In Phase 1 we chose the path of least resistance and used RIF-CS for our registry.

We are not committed to RIF-CS (but it works quite well)

It is not too well known, so here is a quick overview.

- Profile of ISO 2146 (Information and Documentation – Registry Services for Libraries and Related Organizations)
- Optimized for collection services registries
- Maintained by ANDS: see http://services.ands.org.au/documentation/rifcs/1.5/guidelines/rif-cs.html
- ‘Gateway drug’ for CERIF?

By which I mean it moves you away from thinking in terms of a single flat metadata file and starts you thinking about relationships between different entities.

There are only four entities, but they are specialised with types (Figure 1).

![Figure 1: Example entities from the RIF-CS data model](image)

With these you can build up a quite detailed network of records (Figure 2).
Not just about elegance or efficiency: these relations are also browsing pathways.

Whose repository can supply RIF-CS metadata? None. And since at this stage we are not committed to RIF-CS, we couldn’t impose it on our collaborators; that meant performing the crosswalks centrally instead of at each individual site.

So we had to write crosswalks to harvest records in formats that were supported. OAI DC is a fallback that most repositories should support, but we also wanted to benefit from more detailed metadata that many repositories might be able to provide.

<table>
<thead>
<tr>
<th>DataCite 3</th>
<th>EPrints 3</th>
<th>OAI-PMH Dublin Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Archaeology Data Service</td>
<td>- Glasgow</td>
<td>- Oxford Brookes</td>
</tr>
<tr>
<td>- Oxford</td>
<td>- Leeds</td>
<td>- Lincoln</td>
</tr>
<tr>
<td>- Southampton</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DDI Codebook 2.5</th>
<th>MODS 3.5</th>
<th>UK Gemini 2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>- UK Data Archive</td>
<td>- Edinburgh</td>
<td>- NERC Data Catalogue Service (incl. ADS)</td>
</tr>
<tr>
<td></td>
<td>- St Andrews</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Hull</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3 gives an idea of how these work. On the left is an OAI-PMH ListRecords return, and on the right, the RIF-CS XML.

For example, the ‘request’ value (= URL of OAI-PMH endpoint) is used twice:

- group = name of organisation contributing the record (i.e. translated from a URL to a text string).
- originatingSource = ID of organisation holding ‘master copy’ of record (this would be overwritten if the metadata record specifies this explicitly).
You see that it makes a difference whether the record is in an OAI-PMH wrapper or not. So that is why we have two different DataCite crosswalks:

HTTP
- DataciteToRifcs (Single XML record)
- EprintsToRifcs (EPrints XML export)
- Gemini2p2ToRifcs (CSW)

OAI-PMH
- Datacite3ToRifcs
- Ddi2p5ToRifcs
- Mods3ToRifcs
- OaiDcToRifcs

Built into the registry are some automated quality checks that ensure that records have enough useful information.

Quality Level 2
- title
- description
- location (e.g. URL)
- IPR statement
- related party, e.g.
  - P.I./researcher
  - manager

Quality Level 3
- identifier
- citation information*
- subject
- date (e.g. of publication)
- spatial coverage
- temporal coverage
- related activity

* Such as ‘publisher’; other relevant fields are already mentioned.

So what did we learn from this?
- RIF-CS can handle
– ‘stub’ records with minimal information (all we really need is ‘group’, key, and what type of entity it is);
– structured information in structured way;
– unstructured information in unstructured way (e.g. an untyped name parts or several types name parts; full citation or citation metadata).

• We needed to expand the controlled vocabulary for subject schemes, to be able to identify terms from GEMET, HASSET, etc.
• RIF-CS does not describe what web links do unlike ISO 19115 which distinguishes functions like ‘download’, ‘order’, ‘information’.
• Parties need IDs too. If not supplied, we have to generate them, and we can’t guarantee one-to-one mapping.
• There’s no specific, direct relation between a funder and a dataset (it goes via the grant). We could record an arbitrary relation and describe it in English, but that isn’t very Semantic Web.

¶ But there are some questions to which we still need answers.

• Harvesting a new version of a record (as determined by the object key) replaces the old one.
  – How do we merge into the old one? (Important for generated records)
  – How do we conditionally replace the old one? (Important for preferring one source over another)
  – How do we handle deletions?
• Which dates do we really need? ‘Published’, ‘issued’ and ‘available’ are all very similar, do we need them all? They are held separately in the DataCite schema but not in the other standards.
• How do we get ‘boilerplate’ information from user accounts instead of OAI-PMH headers?
• How do we harvest from CRISes in CERIF format?

¶ Now, thinking about the requirements from the eventual service...

We are not committed to RIF-CS; would something else work better?

Alex Ball. DCC/UKOLN, University of Bath.