Cerif4Datasets (C4D) – Utilising Semantics for the Discovery and Exploration of Datasets in Research

Albert Bokma\textsuperscript{a}, Sheila Garfield\textsuperscript{a}, David Nelson\textsuperscript{a}, Esraa Omran\textsuperscript{a}, Oscar Corcho\textsuperscript{b}

\textsuperscript{a} University of Sunderland, UK
\textsuperscript{b} Polytechnic University of Madrid, Spain

Summary

The aim of the Cerif4Datasets (C4D) project is to use CERIF to capture the metadata of research datasets, and integrate this metadata with that held on research projects and research outputs available on a central CERIF cloud. CERIF has so far concentrated on the core entities of organisations, people and publications and has some provisions for research outputs such as patents but is not fully equipped to support datasets, especially when there is a need to make these readily discoverable. C4D therefore aims to explore the capabilities of CERIF and extend where required to support metadata about research datasets from a variety of disciplines building on existing standards where possible and develop a reference implementation that makes use of semantic technologies for improved discovery and navigation of key entities associated with datasets such as people, organisations and publications. In this paper we explore the issues related to suitable metadata standards, classification systems and propose an architecture to afford the required discoverability and navigation layer.

Keywords: datasets, metadata standards, discovery, semantic web, ontologies, classification systems

1 Introduction

The goal of the CERIF for Datasets (C4D) project is to extend CERIF to effectively deal with research datasets. Datasets for the purposes of this paper are collections of data which are usually structured and in tabular form where each row represents a record with a number of fields that hold data on various attributes and that have been generated in the course of research activity, whether they record experimental results or original observations.

CERIF in its core entities has a current focus to record information about people, projects, organisations, publications, patents, products, projects, facilities, and equipment. One aspect which is not specifically covered, though becoming of increasing interest is research datasets as a form of research output. In some disciplines there are considerable amounts of datasets generated which are an important output; these may be of use also to other researchers as a useful input to their research and potentially to help improve research and speed up the rate of discovery and be in itself a useful indicator of impact. Transparency and access are increasingly a feature of the public sector and making results more readily available and discoverability are now the order of the day; especially when these were generated with the assistance of public funds there is also a moral obligation to make results available for inspection and reuse. In addition to the moral
argument there is also increasing legislation in this direction that places specific requirements on researchers and their host organisations.

In Europe one such development has been the INSPIRE Directive on geospatial data which makes it a requirement that datasets in that area be preserved and made accessible. In the UK research council fundholders are now required to make these accessible, develop metadata about them as well as provide Data Management Plans to either curate the datasets themselves or hand them over for curation in dedicated repositories. Entire datacentres, curating and making datasets accessible are appearing more and more and so in the UK the Engineering and Physical Sciences Research Council (EPSRC) as well as the Natural Environment Research Council (NERC) already operate seven datacentres between them. The appropriate management of metadata for datasets is therefore becoming an important topic and there are a number of issues associated with this:

- How datasets can be supported in CERIF alongside other entities and forms of output
- What should these metadata records comprise
- What standards should be used for consistency and interoperability in a distributed environment
- How can the actual datasets be made discoverable especially also across disciplines

Despite similarities and overlaps to other forms of research output, datasets are significantly different to require special treatment. Datasets are often (though not always) generated in the course of research activity, whether as part of funded projects or in the course of the work of individual researchers and research teams. Datasets are an important kind of research output and are also an important input. Significant numbers of datasets are being generated and need to be managed for a number of reasons, including preserving important data for the future, potentially enabling better research, accelerating the rate of discoveries, and assessing and demonstrating impact.

These data collections, beside the activity out of which they arose, can be very useful for future research also by other researchers; they present a formidable resource if properly managed. With the appropriate safeguards in place for the protection of the results and proper acknowledgement and, if necessary, licensing, they have the potential for propelling research. However, due to lack of concrete knowledge as to who holds which datasets in the relevant research community this means that their potential remains currently underutilized. The access to these would be useful for researchers and would have additional benefits for the sharing parties by demonstrating the impact they create. Consequently, there is a need for recording these with easy to use tools and publishing them in a way that facilitates their discovery. To enable these benefits to be realised poses some specific problems:

- The generation of metadata is not always easy and intuitive and can become onerous once lots of datasets are generated on a regular basis and tools are needed to make sure that correct and appropriate metadata can be generated by researchers or support staff. This may sound easy but experience has shown that a considerable amount of checking and data cleaning can be required to achieve this in practice.
- In a distributed and networked world where there is increasing mobility of data and data holders there is a need for standards so that data can be extracted, migrated, integrated or aggregated successfully without requiring significant human intervention.
- As discovery is an important issue to make these resources findable and useful this also requires the use of appropriate classification systems alongside. Relevant datasets may
exist in a variety of repositories and may have arisen in other disciplines. A poignant example is the work of the Sea Mammal Research Unit (SMRU) where in the course of biological research records are created that are of specific interest for climatological research and thus cross-disciplinary cases are not uncommon.

