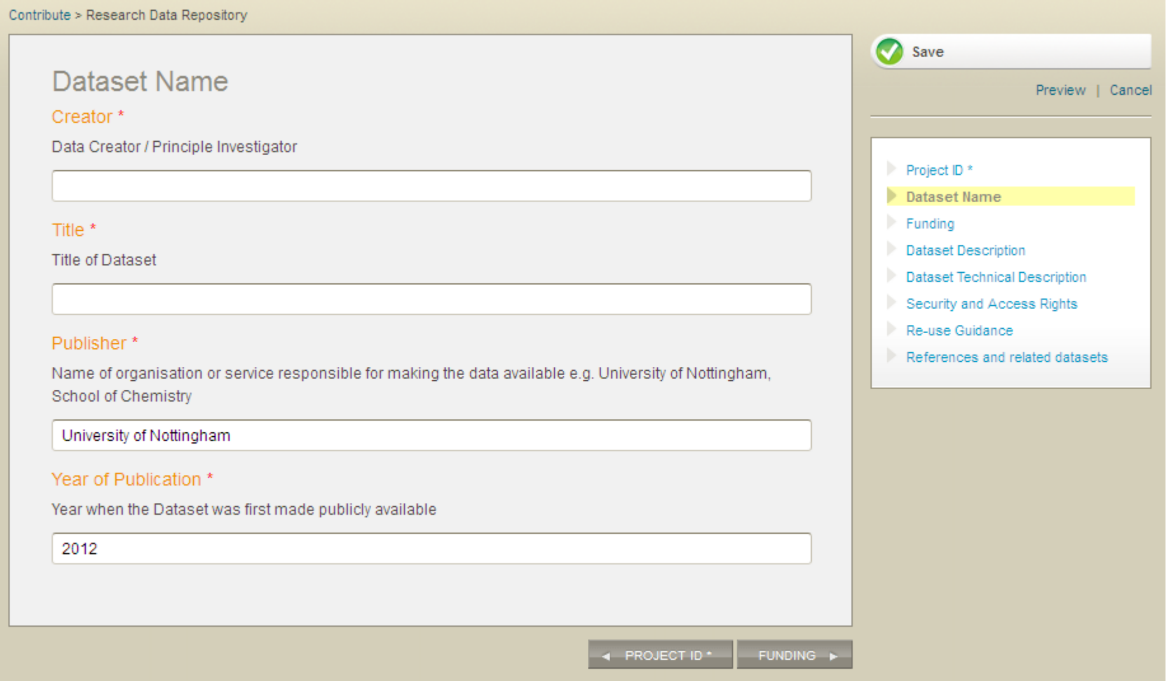
- Is this interface suitable for the use case of a researcher entering metadata for research datasets?
- What should be used as Project ID?

[3] https://dmponline.dcc.ac.uk/

- Are there multiple IDs that might be relevant, and if so, how should this be handled here?
- What about projects that don't have Agresso IDs?
- Is there a risk that researchers might not notice that automatically-populated Project metadata might need to be edited to give similar but slightly different metadata, specific to the dataset?

2.3. Dataset Name

This page (Figure 3) details DataCite's mandatory data fields, necessary for obtaining a DOI. Note that guidance notes can be given for each data field, and this guidance needs to be defined (what appears at present is just a first guess). Also note that mandatory fields are marked with an asterisk, and which fields should be mandatory will also need to be defined.



Contribute > Research Data Repository

Dataset Name

Creator *

Data Creator / Principle Investigator

Title *

Title of Dataset

Publisher *

Name of organisation or service responsible for making the data available e.g. University of Nottingham, School of Chemistry

University of Nottingham

Year of Publication *

Year when the Dataset was first made publicly available

2012

Figure 3: Dataset details

The Creator field is intended to map to the DataCite 'Creator' field, but this may either be the PI of the project, the Data Creator of the dataset, or a list of contributors to the dataset. Some of these fields are also defined on the Dataset Description page. It is unclear what convention should be used for the Creator field and unclear how multiple fields should be combined and mapped into Creator(s) in the Dublin Core metadata.

Crucially, the Title field should be the title **of the dataset**, not the title of the project, and it seems likely this will be a common mistake, especially if the project title is pulled into this field automatically.

