CERIF for Datasets:
Linking and contextualising publications and datasets, and much more …

Scott Brander, Anna Clements, Valerie McCutcheon, Paul Cranner, Ryan Henderson, Kevin Ginty

First Workshop on "Linking and Contextualizing Publications and Datasets", 26th Sep 2013
17th International Conference on Theory and Practice of Digital Libraries: TPDL2013
Sep 22-26, 2013, Valletta, Malta
Many thanks to Nikos Houssos for presenting on our behalf

Slides prepared by Anna Clements with thanks to euroCRIS colleagues: Keith Jeffery, Brigitte Joerg and Jan Dvorak
**C4D Summary**

- JISC Managing Research Data Programme
- Consortium: Sunderland, Glasgow, St Andrews, NERC, EPSRC, DCC and euroCRIS
- “CERIFication” of the *metadata* about research datasets
- Focus on MEDIN* standard: NERC requirement for [http://www.bodc.ac.uk/](http://www.bodc.ac.uk/)

* [http://www.oceannet.org/](http://www.oceannet.org/)
In the UK: The CERIF landscape
CERIF basics

- **Common European Research Information Format**
- A conceptual model for describing the complete research domain
- A standard for the development, implementation and interoperability of current research information systems (CRIS) and their various application
- Est. 1991; maintained by [www.euroCRIS.org](http://www.euroCRIS.org)
- Ongoing work with OpenAire, DataCite, RD-Alliance, ORCID
**euroCRIS basics**

- Not for profit organisation of experts
  - Research organisations; funders; publishers; systems providers; standards organisations

- 109 institutional, 38 personal & 20 affiliate members (*euroCRIS annual report 2012*)

- 41 countries; not just Europe

- Main activity: development, maintenance and implementation of CERIF

- Multiple strategic partners, e.g. VIVO, COAR, CODATA, CASRAI, and others
Linking and Contextualising Datasets and Publications, TPDL2013, Sep 22-26, Malta

CERIF evolution

CERIF 1.5
CERIF 1.4 (XML)
CERIF 1.3

- Data Model
- Infrastructure
- Facility, Equipment, Service
- Measurement & Indicator
- Entities and Link Tables
- Geographic Bounding Box
- CERIF 1.3 Vocabulary
- UUIDs
- Terms
- Schemes
- CERIF 1.4 new XML format
- CERIF 1.5 Federated Identifiers

+ Linked Data

CERIF 1.6

- Data Model
- C4D datasets

CERIF 2006 / 2008 Model

- Data Model
- Model Normalization
- Robust/Consistent Structure
- Extensible Structure
- Semantic Layer
- XML Exchange Specification
- Elaboration on Publication
- CERIF Core Semantics (2008 1.2)

CERIF evolution

Acronym: ERGO
Participants: Keith Jeffery, Anne Asserson, Rutherford Appleton Lab, Univ Bergen, many more

- Networking of DBs
- Exchange of Records
- EC Recommendation to Member States

CERIF 2000 Model

- Data Model
- Multilinguality
- Controlled Vocabulary
- Roles / Types
- User-driven
- EC Recommendation to Member States

CERIF 2000

- Networking of DBs
- Exchange of Records
- EC Recommendation to Member States

CERIF components

**CERIF Entity Types**
- Base Entities
- Result Entities
- Infrastructure Entities
- 2nd Level Entities
- Link Entities

**CERIF Features**
- Multiple Language
- Semantics
- Measures & Indicators
- Geographic Bounding Box
### Mapping MEDIN to CERIF 1.5

<table>
<thead>
<tr>
<th>MEDIN</th>
<th>DataCite v3.0</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Identifier</td>
<td>M</td>
</tr>
<tr>
<td>1</td>
<td>Resource Title</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Alternative Resource Title</td>
<td>Not supported – proposed to CERIF Task Group</td>
</tr>
<tr>
<td>3</td>
<td>Resource Abstract</td>
<td>R (descriptive)</td>
</tr>
<tr>
<td>4</td>
<td>Resource Type</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td>Resource Locator</td>
<td>R</td>
</tr>
<tr>
<td>6</td>
<td>Unique Resource Identifier</td>
<td>R</td>
</tr>
<tr>
<td>7</td>
<td>Coupled Resource</td>
<td>O</td>
</tr>
<tr>
<td>8</td>
<td>Resource Language</td>
<td>O</td>
</tr>
<tr>
<td>9</td>
<td>Topic Category</td>
<td>R (subject)</td>
</tr>
<tr>
<td>10</td>
<td>Spatial Data Service Type</td>
<td>O</td>
</tr>
<tr>
<td>12</td>
<td>Geographic Bounding Box</td>
<td>R (envelope)</td>
</tr>
<tr>
<td>13</td>
<td>Extent</td>
<td>R</td>
</tr>
<tr>
<td>14</td>
<td>Vertical Extent Information</td>
<td>Not supported in CERIF. CERIF has a GeoBox element which can be used to record these attributes, but there is currently no cResProd_GeoBox linking element.</td>
</tr>
<tr>
<td>15</td>
<td>Spatial</td>
<td>R</td>
</tr>
</tbody>
</table>

