CERIF for Datasets (C4D)

Linking and contextualising publications and datasets, and much more...

Scott Brander[1], Anna Clements[1], Valerie McCutcheon[2], Paul Cranner[3], Ryan Henderson[3], Kevin Ginty[3.]

[1] University of St Andrews, St Andrews, Fife, United Kingdom
{scott.brander, akc}@st-andrews.ac.uk
[2] University of Glasgow, Glasgow, United Kingdom
valerie.mccutcheon@glasgow.ac.uk
[3]University of Sunderland, Sunderland, United Kingdom
{paul.cranner, ryan.henderson, kevin.ginty}@sunderland.ac.uk

Abstract. The overall aim of CERIF for Datasets (C4D) is to develop a framework for incorporating metadata into CERIF (the Common European Research Information Format) such that research organisations and researchers can better discover and make use of existing and future research datasets, wherever they may be held. CERIF provides a standardised way of managing and exchanging research information and has been widely used for recording and exchanging information about research projects and publications. C4D looks at the suitability of CERIF for recording datasets, suggests ways that the model could be improved and implements pilot functionality based on the findings of C4D at the three partner Universities in the UK.

Keywords. Datasets, CERIF, Research Information Systems, Current Research Information Systems (CRIS), Institutional Repositories (IR), CERIF for Datasets (C4D).

1 Project context

1.1 Background and aims

The overall aim of CERIF for Datasets (C4D) is to develop a framework for incorporating metadata into CERIF (Common European Research Information Format) such that research organisations and researchers can better discover and make use of existing and future research datasets, wherever they may be held.

This project is funded by JISC, '*... the UK's expert on digital technologies for education and research'* [1] with partners: the Universities of Sunderland, Glasgow and St Andrews, Research Councils (NERC and EPSRC), and euroCRIS as expert advisors. Originally running for 18 months from October 2011, it has now been extended to September 2013.

CERIF provides a standardised way of managing and exchanging research information. The UK is leading the way in promoting and adopting CERIF as the research information exchange format of choice. This underpins efforts to open up access to research publications and data, as signalled in the UK Government's 'Open Data White Paper [2] published in June 2012; and JISC have been providing significant funding and strategic support for projects to investigate and implement practical CERIF solutions since 2009 [3].

1.2 Reporting and assessment landscape

In March 2012, a report by UKOLN found that almost a third of UK Higher Education had implemented a CERIF-compliant CRIS since 2009 [4].
This rapid adoption has been driven by the desire to better support research management at the institutional level and in particular to streamline reporting to funders. The UK operates a dual funding strategy for research [5]; with approximately 1/3rd of funding (quality-related research [QR] money) being distributed as a result of national quality assessment exercises taking place every 5 years or so (the next being REF2014) and a 1/3rd distributed via the 7 discipline-based research councils through competitive application. The remaining 1/3rd comes from charities, business and other sources.
The UK funders are increasingly interested in datasets as well as the more traditional outputs, such as journal articles. In 2011, the research councils, collectively known as RCUK [6], published a set of common principles on data policy[1] and in April 2013, their new policy on open access came into force. This policy requires articles to be made open access and also requires that each article provides a statement on how underlying research materials, such as data, samples or models, can be accessed.
All these developments are part of the national and – and indeed international - agenda to make research more open and accessible. The C4D project looks specifically at three different CERIF-compliant CRIS/IR infrastructures and how these can be extended to include linkages to datasets and so facilitate the existing open access and reporting requirements, anticipate new requirements from other funders, and provide rich contextual information for other researchers and research users looking for relevant research outcomes to explore, reuse and build upon.

1.3 Questions we want to answer

The main questions we wanted to answer in the project were:

1. What metadata do we want to capture?
2. What vocabularies do we need?
3. Can CERIF provide what we need?

[1] RCUK: RCUK Common Principles on Data Policy

Given that our aim is to use the outcomes of the project in our existing research information infrastructure, we had the following principles guiding us in exploring these questions:

- Use established standards where possible; i.e. do not reinvent the wheel
- Be realistic … if we ask for too much we may get nothing and we need to implement **now**
- Design for change as this is a rapidly developing field

All three institutional partners have existing data sets from marine research funded by one of the funder partners, NERC (National Environment Research Council). NERC also runs 6 national data centres which are responsible for the long-term management of research data resulting from their funded research. The marine science data centre is the British Oceanographic Data Centre [7], which expects data deposited to conform to the MEDIN[2] format [8]. See Section 3 for more details on mapping MEDIN to CERIF.