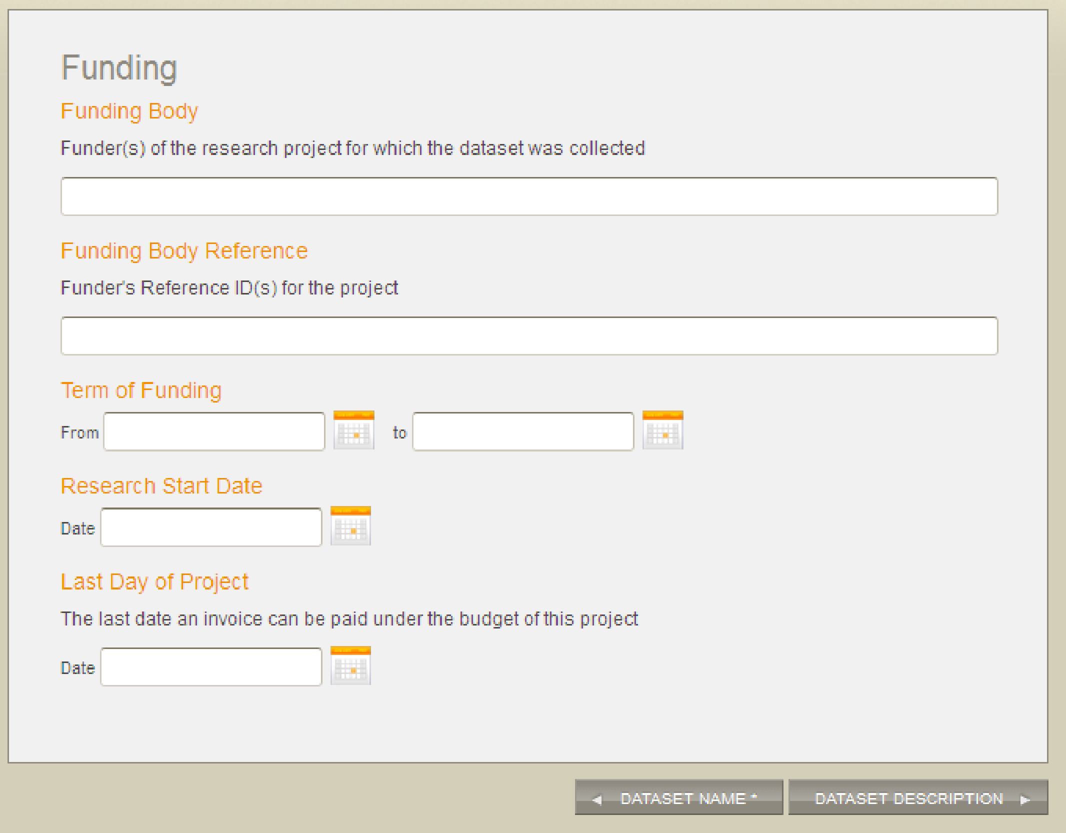
The Publisher may or may not be the University of Nottingham: if the dataset is being surfaced by another repository, perhaps that repository should be the publisher?

2.3.1. Questions

- Who will define the descriptions of the fields, throughout this wizard?
- Which fields should be mandatory?
- What convention should be used for DataCite's Creator field?
- Could/should the Creator field be mapped automatically by combining fields from the Dataset Description details?
- Should the project title be pulled in automatically?
- Who should the Publisher be, and should it be possible or necessary for the researcher to change this?

2.4. Funding

The Funding information on this page (see Figure 4) is not part of the Dublin Core specification, nor is it part of the DataCite information, but it is nevertheless key information for users accessing datasets for Open Access purposes - users may want to know who funded the project, when, in order to know what policies apply to the data. This Funding metadata appears to be (at present) a Nottingham extension to the metadata schema, but work is under way at CrossRef[4] to define a standard for funding metadata.



Funding

Funding Body

Funder(s) of the research project for which the dataset was collected

Funding Body Reference

Funder's Reference ID(s) for the project

Term of Funding

From to

Research Start Date

Date

Last Day of Project

The last date an invoice can be paid under the budget of this project

Date

Figure 4: Funding details

[4] http://www.crossref.org/

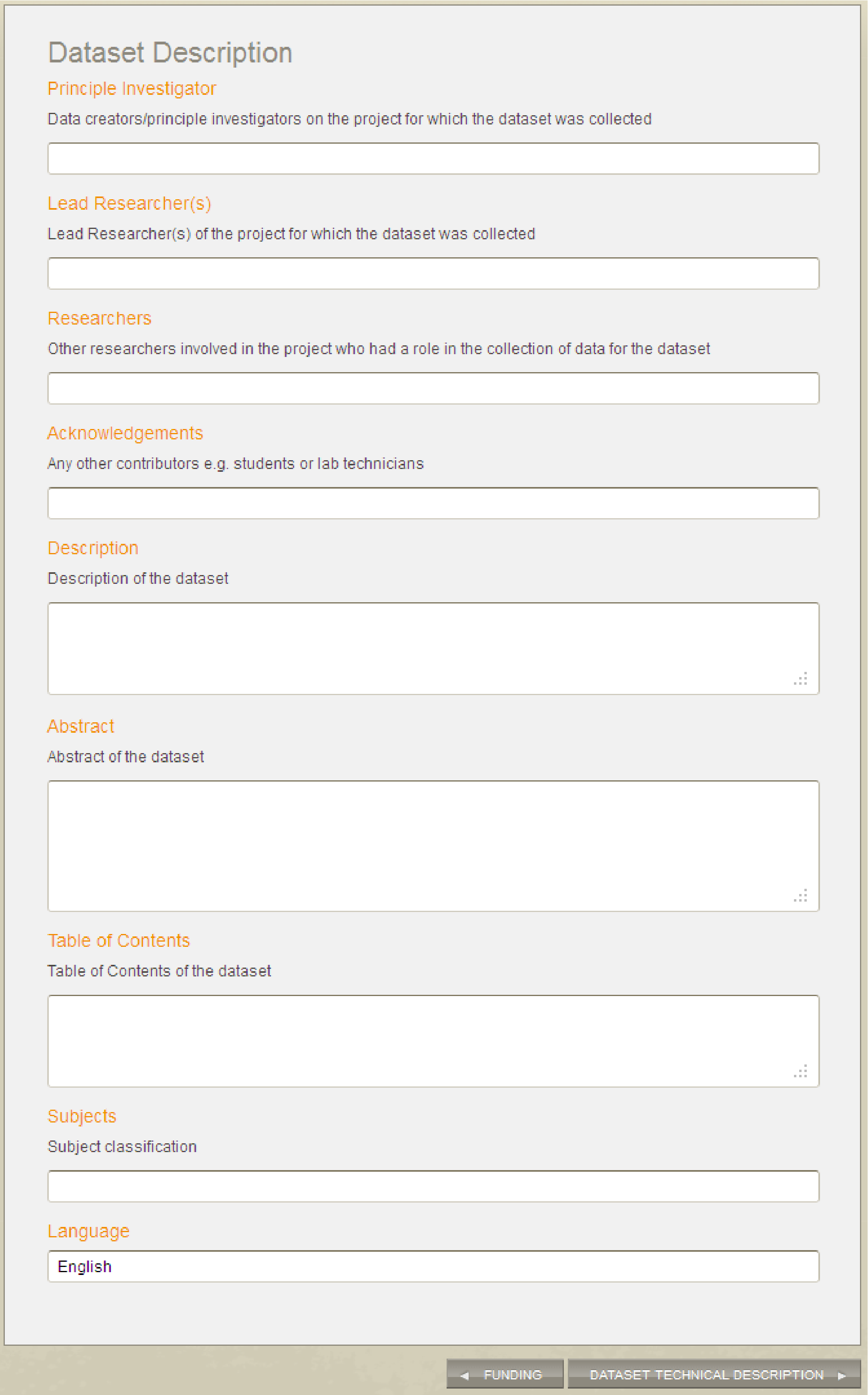
There may also be multiple funders involved: should the funder fields be entered as repeating fields and the data combined? If so, **how** should it be combined? The Funder name probably can't be just a drop-down with options, because the list of options can't be listed definitively - should there be a drop-down plus a free-text field, and if so, (how) can this be achieved in EQUELLA? Should the Reference ID field be checked for correct syntax?