In Figure 1, above, the conceptual framework for C4D is presented that shows where C4D fits into the current research landscape and what it attempts to provide in practical terms. Given the motivations stated earlier C4D attempts to find a solution to this problem and consequently focuses on the following issues:

- Examining the support and extension requirements of CERIF to support datasets
- Proposing a flexible standard for the metadata for datasets based on existing ones to cover a variety of disciplines
- Choosing an appropriate classification system to facilitate storage and retrieval of datasets or the metadata about datasets
- Developing tools for the manual generation of metadata into a CERIF compliant metadata data store and their exchange with other CERIF compliant data stores
- Developing a Discovery and Navigation tool by which the metadata of datasets can be discovered and retrieved as well as explored in relation to other key entities associated with these datasets such as owners, publications, organisations and projects.
In the following we look at the available metadata standards and available classification schemes in section 2 before considering the proposed solution concept in section 3 and draw conclusions in section 4.

2 Metadata and Classification Issues

With rapidly growing numbers of datasets there is a need for efficient ways for dealing with the storage and retrieval of the actual datasets on the one hand and their discoverability on the other.

2.1 Metadata Standards

According to the National Information Standards Organisation (NISO) “Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve, use or manage an information resource” (NISO 2004). A typical example of metadata is the catalogue found in any library, which records key information about the collection such as author, title, subject, as well as the location in the library. Different types of metadata can be distinguished:

- **Descriptive metadata** describes the resource for identification and retrieval purposes, e.g. title, abstract, description, topic, keywords, geospatial location, and taxa; these may be free text fields or fields linked to a classification scheme
- **Structural metadata** to describe the structure of the resource and potentially relationships between elements of the resource, e.g. record structure, fields, units, instrumentation used
- **Administrative metadata** to help manage the resource such as version, when the resource was added to the archive, licensing, encoding (eg XML, CSV, flat file etc), language, data management plan, conformity, standard name, and who can access the resource.

Metadata elements grouped into sets are called metadata schemes; these elements are the descriptor types that are available to be applied to the data. For every element in the scheme the name and the semantics are specified; some schemes also specify in which syntax the elements must be encoded while some schemes do not. Many different metadata schemes have been developed with some of them designed for use across disciplines, while others are designed for specific subject areas, dataset types or archival purposes; they provide for a variety of the needs of a variety of disciplines with some more generic such as Dublin Core (Dublin Core Metadata Initiative 1998), which originated from the domain of librarianship, but has been widely used on its own or with some features absorbed into other metadata standards. Other standards provide more focused support for specific disciplines such as Darwin Core (Darwin Core Task Group 2009) for biology or the Infrastructure for Spatial Information in the European Community (INSPIRE) (European Commission 2008), the Geospatial Metadata Interoperability Initiative (GEMINI) (Association for Geographic Information 2010) and the Marine Environmental Data Information Network (MEDIN) Discovery Metadata Standard (Seeley, Rapaport, Merritt et al 2011) for geospatial applications. The diverging needs of different disciplines, present a dilemma in order to find a single format that fits all these needs. There are a variety of basic descriptions about the dataset that will be the same such as name, description, authorship, location, language, topic and keywords, while other aspects are discipline-specific such as geospatial bounding boxes for geospatial sciences, taxa for biology, chemical structures, etc., that will be important not just for correct classification, but also for eventual discovery.
It is also important to note that there are metadata standards for different levels of abstraction where some of them are designed more for the generation and management of dataset repositories. As in some cases, repositories hold a variety of datasets from different disciplines such as National Archives and Libraries there are metadata standards that are more focused archival and librarianship such as the Encoded Archival Description (EAD), Machine-Readable Cataloging (MARC), Metadata Encoding and Transmission Standard (METS) and Metadata Object Description Schema (MODS) as well as Dublin Core. On the other end of the scale there are also metadata standards for particular types of standardised datasets which focus more on the detailed content of these datasets such as IMMA (Woodruff 2003) for ship records.