| M | Temporal Reference | M | cResProd_Class.cfClassSchemel with temporal reference classification scheme |

| Spatial Reference | M (publication start) | O (date or period of collection) |
| Spatial Resolution | R (notes e.g. method of collection) |
| Limitations on Public Access | O (high) |
| Conditions applying for access and use | O (high) |
| Responsible party | M (Creator, Animator) |
| Data Format | O |
| Frequency of Update | cResProd_Class.cfClassId with Frequency of Update classification scheme |
| Conformity | cResProd_Measurement.MeasId |
| Metadata Date | This is managed by the application |
| Metadata Standard Name | No recommendation by CERIF Task Group, so was mapped to cOrgUnit_cResProd with linking roles |
| Metadata Standard Version | As per 27 |
| Metadata Language | Is cfLang entity but no link to cResProd currently |
| Parent ID | R (related identifier) |

**Notes:**
- Not currently supported – proposed
- Recommendation is cResProd_GeoBox
- Not currently supported – proposed free text
- CERIF TG still discussing
## Representing temporal information: TG discussion

<table>
<thead>
<tr>
<th>Element No.</th>
<th>Element Name</th>
<th>CERIF Entity</th>
<th>Vocabularies/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Temporal Reference</strong></td>
<td><code>cfResProd.cfResProd_Comp_class.cfClassSchemeId</code></td>
<td>Temporal Extent Scheme (The period the data is related to)</td>
</tr>
</tbody>
</table>

### C4D CERIF

<table>
<thead>
<tr>
<th><code>&lt;cfResProd&gt;</code></th>
<th><code>&lt;cfResProdId&gt;02086348-7c61-4be1-8976-2003af66502&lt;/cfResProdId&gt;</code></th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;cfResProd_Comp&gt;</code></td>
<td><code>&lt;cfClassId&gt;temporal_extent&lt;/cfClassId&gt;</code></td>
</tr>
<tr>
<td><code>&lt;cfClassSchemeId&gt;class_scheme_resultProduct_classification_temporalReference&lt;/cfClassSchemeId&gt;</code></td>
<td></td>
</tr>
<tr>
<td><code>&lt;cfStartDate&gt;2010-01-01T00:00:00&lt;/cfStartDate&gt;</code></td>
<td></td>
</tr>
<tr>
<td><code>&lt;cfEndDate&gt;2010-01-02T00:00:00&lt;/cfEndDate&gt;</code></td>
<td></td>
</tr>
</tbody>
</table>

### Notes

- Four dates supported in C4D for this element: **temporal extent** (which has a start and end date), created (single date), revised (single date), published (single date).

  - `cfClassId` – one of: **temporal_extent**/publication/revision/creation
  - `cfClassSchemeId` – always: `class_scheme_resultProduct_classification_temporalReference`

### CERIF TG Notes

The CERIF TG suggests the **temporal interval** of the data itself (the effective datetime range of the observations) is of a different nature than the documentation of the dataset lifecycle. We would therefore suggest:

1. Expressing the **temporal extent** as two links to `cflMeasurements` that hold the `startDateTime`/`endDateTime` of the **temporal** reference.
2. Putting the creation/revision/publication timestamps as `cflDates` on the links to the parties responsible for the respective steps in the dataset’s lifecycle.
Conclusions – what worked well

- Mapping to CERIF pretty straightforward – because it already contains all the entities we need and most of the relationships
- Involving CERIF-TG meant we could give and take ideas – very constructive
- Modelling at the business level first helped resolve questions such as ‘should this be a classification or a relationship to a person or organisation’; this is best practice anyway for sustainability and flexibility; why model a person or an organisation as an attribute … rather than separate entities; this is a fundamental fault with DC and similar ‘flat’ structures
- The separate “semantic layer” ie the classification schemes, allowed us to map different schemes (inspire themes, rcuk subject classifications for keywords) and seavox gazetteer for Extent (ie which bit of ‘water’)
Conclusions – more work needed

- Some MEDIN elements not fully modelled yet but tend to be full text fields so could be better to determine if can be broken down into more structured data, e.g. Lineage (element 17)

- Agreeing semantics e.g. lifecycle stages of a dataset in order to properly model temporal aspects, such as published date, version date, created date, etc.

- Translating conditions of (re)use into structural metadata (element 21); requires modelling at business level first

Actually, these aren’t really issues with CERIF, more on business modelling and agreement on semantics and vocabularies …. Irrespective of data format.
Many thanks for listening

Anna Clements, Head of Research Data and Information Services
akc@andrews.ac.uk  @annakclements

C4D Blog at http://cerif4datasets.wordpress.com/