2.4.1. Questions

- Should the Funding fields be defined within the Description field in Dublin Core, or as an extension to the metadata schema?
- Should the interface be able to handle multiple funders? If so, how should they be stored in the metadata - do they need to be combined into one metadata entry?
- Should there be a list of options for 'Funder' and if so, how can that be combined with a free-text option?
- Should the Reference ID field be checked for correct syntax?

2.5. Dataset Description

The Dataset Description page (see Figure 5) defines project description fields, applied to a dataset. "Dataset Description", like all the other page titles, is a first draft guess, and indeed the division of the metadata into separate pages is just a first attempt to subgroup similar fields together. All of this needs validation and further work.



Dataset Description

Principle Investigator

Data creators/principle investigators on the project for which the dataset was collected

Lead Researcher(s)

Lead Researcher(s) of the project for which the dataset was collected

Researchers

Other researchers involved in the project who had a role in the collection of data for the dataset

Acknowledgements

Any other contributors e.g. students or lab technicians

Description

Description of the dataset

Abstract

Abstract of the dataset

Table of Contents

Table of Contents of the dataset

Subjects

Subject classification

Language

English

Figure 5: Dataset description

The fields for the project staff need further definition. As mentioned previously, there is a question of how these fields should be combined into the 'Creator' field. Conventions for names of individuals are also needed, and in addition there are emerging standards for unique identifiers of academic staff to be considered.

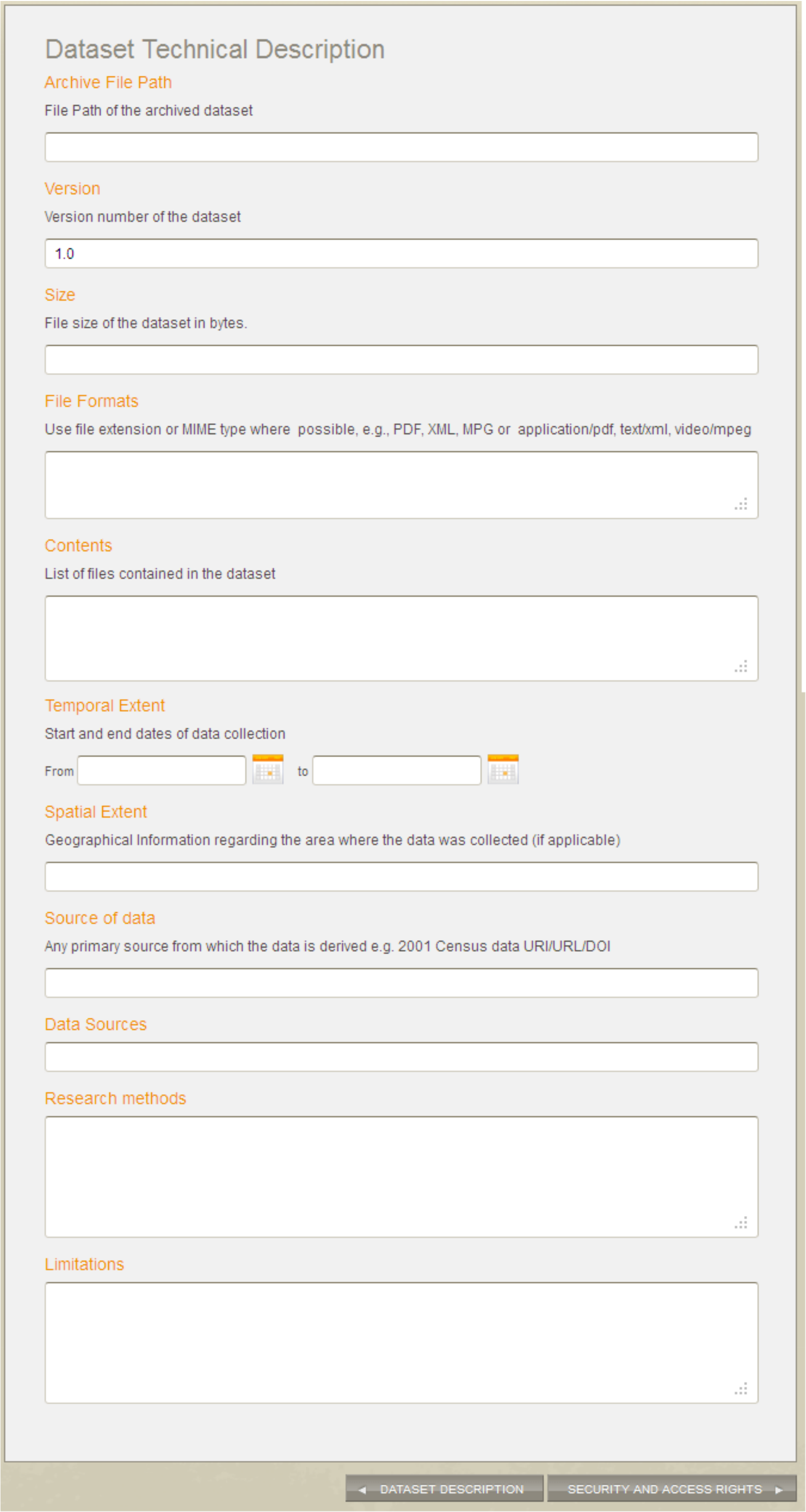
'Description' and 'Abstract' are both listed in the metadata specification, but it's unclear how these differ. What a 'Table of Contents' for a dataset would look like is also unclear (and this may overlap with other fields on other pages). 'Subject Classification' needs a taxonomy - EQUELLA can easily have defined taxonomy standards plugged in and these can appear as drop-down lists, so this needs to be defined for this field. Standards also need to be used for the 'Language' field.

