Toward the integration of datasets in the CRIS environment: A preliminary analysis

Daniela Luzi[a], Rosa Di Cesare[a], Roberta Ruggieri[b]

[a] National Research Council, Institute for Research on Population and Social Policies, Rome, Italy

[b] Senate of The Republic, Rome, Italy

Summary

The paper presents the results of an explorative, qualitative analysis of research data archives listed in OpenDOAR that provide information about projects. Its aim is to analyze whether in the current practice of research data archives metadata describing projects are present, which types of metadata are used, how this information can be retrieved. The variety of archives listed in OpenDOAR allowed us to capture common characteristics in the broad context of research data than can be the basis to identify issues related to the granularity needed in the description of the association between research data and projects.

1 Introduction

The National Science Foundation (NSF) identifies three functional categories of digital data collections: a) research b) resource or community and c) reference data collections. This classification aims to distinguish between research data collected within a project with a certain dimension and budget, as well as different types of funds and funding sources (NSA, 2005). This is also one of the indicators used to evaluate whether the collection should be preserved taking into account the scientific community of reference, their need of data availability and re-use.

Implicitly these three functional categories indicate the close relationship between research data and projects, which are generally the setting prompting data collection according to the plan and methods set up in the research. Thus, metadata describing a project (grant agreement, aim, participating institutions and researchers, duration, budget, etc.) provide the context in which research data are acquired enhancing the range of information as well as improving its discovery and accessibility. From a CRIS perspective, besides these benefits for researchers, an integrated description of standardised metadata related to both projects and research data provides useful information for policy makers and funding organizations to plan future research programs, envision and/or further support the development of infrastructures for data curation and preservation, thus maximising the return of investments in research activities (Lynch, 2007). Moreover, an integrated description enables the evaluation of project outcomes including both scientific publications and research data.

Measurements that are promoting free access to research data at international level as well as its curation and preservation are often connected with projects. Similarly to the European Community measurements that promote better access to publications resulting from the research it funds (see for instance OpenAIRE project), the National Science Foundation is planning to require that project proposals have to include a data management plan (NSF, 2010).

The aim of this paper is to analyse whether in the current practices of research data archives metadata describing projects are present. To this end, an analysis on the archives registered in OpenDOAR (The Directory of Open Access Repositories) and containing both research data and project information was carried out. Whereas other studies aim to analyse how to CERIFy metadata of research data in the context of CRIS (C4D 2012), our objective is to capture the different ways used to make the connection between research data and projects evident. For these reasons we chose to carry out this analysis retrieving information in the OpenDOAR directory that lists a large variety of open access archives and thus can provide an heterogeneous set of cases in the broad context of research data acquisition. Therefore, this is an exploratory, qualitative analysis that can contribute to identifying issues related to the granularity needed in the description of the associations between datasets and projects.

The paper first outlines the methods used in the analysis and then provides the results focusing on the characteristics of the sample (providers, types of archives and holding size), access modality to retrieve information on projects, metadata schema adopted. To conclude a discussion on the issues related to CERIF adoption and/or integration is outlined.

2 Methods

Currently the OpenDOAR directory lists more than 2000 open access archives worldwide providing access to a variety of repositories (institutional, disciplinary, governmental, aggregating) developed by universities, research institutions, funding agencies, governmental institutions and publishers. Information on the archives is submitted by the providers, who register the archives, and then checked and categorized by the OpenDOAR staff. These categories allow users to sort the listed archives according to different criteria. We used the option "dataset" reported in the OpenDOAR content type category to identify our first sample of analysis. The latest update of our survey was completed in March 2012.

To identify archives, that contain both research data and projects we adopted the following strategy:

- Identification of archives that contain research data searching and/or browsing for content type, if the archives had these functionalities. Otherwise a direct analysis of the collections reported in the repository was performed to verify the presence of research data in the archives;
- For each archive that contained research data a search for the term *project* was performed usually using advanced search functionalities. In particular this search strategy was used to verify whether:
    - A specific variable for projects was foreseen, resulting in search functionality and/or a specific field label;
    - The term project was present in any other field of the research data bibliographic description;
    - Any other information describing a project was present either as metadata and/or as a free text.

In archives that provided separate lists of projects, the correspondence of information contained in the dataset bibliographic description and in the project list was verified.

For the purpose of our analysis we did not adopt the OpenDOAR classification of repositories (institutional, disciplinary, governmental and aggregating), but we preferred to distinguish between repository (including both institutional and disciplinary repositories) and data services. In this way we can differentiate between *traditional* repositories that collect research outcomes produced within an institution and/or in a given disciplinary field from those archives that have a more general scope. In fact, the first ones generally collect research data along with other digital objects (journal articles, books, conference papers, reports, etc.) and provide an indication on whether and how the request for free access for research data is achieved. The second ones provide access to multiple resources generally linked in a variety of internal and/or external databases, web pages and tools.

