CERIF for Datasets (C4D) – An Overview

Kevin Ginty[a], Simon Kerridge[b], Paul Fairley[a,] Ryan Henderson[a], Paul Cranner[a], Albert Bokma[a], Sheila Garfield[a]

[a]University of Sunderland, Centre for Internet Technologies, Digital Media Building, St Peter's Campus, St. Peter's Way, North Sands, UK

[b]University of Kent, Research Services, Canterbury, UK

Summary

The JISC funded IRIOS (Integrated Research Input and Output System) project developed a demonstrator integrating research information input (projects) and output (publications) datasets allowing data to be linked and exported in CERIF format. C4D(CERIF for Datasets), extends this platform to manage research datasets as outputs from research projects. Although research datasets vary and are extremely independent, they have some commonality that can be recorded in CERIF. C4D will 'CERIFy' existing research dataset metadata conventions, and provide access to research data in an environment which holds information on research projects and other more conventional research outputs. C4D will also explore the commonality of research dataset metadata, how much of this can be represented in CERIF and explore the possibility of implementing this in a cloud environment. C4D will use live datasets from three UK universities to test the process of 'CERIFying' dataset metadata.

Keywords: CERIF, Metadata, XML, C4D, Datasets, CRIS, Open Access, ERA, RMAS

1 Introduction

The aim of C4D is to use CERIF (Common European Research Information Format) ontologies and syntax to represent the metadata of sample datasets, and integrate this metadata with that held on research projects and other, more traditional, research outputs available at The Universities of St Andrews, Glasgow and Sunderland. In order to demonstrate the facility of the approach a subject specific demonstrator will be built and this approach verified at the three partner Higher Education Institutes (HEIs), each within their own research administration infrastructure.

The JISC funded project, IRIOS[1](Integrated Research Input and Output System) developed a platform for managing research information exchange using CERIF. Providing the C4D service as part of the IRIOS shared service platform should make it easier to promote uptake of CERIF based metadata sharing among the UK institutions using IRIOS already.

C4D will be based on IRIOS which utilises the Universities for the North East Information System (UNIS) platform which was designed to meet the core requirements of rapid customisability and extensibility to satisfy the requirements of user groups within the five campus based Universities in the North-East of England.

A survey of current datasets and metadata was undertaken as a selection process during the early stage of C4D, it was determined that the most likely subject area would be one that all three partner HEI's were active in and have access to; namely marine sciences.

This paper focuses on the C4D[2] project, with reference to the current state of the art within the areas of Research Management, CRIS (Current Research Information Systems) and the exposure and development of the CERIF standard.

- Section 2 focuses on the context of the C4D project
- Section 3 describes the planned architecture for C4D
- Section 4 addresses the various research datasets being used for the C4D project
- The paper reflects on the C4D project with additional focus on possible development and future ventures

2 Context

Most HEIs in the UK have in place numerous systems for handling research information (proposal, project and publication information), with the development of the IRIOS platform a mechanism was implemented that allowed this information to be linked and transformed into CERIF format. This provides the means to transfer and share this information between third party platforms to aid in progression and synchronization of projects and information. At this stage each of the partner institutions acknowledged that the next logical step would be to develop the infrastructure to integrate research data management due to this being largely absent at each of the partners on the IRIOS project.

[1] http://irios.sunderland.ac.uk
[2] http://c4d.wordpress.com

The Engineering and Physical Sciences Research Council (EPSRC) currently identifies research data metadata as a key part of the outputs from its funded activities. EPSRC has clear expectations of organisations in receipt of research funding relating to metadata, research organisations are expected to publish metadata within 12 months of data being generated to improve visibility of their results[3]and made available for 10 years.

Adoption of CERIF has been recommended by JISC[4] as the primary way of integrating research information and management systems across the sector, the C4D project will extend the use of CERIF into the research data management area. The overall aim of C4D is to develop a framework to support the incorporation of metadata into CERIF such that research organisations and researchers can better discover and make use of existing and future datasets, wherever they may be held.

Additionally C4D will add the capacity to store research data metadata onto the existing IRIOS platform, as well as other CERIF compliant systems, and provide a graphical interface allowing users to search the repository. The result of this will be a significant extension to the research information infrastructure, going beyond what is currently available and resulting in an integrated metadata repository which can be sited within the currently developing JISC cloud, and made available to other research organisations enhancing the discovery of datasets by researchers. By developing C4D to integrate CERIF with other research metadata standards, techniques derived from the project will be useful across a wide range of contexts where research data is reported as a project output.