2.5.1. Questions

- Who should define the division of the metadata fields into pages in this wizard?
- Who should define the names of the pages in the wizard?
- What convention should be used for names of individuals, and should this be enforced in code?
- Should IDs for staff members be used, what ID standard should be used, and how can this be plugged into EQUELLA?
- Are 'Description' and 'Abstract' different? Is only one of them needed? If both are needed, how should they to be described?
- What should the 'Table of Contents' refer to? Does this overlap with other fields on the next page?
- What taxonomy should be used for 'Subject Classification'?
- What taxonomy or standard should be used for 'Language'?

2.6. Dataset Technical Description

The 'Technical Description' page (see Figure 6) contains fields that are more specifically descriptive of the dataset itself. Again, the name of this page is a 'first guess' and needs further thought.



Dataset Technical Description

Archive File Path

File Path of the archived dataset

Version

Version number of the dataset

1.0

Size

File size of the dataset in bytes.

File Formats

Use file extension or MIME type where possible, e.g., PDF, XML, MPG or application/pdf, text/xml, video/mpeg

Contents

List of files contained in the dataset

Temporal Extent

Start and end dates of data collection

From to

Spatial Extent

Geographical Information regarding the area where the data was collected (if applicable)

Source of data

Any primary source from which the data is derived e.g. 2001 Census data URI/URL/DOI

Data Sources

Research methods

Limitations

Figure 6: Dataset technical description

The file path is intended to be the path to the archived location of the dataset itself. This field should be used to give the URL which end-users searching the repository can use to access the dataset. As such, its accuracy is crucial. How researchers obtain this archive location is unclear, and it's also not yet defined whether this should be a file path, or whether it needs to be a URL. This depends on how the archived datasets will be served to the public, and whether the URL can be constructed from the file path in a well-defined way. Can some kind of file chooser be used for this file path?

'Contents' seems to overlap with 'Table of Contents' on the previous page. For datasets with many files, this might be a long list, and for large datasets this field might be represented by something like an Excel spreadsheet - it can almost be a dataset in itself. It's also something that would normally be packaged automatically as a descriptor within the SWORD protocol; it probably needs to be generated automatically, somehow.

'Spatial Extent' needs to be expressed in one or more standard forms, and possibly also descriptively. This field can be used to automatically embed Google maps representing the location of collection of the dataset, so it needs to be expressed in the right format - what is that format, and should the format be checked here in code?

'Source of data' and 'Data sources' are both listed in the metadata specification, but appear to refer to the same concept, which needs further explanation. If this is a multiple field, are the multiple entries to be combined into a single field, or stored separately as a list?