Different metadata standards have evolved to serve the needs of their respective communities and a way has to be found to either develop a more generic scheme that can work in a variety of disciplines or to work with these existing standards to develop a higher level metadata scheme that can subsume them. Given the sheer complexity of the available standards it will however be more pragmatic to start with a widely used standard such as MEDIN and assess the need to extend it to incorporate key criteria of other disciplines. Thus, geospatial aspects, in some disciplines, as well as taxa or classifications of compounds/elements/molecules or even extra-geospatial locations for others, will be needed. The discipline specific aspects are largely to do with what the dataset or resource is about and its context. At the same time there are also a considerable number of common features to do with what the resource is, what it is called, where it is located, how it is encoded, who owns it, what conditions are attached to it which are fairly common across the disciplines.

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Description</th>
<th>Format</th>
<th>Metadata Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource title</td>
<td>A brief and precise description of the dataset</td>
<td>Free text</td>
<td>Descriptive</td>
</tr>
<tr>
<td>Resource abstract</td>
<td>A clear and brief statement of the content of the dataset</td>
<td>Free text</td>
<td>Descriptive</td>
</tr>
<tr>
<td>Topic category</td>
<td>Indicates the main theme of the dataset</td>
<td>Controlled vocabulary/classification scheme</td>
<td>Descriptive</td>
</tr>
<tr>
<td>Metadata standard name</td>
<td>Identifies the metadata standard used to create the metadata</td>
<td>Free text</td>
<td>Administrative</td>
</tr>
<tr>
<td>Metadata standard version</td>
<td>Identifies the version of the metadata standard used to create the metadata</td>
<td>Numeric</td>
<td>Administrative</td>
</tr>
</tbody>
</table>

Table 1: Example metadata elements

In order to take this further and create a generic framework there needs to be agreement on which core information is required for data users to discover, use and understand the data. Following on from this, it is essential that a formal definition for each term is agreed upon. It can be seen in the table above, Table 1, examples of some of the metadata elements that are common among the standards chosen. These elements provide coverage of the necessary descriptive and administrative metadata elements required to aid discoverability of datasets.

In Table 2 the Metadata Elements of MEDIN are shown. While this standard is focused on geospatial data in the marine domain, as can be seen from elements 12 to 18, it could easily be modified to allow additional elements such as for place names and regions that might be
imported in other disciplines such as life sciences, humanities or geography or other spatial extraterrestrial references in astronomy and some special elements such as taxa which are particularly important for other specific disciplines such as in biology. The great majority of elements are applicable however independent of discipline. To facilitate discoverability across disciplines the use of a shared classification scheme would however benefit besides making use of user selectable keywords and descriptions. By focusing on a uniform standard adopted across disciplines the interdisciplinary work with datasets would be greatly enhanced as datasets can then be more readily discoverable outside their original discipline.

<table>
<thead>
<tr>
<th>MEDIN</th>
<th>Element</th>
<th>Oblig</th>
<th>MEDIN</th>
<th>Element</th>
<th>Oblig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource title</td>
<td>1</td>
<td>M</td>
<td>Temporal reference</td>
<td>16</td>
<td>M</td>
</tr>
<tr>
<td>Alternative resource title</td>
<td>2</td>
<td>O</td>
<td>Lineage</td>
<td>17</td>
<td>C</td>
</tr>
<tr>
<td>Resource abstract</td>
<td>3</td>
<td>M</td>
<td>Spatial resolution</td>
<td>18</td>
<td>C</td>
</tr>
<tr>
<td>Resource type</td>
<td>4</td>
<td>M</td>
<td>Additional information source</td>
<td>19</td>
<td>O</td>
</tr>
<tr>
<td>Resource locator</td>
<td>5</td>
<td>C</td>
<td>Limitations on public access</td>
<td>20</td>
<td>M</td>
</tr>
<tr>
<td>Unique resource identifier</td>
<td>6</td>
<td>C</td>
<td>Conditions applying for access and use</td>
<td>21</td>
<td>M</td>
</tr>
<tr>
<td>Coupled resource'</td>
<td>7</td>
<td>C</td>
<td>Responsible party</td>
<td>22</td>
<td>M</td>
</tr>
<tr>
<td>Resource language</td>
<td>8</td>
<td>C</td>
<td>Data format</td>
<td>23</td>
<td>O</td>
</tr>
<tr>
<td>Topic category</td>
<td>9</td>
<td>C</td>
<td>Frequency of update</td>
<td>24</td>
<td>C</td>
</tr>
<tr>
<td>Spatial data service type'</td>
<td>10</td>
<td>C</td>
<td>Conformity</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td>11</td>
<td>M</td>
<td>Metadata date</td>
<td>26</td>
<td>M</td>
</tr>
<tr>
<td>Geographic bounding box</td>
<td>12</td>
<td>C</td>
<td>Metadata standard name</td>
<td>27</td>
<td>M</td>
</tr>
<tr>
<td>Extent</td>
<td>13</td>
<td>O</td>
<td>Metadata standard version</td>
<td>28</td>
<td>M</td>
</tr>
<tr>
<td>Vertical extent info</td>
<td>14</td>
<td>O</td>
<td>Metadata language</td>
<td>29</td>
<td>M</td>
</tr>
<tr>
<td>Spatial reference system</td>
<td>15</td>
<td>M</td>
<td>Parent ID</td>
<td>30</td>
<td>O</td>
</tr>
</tbody>
</table>