2.1 IRIOS

The IRIOS project developed a proof of concept demonstrator based on the UNIS platform for a CERIF compliant “grants on the web” system for UK Research Council (RC) funded projects. IRIOS helped to accelerate the shared usage of CERIF by RCs and Research Organisation (RO) by providing access to information on RC funded projects in CERIF format and linking grants.

The aim of C4D is to extend this platform with the overall aim being to develop a single platform allowing for RCs and HEIs to effectively manage and share their combined research activities. By providing C4D as part of the IRIOS service it is likely to make promotion uptake of CERIF based metadata sharing among partners easier as the service is already in use.

3 Architecture

[3] http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx
[4] http://www.jisc.ac.uk/publications/reports/2010/businesscasefinalreport.aspx

C4D will develop a proof of concept demonstrator, using the already existing IRIOS platform. By extending this pre-existing platform to further focus on research metadata, C4D will take advantage of a robust and secure platform already in use and familiar to users.

New elements that are addressed within C4D are:

- Mechanisms for standardising research data metadata into a common format (CERIF)
- Processing of metadata from its source into destination directory services

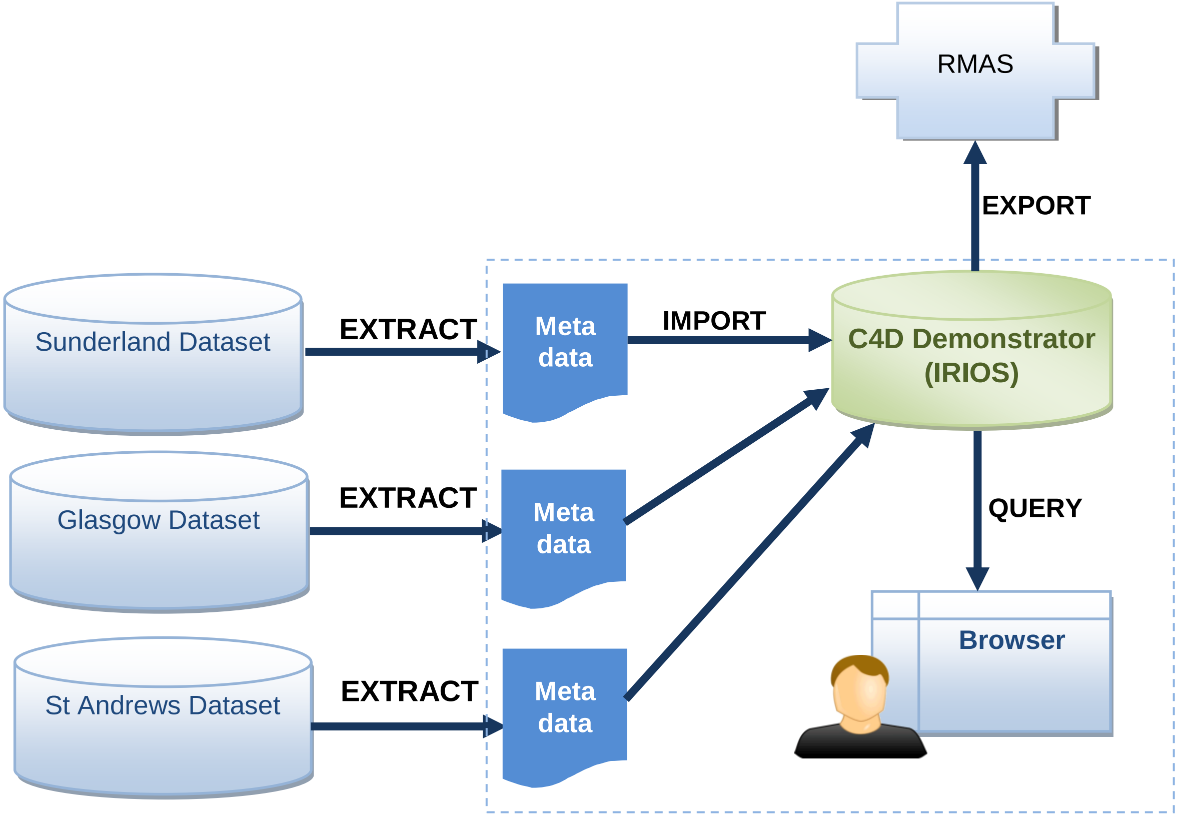

Figure 1: C4D system architecture

C4D aims to address these new elements by creating a standard method of incorporating metadata into CERIF so that it may be transferred and shared, potentially increasing collaboration of research activity and discovery of research data.