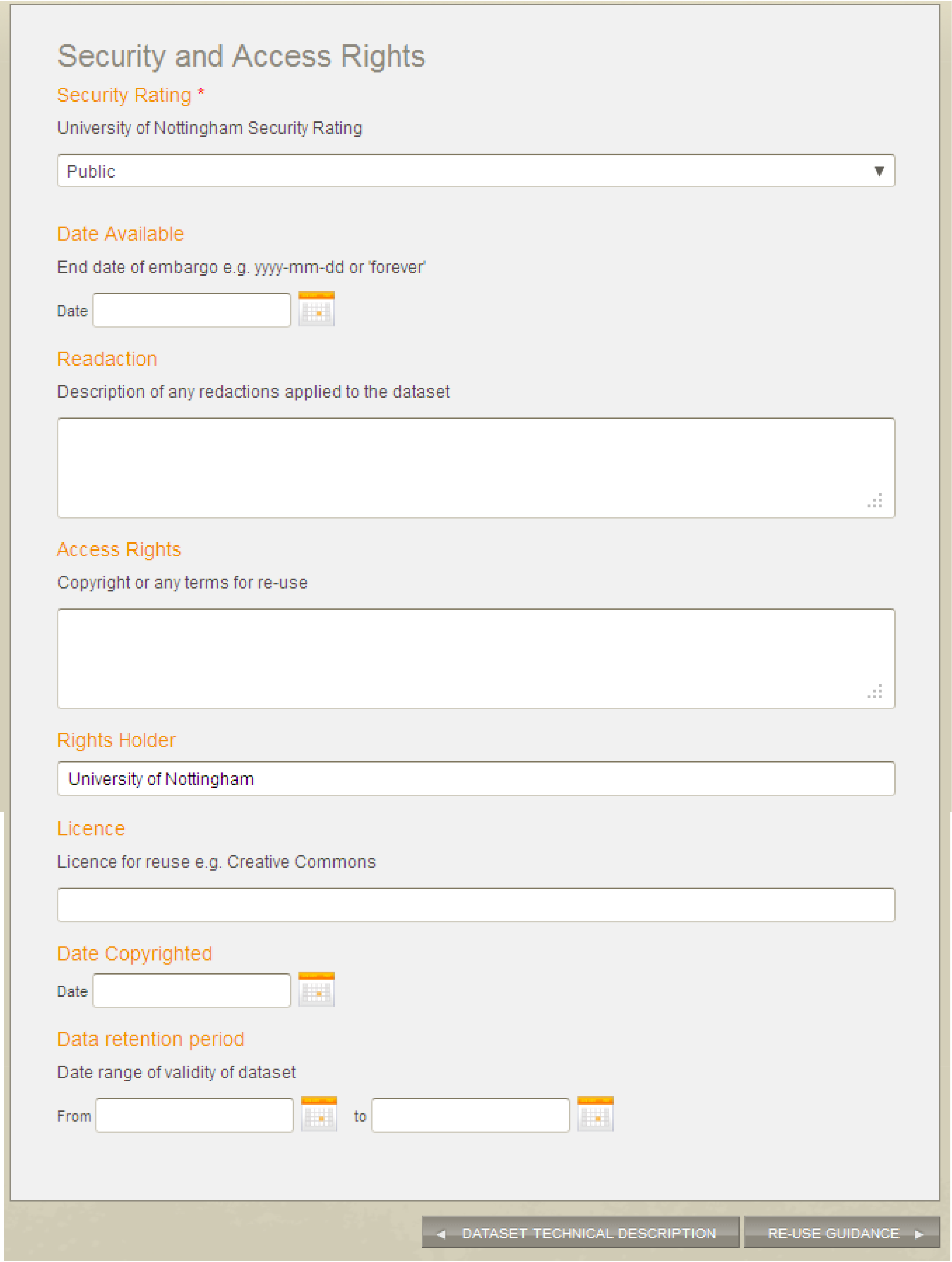
'Research methods' and 'Limitations' require description.

2.6.1. Questions

- What should this page be called?
- What definition should be used for the file path? File or publicly-visible URL? Can the latter be constructed from the former in a well-defined way?
- Is 'Contents' the same as 'Table of Contents'?
- How should a “Table of Contents” be created automatically when the dataset is packaged?
- How can the SWORD protocol be implemented if we aren't using the standard tools which implement it?
- What conventions need to be defined for 'Spatial Extent'? Should they be checked in code? Is this field also a multiple/repeating field?
- Is 'data sources' one field to replace 'Source of data', and if it's a multiple/repeating field, should the multiple entries be combined into one field in the metadata?
- What description is needed for 'Research methods' and 'Limitations'?

2.7. Security and Access Rights

This page (Figure 7) combines security rating, embargo details, redaction, access rights, intellectual property details and definition of the data retention period. As this is all crucial information, perhaps it should appear earlier in the wizard. Maybe it should be on separate pages.



Security and Access Rights

Security Rating *

University of Nottingham Security Rating

Public ▼

Date Available

End date of embargo e.g. yyyy-mm-dd or 'forever'

Date

Readaction

Description of any redactions applied to the dataset

Access Rights

Copyright or any terms for re-use

Rights Holder

University of Nottingham

Licence

Licence for reuse e.g. Creative Commons

Date Copyrighted

Date

Data retention period

Date range of validity of dataset

From to

Figure 7: Security and access rights

The University of Nottingham Security Rating is well-defined, but the options probably need to be explained to the end-user. These ratings probably map well onto the requirements for datasets, but that has not been confirmed. Security Rating, Data Available (Embargo) and Data Retention Period need to be used to automatically determine further actions concerning access to the dataset; further work is needed to confirm how this should work. Only Open datasets can obtain a DOI. How access to the data archive is controlled based on embargo date needs to be defined.