M = mandatory     C = conditional     O = optional

Table 2- MEDIN metadata elements

### 2.2 Classification Schemes

Classification is important from the point of view of organising collections in a manageable way and to enable users to search collections in a meaningful way. Classification schemes have been developed and used in a variety of contexts and vary in scope and methodology, but can be broadly divided into universal, national general, subject specific and home-grown schemes. There are a number of classification schemes in use with a marked focus on library and information science, which are designed to organise and classify the world of knowledge and its contents. This has led to the emergence of universal classification schemes, such as Dewey Decimal Classification (DDC), Universal Decimal Classification (UDC) and the Library of Congress Classification (LCC); these are based on disciplines which are recognised fundamental fields of
study, such as Philosophy, Social sciences, Science, Technology, the Arts. Disciplines have their sub-disciplines, e.g. sciences include Physics and Chemistry, and social sciences include Sociology and Economics. In all of them the arrangement of concepts is hierarchical.

The DDC, UDC, and the LCC schemes mentioned above, while universal, do however have a library bias, which is also to do with correctly shelving and finding items in a library. However what is useful to learn from them is the benefit of a taxonomic/hierarchical system that allows expanding or restricting retrieval and browsing to suit the users’ needs. There are in addition other schemes that have emerged such as the United Nations Educational, Scientific, and Cultural Organisation’s Science and Technology Policies Information Exchange System (UNESCO-SPINES) thesaurus (Unesco Secretariat & de Padirac 1976) or Frascati (Organisation for Economic Co-operation and Development 2002), which have the benefit of being more information and information retrieval oriented as opposed to a library focus, but they do not provide the right level of granularity to classify large collections of datasets in a way where interested parties can drill down to classify datasets for later discovery and reliably find relevant datasets for a given task.

These existing classification systems despite appeal in terms of universality are essentially not well suited to a scientific focus and with enough variety to cover a vast array of disciplines and sub-disciplines and fields of research that have grown around them. However, the research councils have, for other reasons, already been developing classification schemes of their own for managing research activity for some time and which are focused on disciplines and sub-disciplines. This has lead to two systems currently in use:

- the Research Councils UK (RCUK) Subject Classification Scheme (and the Research Output System (ROS))
- the Joint Academic Classification of Subjects (JACS) subject coding system used by the Higher Education Statistics Agency (HESA)

Both schemes are hierarchical with RCUK being more sophisticated in terms of the number of classifications. The RCUK classification scheme has 78 first level categories and then subcategories to level 3, where the resolution at level 3 can vary from a single subcategory to 30 or more. Given its widespread use in managing funding and classifying and the fact that research councils are increasingly starting to develop datacentres for the curation of datasets it seems reasonable to adopt the RCUK system for the purpose of C4D.