Throughout the project we have engaged with others working in this area, including the Universities of Oxford and Bristol, and reviewed the emerging consensus in the minimum metadata required for identifying and discovering research data. The University of St Andrews also consulted with the other Pure Users and Atira, the makers of Pure, to agree a common set of metadata that was sufficient to provide the level of contextual reporting required by the funders but not so burdensome that researchers would be reluctant to or be unable to provide the information. Similarly, the University of Glasgow has consulted with other ePrints Users.

In term of vocabularies, we were particularly keen to find a common way of categorising research areas or themes. The related JISC-funded Engage project [9] looking at research clusters[3] at the University of Glasgow looked at this topic [10] and recommended the use of the recently harmonised RCUK Classification scheme.[4] In addition, there are further vocabularies required such as those for defining data types and access to research datasets.

The third question, i.e. whether CERIF can provide what we are looking for, is the main focus of the rest of this paper.

2 What is CERIF and what is a CERIF-CRIS?

2.1 Description

CERIF is a conceptual model describing the research domain and is maintained formally as an entity relationship model (ERM). It is a standard, living model for the development, implementation and interoperability of current research information systems (CRIS). There have been several iterations of the CERIF model since 1991 [11], with the latest version (v1.5) released earlier in 2013.

---

2 Marine Environmental Data and Information Network (MEDIN); http://www.oceannet.org/

3 http://researchclusters.wordpress.com/

4 http://eprints.gla.ac.uk/69724

The main entities of CERIF are represented in Fig 1 below:

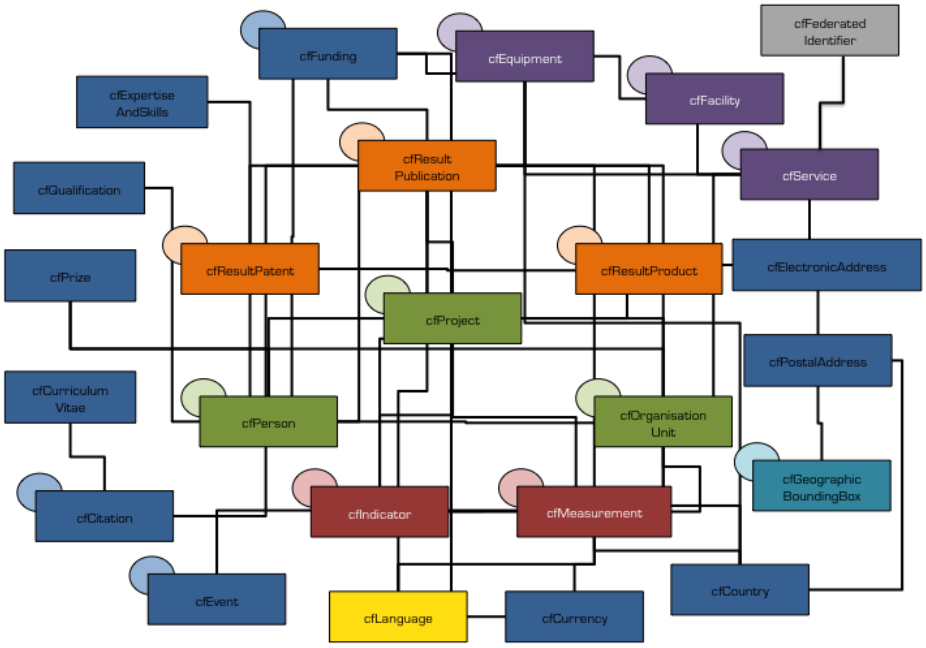

Fig. 1. CERIF research information entities

A Current Research Information System (CRIS) is a database or other information system storing data on current research by organisations and people.
In order to gain a holistic view of research information, we have presented a view of the research information system as the centre of research-related activity within an institution (or region, or country…) which integrates and interoperates with other systems.
For example, at the University of St Andrews we have the model represented in Fig 2. The dashed lines indicate the current work in C4D to integrate research datasets.