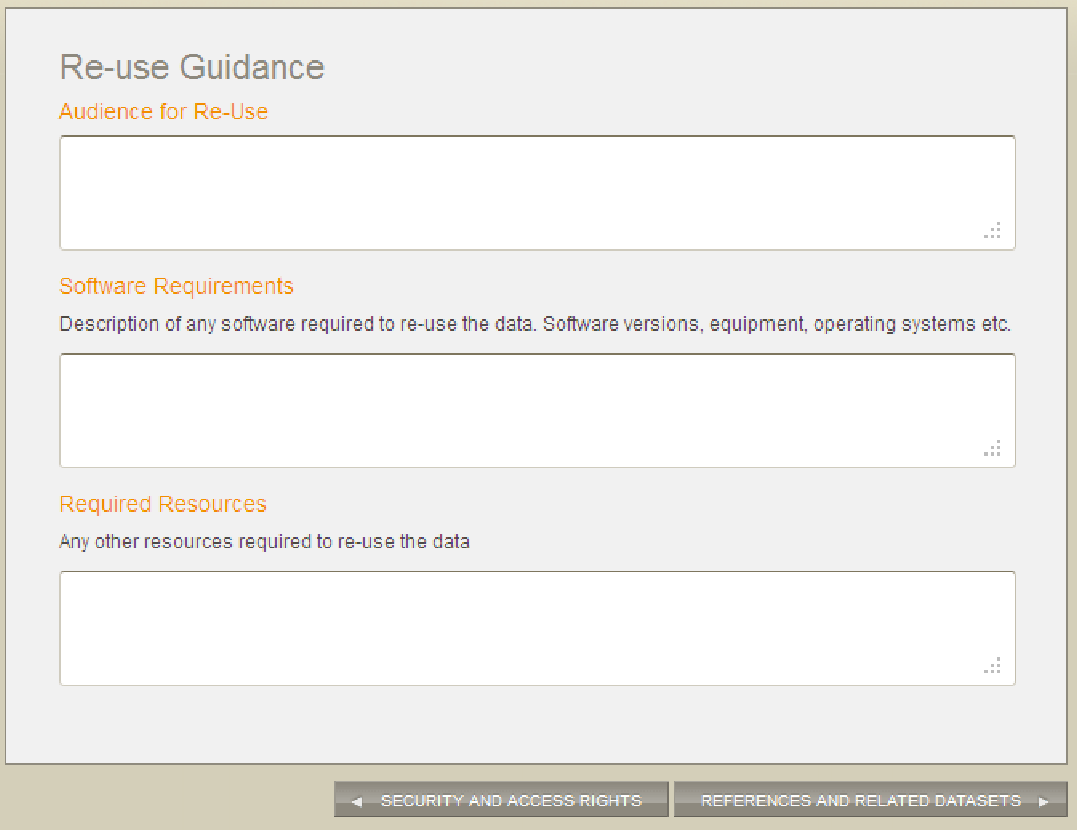
Licence could perhaps provide a drop-down of options; in final presentation to the end-user, a link to the definition of the licence terms is probably necessary.

2.7.1. Questions

- Should this page appear earlier in the wizard?
- Are different fields required or not required elsewhere in the wizard depending on the Security Rating? Are some fields that are mandatory for Open datasets non-mandatory for internal or confidential datasets?
- Are the Nottingham Security Rating fields sufficient for this purpose?
- How will access to embargoed datasets be controlled?
- What identifier is needed for non-open datasets? Is a Nottingham identifier scheme needed for these datasets, in addition to using DOIs for Open datasets?
- Is a well-defined list of options for Licence terms available? Can the licence agreement text itself be linked to? If so, what are the implications for how this data entry field should be defined?

2.8. Re-use Guidance

This page (Figure 8) details guidance to the end-user for re-using the data, including any pre-requisites. There are only three fields; possibly there are other fields that could be added here.



Re-use Guidance

Audience for Re-Use

Software Requirements

Description of any software required to re-use the data. Software versions, equipment, operating systems etc.

Required Resources

Any other resources required to re-use the data

◄ SECURITY AND ACCESS RIGHTS
REFERENCES AND RELATED DATASETS ►

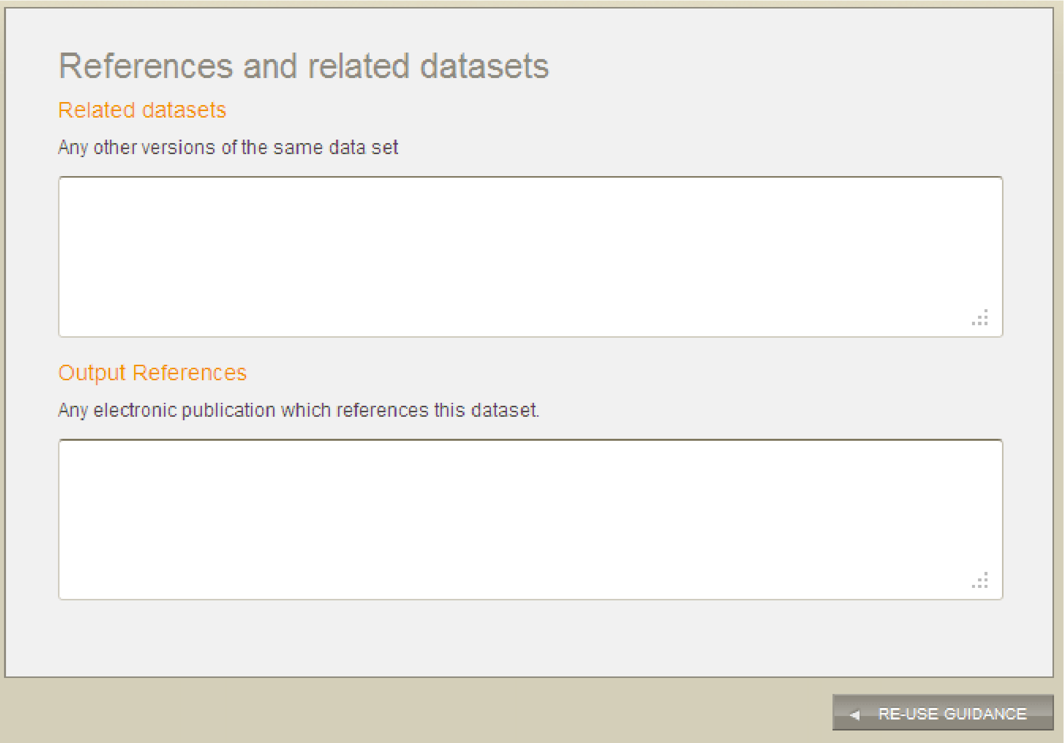
Figure 8: Re-use guidance

2.8.1. Questions

- Are there any other fields that need to be added to this page?

