Work Package 3: Metadata Development and Standardisation

Project Information			
Project Identifier	PID TBC		
Project Title	UK Research Data (Metadata) Registry Pilot		
Project Hashtag	#jiscrdr		
Start Date	1 October 2013	**End Date**	31 March 2014
Lead Institution	Jisc		
Project Director	Rachel Bruce		
Project Manager	—		
Contact Email	—		
Partner Institutions	Digital Curation Centre (Universities of Edinburgh, Glasgow, Bath); UKDA (University of Essex)		
Project Webpage URL	http://www.dcc.ac.uk/projects/research-data-registry-pilot		
Programme Name	Jisc Capital Programme		

Document Information			
Author(s) and Role(s)	Alex Ball (Metadata Coordinator)		
Date	13 May 2014	**Project Refs**	WP3
Filename	uk-rdr-wp3-report-v01.pdf		
URL	http://www.dcc.ac.uk/sites/default/files/documents/registry/uk-rdr-wp3-report-v01.pdf		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
01	13 May 2014	Initial version.

One of the purposes of this work package was to agree a metadata standard for use with the registry pilot. The decision to use the ANDS ORCA software as the platform for the registry considerably simplified this task: since ORCA is set up by default to use RIF-CS as its underlying metadata scheme, the effort required to switch it to another would have been greater than any potential benefit. Indeed, working with RIF-CS has shown it to have some admirable qualities:

- with its four-entity data model, it has some of the power of a highly structured and relational scheme like CERIF while having a much gentler learning curve;
- it is has a minimal number of required fields, meaning that records can be generated even if little information is available to populate them;
- it allows information to be recorded in a structured way, with semantic markup and controlled vocabularies, making it easier for automated tools to interact with it downstream, but provides distinct ways of providing unstructured information if that is all that can be harvested.

Even so, the process of using RIF-CS in practice raised some questions that need to be resolved.

- *Controlled vocabularies.* RIF-CS uses a large number of controlled vocabularies, which has its advantages for downstream reuse of the information, but can be problematic when the vocabulary is not expressive enough for a given context. For example, we found that the (meta) vocabulary used for subject vocabularies, while extensive, did not contain terms for the subject schemes used by our data contributors. This left us with the dilemma of whether we should follow the standard exactly, and thereby lose knowledge of which schemes were used under the catch-all 'local' term, or extend the vocabulary with additional terms. We decided to take the latter approach, on the understanding that if our extensions were used in the eventual Jisc service, we should attempt to feed our terms back into the RIF-CS standard.
- *Dates.* RIF-CS allows many different types of date to be recorded. Some of them have overlapping semantics, e.g. 'published', 'available', 'issued'. In such cases it is not clear when harvesting a date whether to represent it using all three dates or just one, when the semantics appear to match all three. The former approach might assist downstream applications, but also appears to clutter the interface of the registry application: indeed, it seems the generated sample citation makes use of all available dates in a rather unhelpful manner.
- *Links.* In contrast, RIF-CS does not provide a way of distinguishing different types of links, beyond treating an alternative metadata record for the object as a separate, related entity. Thus there is no way of distinguishing a direct download link from a landing page link if both are provided. The registry software does, however, know how to generate a link from a plain DOI, so if the DOI in link form is separately provided, this leads to duplication of the link in the web page for the record.

The other objective for this work package was to develop crosswalks from the metadata schemes used by our data contributors to the one used by the registry. This was a three stage process:

1. We had to decide which metadata schemes to support, as several data contributors could supply metadata according to multiple schemes.
2. We determined how, in theory, a RIF-CS record should be populated from information encoded in these supported metadata schemes. First, we matched the semantics of

elements in RIF-CS with those of a particular source scheme. Then, we determined how to transform information encoded according to the conventions of the source scheme into the formats and vocabulary terms expected by RIF-CS.

3. We wrote PHP crosswalks that instantiated these transformations and tested them against sample records provided by our data contributors. In many cases we had to write more complex transformations than originally planned in order to support real-world usage of the source metadata schemes.

The schemes we chose were as follows:

- DDI Codebook 2.5, the most detailed scheme available from the UK Data Archive;
- UK GEMINI 2 (version 2.2 was used as reference), the scheme used by the NERC Data Catalogue Service;
- OAI Dublin Core (oai_dc), the fallback scheme supported by all OAI-PMH endpoints;
- The EPrints metadata export scheme, with support for extensions provided by the ReCollect plugin and other local variations used by our data contributors;
- DataCite (version 3 was used as reference), used by the Archaeology Data Service and other repositories registering DOIs for their datasets, but in practice only available from DataCite's own OAI-PMH endpoint;
- MODS (version 3.5 was used for reference), a scheme commonly available from DSpace-powered OAI-PMH endpoints.

CERIF was considered, but as support for datasets was still immature in CERIF, and there were few sample records with which we could work, we felt it would be better to postpone writing a CERIF crosswalk until a later phase.

In the majority of cases we found it possible to create high quality RIF-CS records from harvested metadata, even with OAI Dublin Core, our simplest source scheme. While much depended on how much information was actually provided, of course, in many cases records only missed ORCA's Quality Level 3 (the highest) due to a missing association with an activity (project) record. The more permissive and flexible the source scheme, though, the more work had to be done in the crosswalk to achieve a high quality record. An extreme example of this is the oai_dc 'coverage' element, which can contain geospatial or temporal information in any number of formats. In such cases we chose to write an algorithm for determining the type of information and represent it appropriately, rather than ignore the field, as we felt it was better to risk the occasional mistake than to lose the information entirely.

There were some instances where we could not in practice complete the mappings as originally intended. Both DDI and UK GEMINI can record information about related publications, but they do so in a free-text manner; this makes it hard to extract the title and identifier information needed to record the information in RIF-CS. We did not have the resources to write a reliable routine to achieve this. Furthermore, several schemes can record funder information, but without information about the activity (project) through which the dataset was funded it is not possible to record this in RIF-CS in a meaningful way. It may be hoped that in a future development of this service more comprehensive information might be harvested from Current Research Information Systems.

There were some technical challenges to implementing the crosswalks. The relevant functionality was newly implemented in ORCA and there were some coding issues that needed to be addressed before the crosswalks could be written. Also, since the crosswalks work on a whole XML file level, rather than on extracted records, they had to take account

of how records were harvested as this affects if and how they are ‘wrapped’ in XML. Lastly, the crosswalks could not be tested on a live OAI-PMH harvest, as the OAI harvest method provided by ORCA was hard-coded to request RIF-CS metadata.

Concentrating on record quality meant we did not have the resources to address some other points of concern we had, such as how to handle record deletions, and how to deal with records that might be harvested from several sources. We believe the ORCA APIs would be useful in finding appropriate solutions, but have not been able to verify this in practice.

Another point we would like to consider in a future phase is whether the workflow used by ORCA – harvesting metadata and transforming it to RIF-CS before storing it – is optimal. A different workflow to consider would be to harvest and store the ‘original’ metadata records and normalise them either on the fly or as a regular internal registry operation. That way, if improvements are made to the crosswalk, the benefits can be realised immediately instead of waiting for the next harvest.

In summary then, we found it possible to normalize metadata from a wide variety of data sources into high quality records that could be used to operate a data discovery service. There are certainly some quirks to be addressed, and the solutions we have found may not be robust enough in their current state to cope with all possible implementations of the chosen metadata schemes. Nevertheless, the progress we have made in this short phase bodes well for future development.