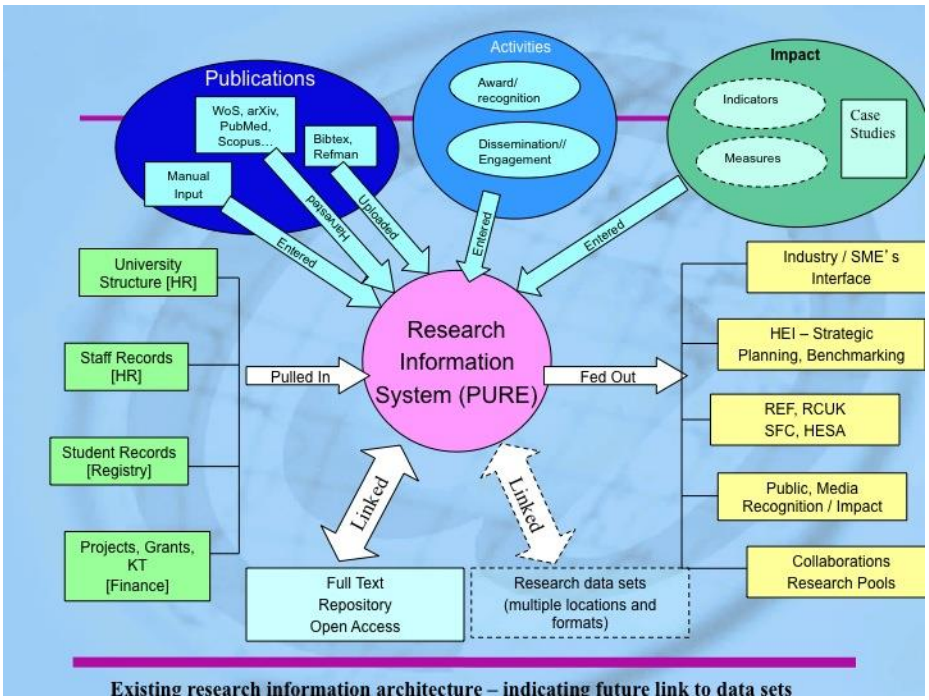

Fig. 2. CRIS model at University of St Andrews

CERIF is maintained by euroCRIS, a not-for-profit community of users, experts and developers of research information systems and dedicated to supporting all members of the research community by advancing interoperability between CRIS and related systems using CERIF. At the end of 2012, euroCRIS had over 170 members (institutional and personal) in 43 countries, including in North America, Asia and Africa, as well as Europe [12].

2.2 CERIF advantages

CERIF has many advantages as the canonical model (the research information entities, attributes, associations and semantics) for contextual metadata for datasets:

- Covers all aspects of research information: researchers, projects, organisations , funding, outputs (publications, patents, products including data sets) , equipment, services, and so on;
- An optimal (relational) architecture allowing the expression of any kind of relation between entities/attributes with every relation "time-stamped" and semantically defined;
- Very fine-grained structure, allowing output of the metadata to virtually any format;

- A separated “semantic layer” allowing the use of multiple (any) controlled vocabularies (classifications, typologies) as well as their cross-linking and mapping;
- Ability to cope with multiple languages.

2.3 Current use cases

Pure at the University of St Andrews
The University of St Andrews purchased Pure in 2009 (the first to do so in the UK as part of a joint project with the University of Aberdeen) having identified the need for a fully functional and integrated research information system to replace their existing in-house research expertise database. Pure is a user-driven enterprise-class CRIS based on CERIF and currently covers projects, outputs, staff, students, organisations (internal and external), equipment, activities and awards and the relationships between them.
Datasets are not currently captured sufficiently in Pure, although there is a growing need by funders and institutions for them to be preserved, alongside sufficient metadata to enable the data to be understood and discovered.

ePrints at the University of Glasgow

The University of Glasgow has a long history of integrated core systems. The research support system (where projects, applications, awards, internal and external organisations are stored) has been linked to the human resources (staff), finance, and student systems for many years and the repository (where research outputs are stored) was linked to the Research Support System 2010 under the JISC funded ‘Enrich’ project [13]. CERIF export facilities are included in the repository which is fully accessible via the web [14].
As part of the C4D project, the University of Glasgow have set up an ePrints data registry and are working with other ePrints users to standardise the common ePrints CERIF-compliant metadata fields to satisfy ePrints user and stakeholder requirements including funder terms and conditions such as the RCUK policy. An ePrints UK User Group has recently been formed and further consultation will take place via this. The University has recently signed up to the DataCite [15] service so that unique Digital Object Identifiers can be assigned to datasets.