2.9. References and related datasets

This page (Figure 9) refers to other related datasets, and publications that reference the dataset. Several types of relationship are defined in Dublin Core, and perhaps they should be defined somehow on this page. References to other versions of the same dataset should probably use some unique identifier, so that should perhaps not be a free text field. It's not clear how the 'output references' can be determined at the time of publication of the dataset.



References and related datasets

Related datasets

Any other versions of the same data set

Output References

Any electronic publication which references this dataset.

Figure 9: References and related datasets

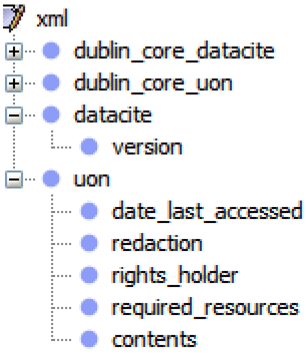
Crucially, there is a relationship between publications and datasets. The publication and datasets need to reference each other, both in the publication text and in the metadata associated with both, using well-defined identifiers. There may be a 'chicken and egg' problem here. It would perhaps be easier to connect publications and datasets if both were stored in the same repository, and if datasets were an additional feature of the publication record.

2.9.1. Questions

- Should multiple types of relationship be defined on this page?
- Are identifiers needed for related datasets, and should they be checked in code?
- How can 'output references' be determined at publication time?
- How should publications and datasets reference each other?
- Would this be easier to manage if publications and datasets were both catalogued in the same repository?

2.10. Metadata

Figure 10 shows the high-level structure of the metadata definition:



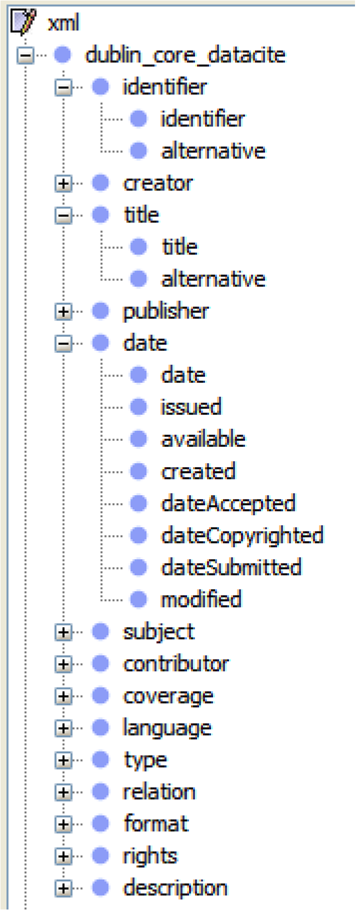
```
xml
  dublin_core_datacite
  dublin_core_uon
  datacite
    version
  uon
    date_last_accessed
    redaction
    rights_holder
    required_resources
    contents
```

Figure 10: Metadata

The first two categories are mapped to Dublin Core. These are separated at this stage to help identify the issues that need to be resolved regarding the mapping to Dublin Core. DataCite have provided a mapping of their mandatory and optional fields to Dublin Core, and this mapping is used in the dublin_core_datacite section. The University fields which are **not** already covered by that mapping are also mapped to Dublin Core, in the dublin_core_Nottingham section, but that mapping is not fully-defined and there are several questions to be answered about how that mapping should work. Ultimately, when those questions are answered, these two Dublin Core sections should be combined, and our Dublin Core schema should probably be defined at a more fundamental level in EQUELLA.

The other two categories define fields which do not have an equivalent in Dublin Core. DataCite defines a Version number of the dataset, but does not map that field to Dublin Core. Nottingham proposed metadata fields for Date Last Accessed, Redaction, Rights Holder, Required Resources and Contents have no obvious mapping in Dublin Core. Such fields can be defined in additional non-Dublin Core sections of the metadata schema, specific to research datasets, as has been done here.

Figure 11 shows some illustrative elements of the structure of the DataCite Dublin Core mapping:



- xml
- dublin_core_datacite
  - identifier
    - identifier
    - alternative
  - creator
  - title
    - title
    - alternative
  - publisher
  - date
    - date
    - issued
    - available
    - created
    - dateAccepted
    - dateCopyrighted
    - dateSubmitted
    - modified
  - subject
  - contributor
  - coverage
  - language
  - type
  - relation
  - format
  - rights
  - description