3 Results

3.1 The sample

In OpenDOAR there are 79 archives that claim to contain research data in their content type. Our analysis verified that 43 out of 79 archives actually contain research data, while references to projects are present only in 12 archives out of 43. 6 archives listed in OpenDOAR were not accessible. Thus, our sample of analysis focuses on the analysis of 12 archives registered in OpenDOAR containing both research data and references to projects (fig. 1).

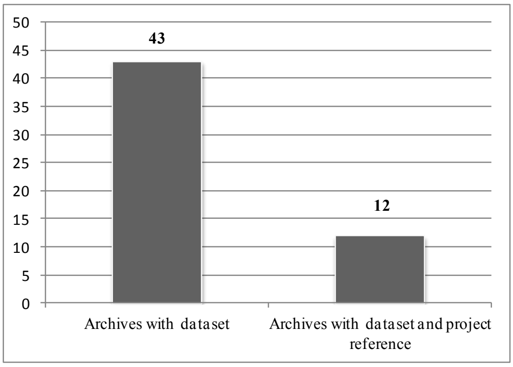

Figure 1: OpenDOAR data archives

3.2 Characteristics of the sample

Table 1 provides an overview of the 12 archives analysed. As previously mentioned we classified our sample in two broad categories: data services and repositories to distinguish between archives developed in response to the free diffusion of different types of institutions' research results and those that tend be a reference point for a specific scientific community and provide access to different information sources and services. In our sample data services and repositories are equally distributed. The majority of the archive providers are universities and research institutions, but the sample also encompasses an indexing abstracting service and a publisher consortium. In the classification of providers we have also introduced the category consortium, as collaboration between different institutions may become a common practice in the development of research data archives. In our sample consortia have been developed among publishers for the development of Dryad, an archive that connects dataset and peer-reviewed articles, and among research institutions as in the case of PANGAEA, an important archive in the field of environmental sciences.

Table 1: Major characteristics of the sample of analysis

Type of Archive	Dataset Archive	Provider	Type of Provider	Disciplinar fields	Funded
Data Service	ADS	Archaeology Data Service	Indexing/abstracting service	Archaelogy	yes
	Dryad	NESCent (National Evolutionary Synthesis Center)	Publisher corsortium	Biochemistry	yes
	IFPRI	International Food Policy Research Institute (IFPRI)	Research institution	Agriculture/ Environmental sciences	no
	OSTI	DOE (U.S. Department of Energy)	Research institution	Multidisciplinary	no
	PANGAEA	Alfred Wegener Institute for Polar and Marine Research (AWI), Center for Marine Environmental Sciences (MARUM), University of Bremen, Germany	Research consortium	Environmental sciences	yes
	Verkehrsmodelle	Deutschen Zentrum für Luft- und Raumfahrt	Research institution	Mobility/Transport	no
Repository	CEDA	Centre for Environmental Data Archival (CEDA), STFC Rutherford Appleton Laboratory	Research institution	Multidisciplinary	yes
	dLIST	School of Information Resources & Library Science, University of Arizona (UA)	University	Multidisciplinary	no
	Dspace @ Cambridge	Cambridge University Library and Computing Service, University of Cambridge	University	Multidisciplinary	no
	Edinburgh DataShare	Data Library, University of Edinburgh	University	Multidisciplinary	yes
	IAI Search	Inter America Institute for Global Change Research (IAI)	Research institution	Environmental sciences	no
	WHOAS	MBLWHOI Library Marine Biological Laboratory & Woods Hole Oceanographic Institution (MBL & WHOI)	Research institution	Environmental sciences	no

If we consider disciplinary fields, there is a consistent number of archives that cover different aspects of environmental sciences. ADS contains data on archaeological excavations and English heritage assets, while the German data service Verkehrsmodelle provides access to data on mobility and transport. Repositories usually have a multidisciplinary vocation, as they tend to reflect research activities and results within the whole institution.

Interestingly a significant part of these archives (5 out of 12) have been developed through specific funding initiatives and this indicates the increasing attention going to the curation and preservation of research data.

It is quite difficult to give an idea of the holding size of the archives we analyzed, since this type of measurement depends on many factors, such as the size of the institution and related community (Marcial & Hemminger 2010) or how long the archive has been operative. Moreover, in research data archives, the size is also related to *what constitutes a dataset* (DOE, 2012*)*. According to DOE definition “A dataset may be one file or may contain many files and the files may include information in various media and formats”. The latter case is generally related to projects in which “continuously-running instruments are acquiring data over a period of time and are stored “together” as a datastream”. Therefore in the classification of the archive size we considered both the number of dataset items and the datastream that generally contains a large quantity of research data.