UNIS at the University of Sunderland
UNIS is a collaborative project management/CRM tool which is used by the five universities in North East England, including the University of Sunderland, to manage reach-out activity. The system was designed to meet the core requirements of rapid customisability and extensibility to satisfy the requirements of user groups within those five universities. Adapting UNIS for research management purposes took advantage of a robust and secure platform already familiar to some users.

Over a period of time the UNIS platform has been adapted and extended to include research information by supporting the import and export of Research Council data in

CERIF format, and the linking of research grant information and publications together. In C4D, Sunderland extended the platform adding the capacity to store research data metadata in an environment which already holds data on research projects and research outputs. C4D also provides functionality to link the metadata to grant information and an interface to search the repository.

3 CERIF: Where datasets fit

3.1 CERIF – the current model

There are numerous ways of defining a dataset, but in its simplest form a dataset is a set of data that is collected for a specific purpose. The dataset can be collected in many ways, and may take the form of surveys, interviews, observations, census data, raw data from equipment, and so on. A dataset may also be a research input as well as an output.

For the purposes of C4D we concentrated on datasets resulting from funded research; with the emphasis on data underlying research publications - this being a pragmatic approach to allow us to fulfil existing funder policy requirements.

The key aspects of such datasets are that they should be discoverable and citable; therefore easily identifiable, stable, complete and be seen in context i.e. related to the project and funder, the researchers, the output(s) and publisher(s), even the equipment and other activities involved in the research which produced the research data.

The central CERIF entity behind the concept of a dataset is cfResultProduct and it maintains multiple relationships with other entities such as publications, persons, organisations, patents, projects, equipment and funding in line with CERIF's fully relational capability. Fig 3 below is a subset of CERIF entity relationship model showing the ResultProduct entity and its relationships.

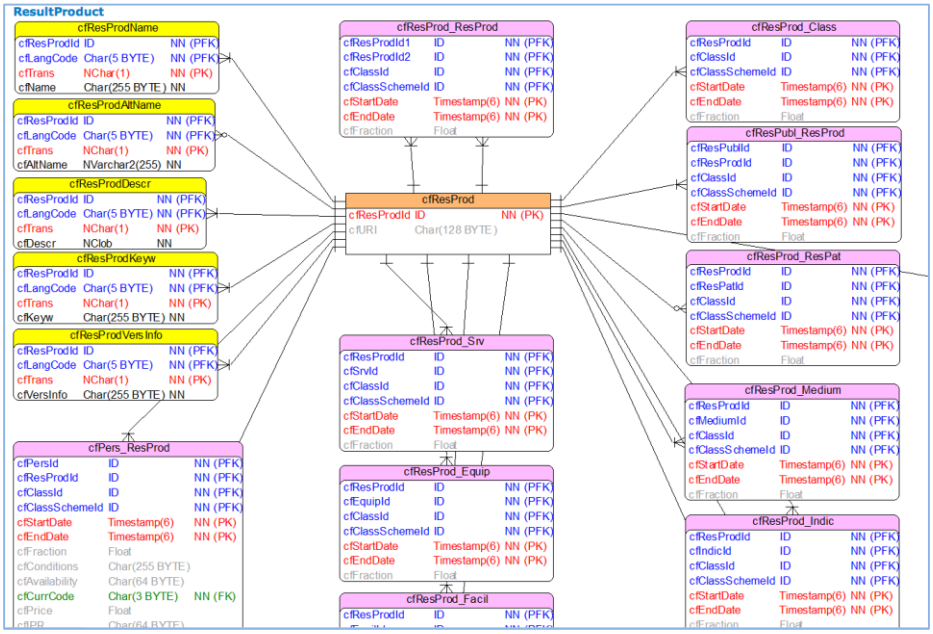

Fig. 3. Extract from CERIF 1.6 entity relationship model

3.2 Mapping MEDIN to CERIF

A detailed mapping exercise was conducted from the C4D use case – marine metadata from the Marine Environmental Data and Information Network (MEDIN), a profile of the GEMINI2 metadata standard – and the elements were mapped largely to cfResultProduct. The full mapping is available at the C4D project blog [16]. An abridged version can be viewed in Appendix A.

This exercise was largely successful with 24 of the 30 elements mapping across directly. The remaining 6 elements require extensions to CERIF and these have been recommended to euroCRIS and discussed in the CERIF Task Group which maintains the standard. Three of the extensions have already been approved and will appear in the next release of CERIF v1.6. The remaining items are still being discussed within the CERIF TG. Further detailed documentation on the C4D mappings and recommendations can be found at the project blog including an implementation example [17].