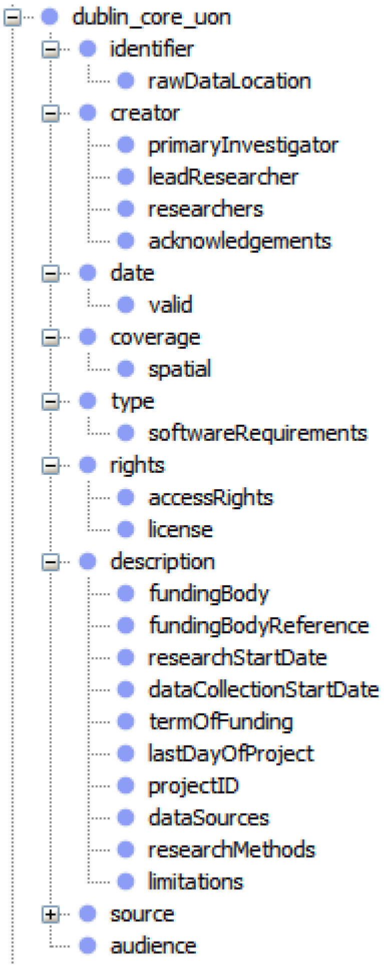
Figure 11: DataCite mapping

The top-level in this tree lists the Dublin Core elements, and within each element are Dublin Core terms. This may not be the right way to do this in EQUELLA, but only the bottom level elements in the tree can be mapped to data-entry fields in the wizard, so all DC elements have been expanded to include their subsidiary terms.

The mapping of these terms follows the definition in DataCite's own documentation, but it should be noted that, as a result, the names in this metadata definition are the names of the Dublin Core elements, and there is no definition within this EQUELLA pilot of how these terms are to be interpreted, i.e. what their meaning is within Dublin Core. For example, DataCite define that 'identifier' maps to 'DOI' and 'alternative' is some other identifier (perhaps an institutional ID). One has to know this mapping in order to make sense of the schema, and so the full mapping will need to be documented somewhere.

Some of the fields which are mapped to Dublin Core, it seems, are composite or multiple fields. For example, 'creator' maps to 'creators' and may perhaps need to be composed from a combination of other terms listed under the 'creator' element. 'Relation' lists related datasets - but how Dublin Core is intended to handle multiple entries in fields like this is not clear. These are key questions that need to be understood before the metadata schema in EQUELLA can be properly defined.

Finally, Figure 12 provides some illustrative aspects of the Nottingham Dublin Core mapping:



- dublin_core_uon
  - identifier
    - rawDataLocation
  - creator
    - primaryInvestigator
    - leadResearcher
    - researchers
    - acknowledgements
  - date
    - valid
  - coverage
    - spatial
  - type
    - softwareRequirements
  - rights
    - accessRights
    - license
  - description
    - fundingBody
    - fundingBodyReference
    - researchStartDate
    - dataCollectionStartDate
    - termOfFunding
    - lastDayOfProject
    - projectID
    - dataSources
    - researchMethods
    - limitations
  - source
  - audience

Figure 12: Nottingham Dublin Core mapping

It is suggested in the specification that the Raw Data Location (location of the archived dataset) can be mapped to Dublin Core's 'identifier' field. This field is already defined by DataCite for use for the DOI, and 'alternative' is proposed for institutional identifiers. If the raw data location is also going to use 'alternative', or if there are multiple 'alternatives' for 'location', then how can it be defined which is which?

The roles listed under 'creator' are project roles that are proposed to be mapped to the 'creator' element. But these terms - primaryInvestigator, leadResearchers, etc - are not defined Dublin Core terms. It is not clear how these terms can be used within a Dublin Core schema. If they are all to be combined into a single composite 'creator' field, then information will be lost and there seems little point in entering them as separate fields in the first place. If they are defined as separate fields, then again, how does DataCite handle multiple fields with the same term and how can one distinguish between these different roles and determine which of the 'creator' terms refers to the Primary Investigator? The implementation as shown in the images is an attempt to implement all the metadata fields required within a Dublin Core structure.

Some of the remaining terms, as shown in the mapping, are mapped using valid Dublin Core names, but others are not. For example, the 'spatial' term is a valid Dublin Core term under the 'coverage' element, which (unlike temporal coverage) is not a DataCite mandatory or optional field. On the other hand, the metadata fields which have been proposed as being mapped to the Description element are not Dublin Core terms. The 'description' element terms are illustrative of the issues regarding how these fields can be validly mapped to Dublin Core elements and terms while still preserving the meaning of these fields.

Thus, the metadata mapping implemented in this demo is a very partial and incomplete first attempt at mapping metadata fields to Dublin Core, but hopefully this work, like the wizard pages above, will be a useful first step in visualising the questions that need to be answered in order to take this project forward.

3. Further work for the project team

Lots of questions are listed below; here's a summary of some of the key issues:

- Is this interface suitable for end-users (researchers) to use, or would a separate data-entry web application need to be written as a front-end to an EQUELLA metadata store?
- What should be used as Project ID?
- Is there a risk that researchers might not notice that automatically-populated Project metadata might need to be edited to give similar but slightly different metadata, specific to the dataset?
- Who will define the descriptions of the fields, throughout this wizard?
- Who should define the division of the metadata fields into pages in this wizard?
- Who should define the names of the pages in the wizard?
- What taxonomies and format definitions are needed, and should these formats be verified in code?
- What definition should be used for the resource file path? File path, or publicly-visible URL? Can the latter be constructed from the former in a well-defined way?

- Could a 'Table of Contents' be created automatically when the dataset is packaged?
- How can the SWORD protocol be implemented, to enable researchers to package the dataset, if we aren't using standard tools which implement it? Should our own datasets be packaged, using metadata from EQUELLA, at the time when they are archived? How does this requirement fit in with the architecture defined thus far?
- How will access to embargoed datasets be controlled?
- What identifier is needed for non-open datasets? Is a Nottingham identifier scheme needed for these datasets, in addition to using DOIs for Open datasets?
- How should publications and datasets reference each other?
- Would this be easier to manage if publications and datasets were both catalogued in the same repository?