The C4D architecture, shown in figure 1, will enable users of the system to import, export and browse metadata as core functionality, discoverability will be greatly increased due to the metadata repository being created as well as the system being CERIF compliant. As EPSRC has identified research data metadata as a key output from its funded activities, the publication of datasets so that they can be found is a priority among research organisations. C4D will facilitate this process, and increase community engagement with metadata production and publication.

As can be seen in figure 1 (Representative of the Sunderland system) the C4D platform will be piloted with connection to an external CRIS system, this system will be individual to each pilot institution. This will enable the connection of C4D metadata to project information, further increasing potential interest from external research organisations through wider compatibility.

The system will be piloted by the three partner institutions (Universities of Sunderland, Glasgow and St.Andrews), each with their own CRIS. St.Andrews will integrate with their Pure[5] CERIF-CRIS with research data metadata from selected datasets, the experience of this integration will be relevant to the other UK institutions who have so far purchased Pure, as well as several others using third party acquired CERIF-CRIS's such as Converis[6] and Symplectic[7]. Sunderland will pilot the C4D service with the RMAS platform[8]. Glasgow will pilot C4D with their own in house system incorporating ePrints[9]. By employing three different pilot approaches, C4D will be developed in a system agnostic way further extending the possible future developments and third party organisations that may be interested in development and use of such a system in the future.

4 Research Datasets

The University of Sunderland currently has access to several suitable research datasets for C4D. Specifically these include the Climatological Database for the World's Oceans (CLIWOC)[10] and the JISC-funded UK Colonial Registers and Royal Navy Logbooks digitisations project (CORRAL)[11] datasets, providing historical and marine climatology datasets.

At the University of St Andrews, the Sea Mammal Research Unit[12] is an NERC collaborative centre and provides the UK's main capability in the field of marine mammal biology collecting data on seal and cetacean populations and related oceanographic data.

[5] http://www.atira.dk/en/pure/
[6] http://www.avedas.com/
[7] http://www.symplectic.co.uk/index.html
[8] http://www.exeter.ac.uk/research/rmas/
[9] http://www.eprints.org/openaccess/
[10] http://www.ucm.es/info/cliwoc/intro.htm
[11] http://www.corral.org.uk/
[12] http://www.smru.st-and.ac.uk/

At the University of Glasgow, the School of Geographical and Earth Sciences[13] currently has access to marine Climatological datasets, primarily for the North Atlantic, extending back up to 650 years.

At present CERIF does not have the ability to support metadata fully, this is the issue C4D aims to address to allow for CERIF to manage and aid in the sharing of research data.

5 Metadata Standards

To effectively incorporate metadata in CERIF it is vital to investigate multiple metadata standards. Due to the differences in data and varying needs in different research areas, no single metadata standard exists. The C4D project will focus on marine datasets for the initial demonstrator, with the more long term plan being to create a mapping into CERIF that is applicable across the widest possible range of subject areas. Although the project will focus mostly on standards that are subject area specific to ensure that the sample datasets are thoroughly covered by the mapping, it will also be necessary to look at some of the more generic metadata standards.

Metadata descriptions from different standards are not semantically linked but overlap and relate to each other in various ways. As the number, size and complexity of metadata standards grow, the task of facilitating metadata becomes increasingly difficult. In order to maximize the usefulness of metadata to the wider audience use of a unique metadata standard should be worked towards (Nogueras-Iso, 2004).

To create a system that will be highly generic with the ability to cover management of metadata in multiple subject areas, C4D looks at selected metadata standards based on marine sciences as well as generic model with the intention of creating a universal mapping into extensions of CERIF-XML. With the constant development of metadata standards for specific subject areas (some of which will become widely used) it is vital that mappings into CERIF cover as many aspects of metadata as possible (Hirwade, 2011). C4D will take multiple standards and, aiming at maximum interoperability, develop a mapping with common terms between the selected metadata standards for the management of metadata.

Some of the standards that have been included in a CERIF mapping for C4D are:

- Marine Environmental Data Information Network (MEDIN) Discovery Metadata Standard - The MEDIN metadata schema is based on the ISO 19115 standard, and includes all core

[13] http://www.ges.gla.ac.uk:443/

INSPIRE metadata elements. It also complies with the UK GEMINI 2.1 metadata standard. The xml produced conforms to the ISO 19139 standard for xml implementation.