In making their recommendations, the CERIF Task Group has also looked at other dataset metadata schemas, including DCAT and eGMS, as well as receiving input from OpenAire and UK REF (Research Excellent Framework) requirements. These include the need to identify sensitive outputs, including datasets, and the requirement to link datasets to projects and related funding. More information can be found at Brigitte Jörg's CERIF Support blog [18].

4 Progress in implementation

Overarching aim to link research data sets to the other research information already in the institutional research information systems.

4.1 University of St Andrews : commercial CERIF-CRIS

The C4D application profile is being implemented in phases in the commercial Pure CRIS with first phase due for release in v4.17, Oct 2013.

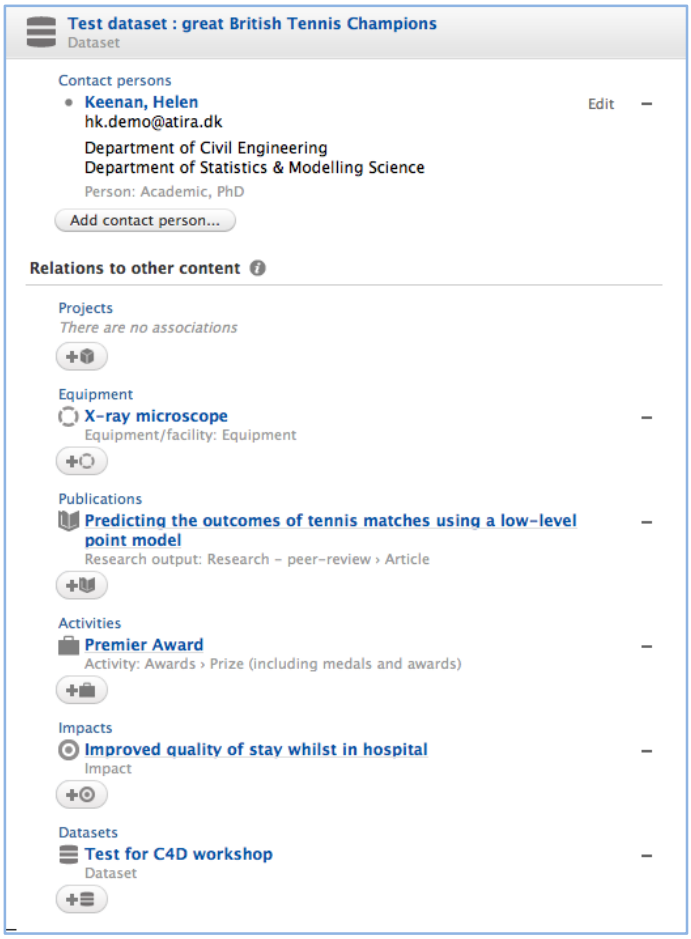

Fig. 4. Extract from Pure 4.17 beta: showing dataset links to contextual information

4.2 University of Glasgow: open source IR software

A sub-set of the key fields was implemented in January 2013. Discussions with other ePrints sites and feedback from users allow ePrints to modify the look and feel periodically [19].

Fig. 5. Example from Glasgow Research Data Repository

4.3 University of Sunderland: in-house CERIF-CRIS

The UNIS/C4D system extended the research information infrastructure, going beyond what was available and resulting in an integrated metadata repository. The current Sunderland platform is available as a beta demonstrator system.