3 Proposed C4D Discovery Architecture

Once datasets are fully supported in CERIF, and CERIF data stores, whether distributed or centralised, these data stores will be a formidable information resource that will serve a variety of users and uses. To be able to fully unlock this potential a number of issues have to be addressed:

- generation of correct metadata and integration with existing records where necessary (potentially also completing missing information in related records)
- facilitating the discovery of relevant records through appropriate retrieval mechanisms (even across distributed repositories)
- enabling the browsing and navigation of relevant datasets and their associated people, organisations, projects and publications
The generation of correct metadata and the correct insertion of records into a CERIF repository will require linking it correctly to already existing information such as that about people, organisations and projects or alternatively initiate the additional records that need to be created (e.g. a dataset needs a person who generated and owns it and should be linked to a project through which is was created and an organisation who hosted the project etc.). This will draw on some earlier work of the authors such as (Bokma 2004) and (Tektonidis & Bokma 2006). Alternatively, it might also require links to be made to data in other repositories in a distributed environment where data is to be subsequently searched and retrieved. Finally, the navigation of datasets and related entities in a relational/graphical fashion will require connections to be made that might not be explicitly present in the existing record. For this reason it is necessary to have a mechanism that can look at relationships and expectations with respect to correct records and be able to reason about it, be it for the purpose of generation of correct records, data cleaning or browsing. C4D adopts a semantic web oriented approach to provide a solution for these issues and aims to develop an ontology and an ontology-driven interface to CERIF data stores which will provide the key features of import, discovery, exploration and linking.

By modelling the key concepts of people, projects, publications, organisations and results an interface can be built that has the desired functionality of enabling the non-invasive retrieval of records also across data stores and the representation and navigation of links between entities so as to for example visualise related records surrounding a particular dataset, such as other datasets, researchers, publications, projects, etc. This will enable the vision presented in Figure 1 above to provide for powerful retrieval and navigation of these intrinsically related entities. For the thematic retrieval of datasets and as a starting point for showing related entities the use of a classification system will also be important and hence the integration of the RCUK classification system into an eventual ontology as discussed in section 2.2. When dealing with distributed repositories and where there may be inconsistencies and there is the possibility to enforce data-cleaning it is important to have a flexible mechanism that can resolve these in a non-invasive way at the point of combining and aggregating retrieval results.

In order to make use of existing work to develop a comprehensive solution more quickly we propose to base the application on R2O and ODEMapster – see (Bui-Aranda, Corcho et al 2009). R2O & ODEMapster is an integrated framework for the formal specification, evaluation, verification and exploitation of the semantic mappings between ontologies and relational databases. This integrated framework consists of:

- R2O, a declarative, XML-based language that allows the description of arbitrarily complex mapping expressions between ontology elements (concepts, attributes and relations) and relational elements (relations and attributes)
- ODEMapster processor, which generates Semantic Web instances from relational instances based on the mapping description expressed in an R2O document.

Figure 2 below shows how such a system architecture would work by linking to an existing metadata repository (CERIF) which holds the records on researchers, projects, publications and datasets; this would be accessed using ODEMapster. The application then uses the C4D ontology and maps it to the database using R2O mappings and then allows the database to be searched and data to be retrieved by interrogating the domain ontology. Depending on user preference this can then be accessed through a traditional tabular search facility or through an interactive graphical representation that allows the collection of information to be navigated and browsed. Also, the benefit of this approach is that it could be easily extended to deal with a variety of repositories.
and allow the data to be combined for more unified and transparent navigation. However this would require additional effort to deal with merging of data from different sources.

The use of a common classification scheme in combination with these key concepts will allow any of these categories and existing records to be managed for the purpose of discovery and integration. Thus the ontology will also need to contain the classification scheme as a topic hierarchy to be used to classify individual researchers, projects, datasets and publications so that the activities in particular subject areas become extractable and so that the proposed vision of being able to explore activity surrounding particular datasets, as depicted in the graphical form reproduced above, is achieved. The fact that the RCUK classification scheme is also hierarchical is useful as it allows a richer set of functionality to be generated rather than out of a flat dictionary.

### Conclusion

The support of a variety of disciplines requires a flexible approach that can capture key aspects of different disciplines. Using a uniform classification system across different datasets allows records to be correlated and collated. There is the need to be mindful of different levels of metadata, to be able to distinguish the levels to ensure a clear separation between levels and to enable eventual integration between repositories holding datasets and repositories holding research information for administrative purposes. The semantic component of C4D will add a discovery layer to the dataset enhanced CERIF repository through the development of a C4D ontology, which will be mapped to the enhanced CERIF repository. The proposed semantic navigator will enable the user to discover datasets based on the respective metadata that is stored in the enhanced CERIF repository, through a search facility as well as a graphical browsing facility.
References


Contact Information

Dr. Albert Bokma

Faculty of Applied Sciences – DCET
University of Sunderland
St. Peter’s Way
Sunderland
SR6 0DD
United Kingdom
albert.bokma@sunderland.ac.uk