- UK GEMINI2 – UK GEMINI2 specifies a core set of metadata elements for use in a geospatial discovery metadata service. It is the definitive metadata standard for describing geographic information. This revised version of the GEMINI standard is compatible with the requirements of the INSPIRE metadata Implementing Rules (IR), conforming to the ISO 19155.
- Dublin Core (DC) – It is a generic metadata standard which is used extensively in digital-libraries, it is currently being adopted to be compatible with geographic information in order to enhance the level of compatibility across digital cataloguing systems. (Nogueras-Iso, 2004)
- Darwin Core - primarily based on taxa, an ordered system that indicates relationships among organisms, their occurrence in nature as documented by observations, specimens, and samples, and related information.

Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve, use or manage an information resource. Metadata is often called data about data (Understanding Metadata, 2004). Consistent use of standard terminology for metadata will help the understanding, discovery, integration and use of research data sets.

Work carried out within the scope of C4D includes a mapping between (some) selected standards and potential CERIF entities to allow for the comparison of terminology and its usage. C4D will produce a mapping that identifies metadata content needs to include elements from multiple standards and identify any areas where cross over will be possible. The standards chosen cover a wide range of disciplines and will thus support the development of a generic solution.

By using the afore mentioned projects and the discussed metadata standards we will evaluate the feasibility of using CERIF as a standard protocol to represent ‘higher-levels’ of abstraction of research data metadata, and demonstrate how this data metadata can be used alongside other CERIF entities for projects, people, institutions and outputs.

6 Anticipated Outcomes

C4D will produce a framework for incorporating metadata into CERIF by building a subject specific demonstrator (system) allowing for the import and management of metadata related to the selected research datasets. The C4D demonstrator will be built on a populated IRIOS platform to allow for testing of the specified functions.

The demonstrator and framework will be produced to handle marine datasets as they are common to all three institutional partners, however various standards will be incorporated into a Metadata-CERIF mapping to ensure a certain level of generic coverage. C4D will facilitate the process of publishing metadata to enable research datasets to be more discoverable. Due to IRIOS already being populated with research information relating to projects it will be possible to incorporate metadata with the information already held.

Definition of a CERIF common metadata ontology will take place to relate research datasets to other research information and provide a standard means to support CERIF among institutions. Delivery of an ontology and ontology driven interface will provide the ability to import, export, discover, explore and link metadata.

Finally the result of the C4D project will be the production of an open access metadata repository for use by researchers and those maintaining research datasets. The central repository will house relevant metadata-information to enable researchers to identify required archive data and will be built on the existing IRIOS infrastructure. The Anticipated outcome of this is that part of the service will reside in the JISC cloud currently under development, and in turn allows for the entire C4D platform to eventually reside within the cloud.

7 Future Development

Future developments around C4D can be widely discussed. There exists considerable scope to extend the service beyond that of the project. Primarily scope for synergy exists between the C4D project and the currently developing IRIOS-2[14] platform; both will share an underlying database platform and will therefore overlap at certain points during the lifetime of both projects.

[14] http://irios2.wordpress.com/

Although still in the development phase the JISC cloud provides a multitude of options for further developing C4D, deployment of the final system in the cloud computing environment currently in development by JISC will be planned, and costs attached to developing and harvesting metadata estimated for a range of target projects . The continuation of liaising between institutions and EuroCRIS will go towards the evolution of CERIF, extended interest in the C4D platform and any future developments that could be made to the service.

At present C4D is a UK based project, and therefore restricted in the level of exposure it will receive. At present the University of Sunderland are part of a consortium developing a project proposal that will make use of the lessons learned within C4D to further develop the idea of producing a collaborative data infrastructure which combines the ideas originally discussed in C4D with those of partners within the European Union.

References

J.Nogueras-Iso, F.J. Zarazaga-Soria, J. Lacasta, R.Bejar, P.R. Muro-Medrano, (2004): Metadata standard interoperability: application in the geographic information domain. *Computers, environment and urban systems*, Vol.28, p.611-634

Hirwade, M.A (2011): A study of metadata standards. *Library High Tech News,* No.7, p.18-25

Seeley, B, Rapaport, J, Merritt, O, Charlesworth, M (2009): Guidance notes for the production of discovery metadata for the Marine Environmental Data and Information Network. Retrieved: 9th March 2012 from: http://www.oceannet.org/marine_data_standards/medin_approved_standards/guidance document.html

Understanding Metadata (2004). USA: NISO Press. Retrieved: 10th March 2012 from: http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Contact Information

Kevin Ginty, University of Sunderland, Centre for Internet Technologies, Digital Media Building, St Peter’s Campus, St. Peter’s Way, North Sands, UK e: gintyk@gmail.com
t:0191 515 3236