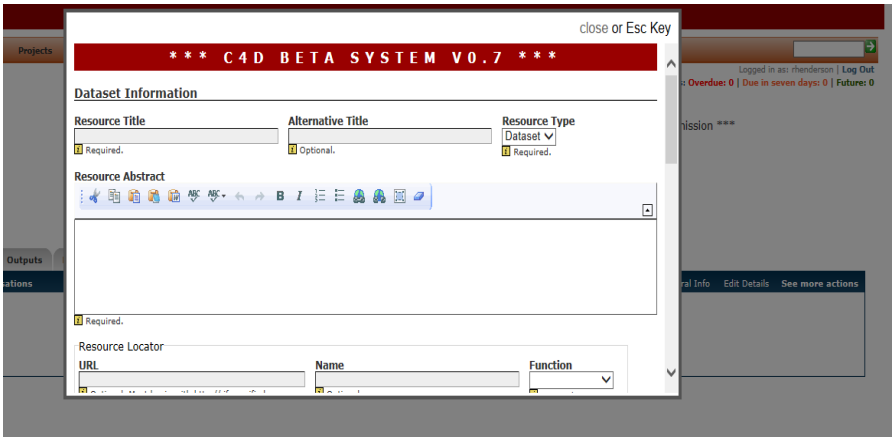

Fig. 6. Datasets implementation in UNIS

4.4 Next steps

The C4D project has demonstrated that CERIF can be used to record rich metadata about datasets and, crucially, relate these datasets to the many other pieces of the information jigsaw in the research landscape. All three partner institutions are implementing some or all parts of the C4D profile within their respective research information frameworks as described in section 4.3.

The flexibility of CERIF does mean that there are some areas which still need more consultation with euroCRIS in order to agree the best approach. In some cases this is because it is not clear what type of information needs to be recorded e.g. element 21: "*Conditions applying for access and use*" - should this be picked from a pre-defined list (i.e. a classification) or a free-text field? In another example e.g. element 18: "Lineage", more comprehensive modelling work is required, although the building blocks, such as cfMeasurement and cfGeoBBox, exist to capture this information; but not all the necessary linkages. It is also questionable as to how generic such a metadata element is.

The C4D project also includes members of the UK-based Digital Curation Centre [20] who are working on proposals for a national data register and working with euroCRIS on ensuring CERIF-compliance. The work done in C4D and other projects, such as the EU FP7 Engage project [21] which recommends a three layer model, with CERIF as the middle context-rich layer, will feed into this national data register model.

5 Conclusion

The project has delivered working pilots of data registries at the three partner Universities, with links to existing contextual research information infrastructure. However, the project has also indicated that there are areas where CERIF needs to be

extended, or potentially remodelled, and this work is ongoing within the CERIF Task Group.

5.1 The future of datasets in CERIF

C4D has looked at datasets as a research output and this is how it is currently modelled in CERIF i.e. as cfResultProduct. However datasets can also be inputs to research. Should a new entity, cfDataSet, for example be introduced? C4D has also demonstrated that the current dataset model requires additional linkages to areas of CERIF that cfResultProduct does not currently provide - again, an argument that a new 'cfDataset' entity may be required.

As with other entities in CERIF, there is a need for agreed definitions and vocabularies (CERIF classification schemes [cfClassScheme]) in order to allow true interoperability between systems. euroCRIS works closely with CASRAI [22], the standards organisation which develops and maintains a common data dictionary for full lifecycle of research activity. One of the profiles currently being looked at within the CASRAI UK chapter is on data management.

Funding Acknowledgement
This work was supported by JISC [grant number DIINNAA].

References

1. JISC, http://www.jisc.ac.uk
2. HM Government, Open Data White Paper : Unleashing the Potential https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/78946/CM8353_acc.pdf
3. JISC, Reseach Information Management : Towards a common standard for exchanging research information http://www.jisc.ac.uk/publications/briefingpapers/2010/bpexriv1.aspx
4. UK UKOLN, http://www.ukoln.ac.uk/isc/reports/cerif-landscape-study-2012/CERIF-UK-landscape-report-v1.1.pdf
5. UK Innovation Research Centre https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/181652/bis-13-545-dual-funding-structure-for-research-in-the-uk-research-council-and-funding-council-allocation-methods-and-the_pathways-to-impact-of-uk-academics.pdf
6. Research Councils UK (RCUK) http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
7. NERC, British Oceanographic Data Centre http://www.bodc.ac.uk
8. Marine Environmental Data and Information Network, MEDIN http://www.oceannet.org
9. JISC Engage Project http://researchclusters.wordpress.com/
10. Research Councils UK, RCUK http://www.rcuk.ac.uk/research/Efficiency/Pages/harmonisation.aspx

11. CERIF Support Blog, Brigitte Joerg, CERIF in Brief
http://www.cerifsupport.org/cerif-in-brief/
12. euroCRIS Annual Report 2012
http://www.eurocris.org/Uploads/Web%20pages/annual_report/ANNUAL_REPORT_2012.pdf
13. JISC Enrich Project
http://www.gla.ac.uk/enrich/
14. Glasgow University ePrints repository
http://eprints.gla.ac.uk/
15. Datacite
http://www.datacite.org/whatisdatacite
16. Cerif for Datasets C4D project blog
http://cerif4datasets.files.wordpress.com/2012/09/c4d-cerif-mapping-v0-1.xlsx
17. Cerif for Datasets C4D implementation example
http://cerif4datasets.files.wordpress.com/2013/03/c4d-cerif-metadata-implementation.pdf
18. CERIF Support Blog, Brigitte Joerg, Datasets in CERIF
http://www.cerifsupport.org/2013/04/02/data-in-cerif
19. Glasgow University research data repository
http://researchdata.gla.ac.uk/
20. Digital Curation Centre DCC
http://www.dcc.ac.uk
21. Houssos, N., Jörg, B., Matthews, B. A multi-level metadata approach for a Public Sector Information data infrastructure. In: Jeffery, Keith G; Dvořák, Jan (eds.): Proceedings of the 11th International Conference on Current Research Information Systems. Prague, Czech Republic. pp. 19-31. (2012)
http://www.eurocris.org/Uploads/Web%20pages/CRIS%202012%20-%20Prague/CRIS2012_2_full_paper.pdf
22. CASRAI, http://casrai.org

Appendix A: Table mapping MEDIN (GEMINI) metadata to CERIF v1.5

	MEDIN	DataCite v3.0 Mandatory Recommended Optional	CERIF v1.5	Notes
0	Identifier	M	cfResProdId	
1	Resource Title	M	cfResProd, cfResProdName.cfName	
2	Alternative Resource Title		Not supported – proposed to CERIF Task Group	Approved and due v1.6, summer 2013
3	Resource Abstract	R (Description)	cfResProd.cfResProdDescr.cfDescr	
4	Resource Type	R	cfResProd.cfResProd_Class	
5	Resource Locator		cfResProd_Srv.Srvid	
6	Unique Resource Identifier		cfResProd.URI	
7	Coupled Resource		cfResProd_ResProd.classId	
8	Resource Language	O	cfResProd_Class.cfLang with appropriate cfLangCodes	
9	Topic Category	R (Subject)	cfResProd_Class.cfClassId with appropriate classification scheme	
10	Spatial Data Service Type		cfResPubl_Srv.cfClassId linked to cfResProd	
11	Keywords		cfResProd.ResProdKeyw.Keyw and cfResProd.cfResProd_Class.cfClassId	
12	Geographic Bounding Box	R (GeoLocation)	cfResProd_GeoBBox.GeoBoxId	
13	Extent		cfResProd_Class.cfClassId	
14	Vertical Extent Information		Not supported in CERIF. CERIF has a GeoBBox element which can be used to record these attribute, but there is currently no cfResProd_GeoBBox linking element.	Approved and due in v1.6, summer 2013
15	Spatial Reference		cfResProd_Class.cfClassSchemeId with spatial reference system	

			classification scheme	
16	Temporal Reference	M (Publication Year) R (other dates e.g. period of collection)	cfResProd_Class.cfClassSchemeId with temporal reference classification scheme	
17	Lineage		Not currently supported – proposed	CERIF TG still discussing
18	Spatial Resolution		Not currently supported – proposed	Recommendation is cfResProd_GeoBBox
19	Additional Information		cfResProd_cfResPubl.ResPublId with classification scheme	
20	Limitations on Public Access	O (Rights)	cfResProd_Class with appropriate classification scheme	
21	Conditions applying for access and use	O (Rights)	Not currently supported – proposed free text	CERIF TG still discussing
22	Responsible party	M (Creator, Publisher) O (Contributor)	cfOrgUnit_ResProd.OrgUnitId cfPers_ResProd.PersId	
23	Data Format	O	cfResProd.cfResProd_Class with Data Format classification scheme	
24	Frequency of Update		cfResProd_Class.ClassId with Frequency of Update classification scheme	
25	Conformity		cfResProd_Measurement.MeasId	
26	Metadata Date		This is managed by the application	
27	Metadata Standard Name		No recommendation by CERIF Task Group, so was mapped to cfOrgUnit_cfresProd with linking roles	
28	Metadata Standard Version		As per 27	
29	Metadata Language		Is cfLang entity but no link to cfResProd currently	CERIF TG still discussing
30	Parent ID	R (Related Identifier)	cfResProd_ResProd with appropriate classification scheme	

Table 1. Mapping of MEDIN elements to CERIF and DataCite