Table 2: Distribution of archives by size, project references and proportion between archive size and project references

Type of Archive	Dataset Archive	Dataset item	Datastream	Dataset size	Project references	Indicator of correlation
Data Service	ADS	√	√	large	many	high
	Dryad	√		large	few	low
	IFPRI		√	large	few	low
	OSTI		√	large	few	low
	PANGAEA	√	√	large	many	high
	Verkehrsmodelle	√		large	few	low
Repository	CEDA	√		small	few	low
	dLIST	√		small	few	high
	Dspace @ Cambridge	√	√	small	few	high
	Edinburgh DataShare	√		small	few	high
	IAI Search		√	large	many	high
	WHOAS	√		large	many	high

Given the heterogeneous sample of analysis, we categorized the holding size very broadly in large archives (containing more than 1000 datasets and/or more than 100 datastreams) and small (less than 1000 and/or less than 100 datastreams). Similarly, we considered the amount of project references and categorized them under many project references (more than 50 occurrences) and few projects (less than 50 occurrences). To find out whether there was a relation between size of the archive and project references we calculated the proportion between the two values (high = more than 0,5, low = less than 0,5). The majority of archives in our sample can be qualified as large research data archives, most of which are data services. The amount of referenced projects is high in a minority of archives, but this does not depend on the type of archive since they are equally distributed between repositories and data services. The proportion between the size of archive and project references, that can be considered an indicator of correlation, shows that a high proportion does not depend exclusively on the size of the archives. In our sample the majority of archives that make datastreams available have large holding size, a high indicator of correlation, while project references are equally distributed.

3.3 Accessing project information

Each of the analysed archives has its own way of organising information. For instance repositories are generally developed using software such as E-print, DSpace or Fedora so that the archive content is presented according to the information model envisioned by each of them (community collection in Dspace, or content type in E-print). However, archives that have developed in-house software, that address specific requirements depending on the type of content and users may have developed specific information models as well as their own metadata schema. This is often the case in data services that may allow access to a variety of information sources available for a specific scientific community that may have developed common metadata schema to share research data.

Table 3: Access modality distributed by system functionality

Type of Archive	Dataset Archive	Direct access			External link
		Separate project list	Search for project title	Browsing for project	Project' web page
Data Service	ADS	√			
	Dryad				
	IFPRI				√
	OSTI				√
	PANGAEA	√	√	√	√
	Verkehrsmodelle		√		
Repository	CEDA				√
	dLIST				√
	Dspace @ Cambridge				√
	Edinburgh DataShare				√
	IAI Search	√			√
	WHOAS				√

Table 3 shows how project references were retrieved in our analysis and indirectly it also outlines how the different archives organise information related to both projects and research data. This analysis led to the distinction between archives that have direct access modality to retrieve information about projects and those where references to projects are contained in the research data bibliographic description, but do not have specific metadata to make them retrievable (cfr. Table 4). In a minority of cases project information can be retrieved accessing separate archives and/or lists made available through specific menus. This is the case of ADS and PANGAEA that allow users to access projects from a clearly visible menu. However, some differences are worth noting, and these may depend not only on the information model adopted, but also on the disciplinary field covered by those archives, that is archaeology and environmental sciences. In PANGAEA the retrieval of the project list provides an overview that details the acronym and name of the project, the name of a contact person and the research data associated to the project. Each of these variables provides access respectively to the project's web page, to the e-mail of the contact person and to the bibliographic description of the research data. In ADS the project archives provide a list of titles associated with the name of the institution and/or author as well as with the date of the latest additions. The selection of the project from this list gives access to a description that depends on the type of item described (for instance the collection of archaeological objects found during an excavation, or a report describing special excavation methods and measurements). Therefore, information about projects is not uniformly reported within this special project archive.

Similarly to PANGAEA, the repository IAI and the Verkehrsmodelle archive always include the project variable in the bibliographic description, providing a uniform metadata scheme of research data. A specific variable for project title also enables its inclusion in the search functionality, thus allowing users to retrieve project titles associated with research data and related documents. It is noteworthy that in the majority of the archives in our sample (8 out of 12 archives) references to projects, in whatever form, have an external link to the project web page, thus providing direct source information.

When the archives in our sample do not foresee a specific variable for project title, project references can be retrieved indirectly, i.e. from a search of the term *project* performed in simple and/or advanced search modes (table 4). In most cases the term project and its related title is part of the title of the research data described. Abstracts are also fields in which the project is mentioned, sometimes providing its title and brief description, or reporting its grant number. However the title of the project can be retrieved in other description fields, such as producer, publisher, copyright owner, or subjects and notes. Note that the fact that the same archive may use different fields to report the project title is to the detriment of bibliographic uniformity but also biases information exchange among different archives.

Table 4: Project references distributed by bibliographic fields

Type of Archive	**Dataset Archive**	**Fields with project occurrences**						
		title	**abstract/ description**	**publisher**	**producer**	**subject**	**right**	**note**
Data Service	ADS	√	√					
	Dryad		√					√
	IFPRI		√		√			
	OSTI	√	√					
	PANGAEA		√					
	Verkehrsmodelle	√						
Repository	CEDA	√	√					
	dLIST	√	√					
	Dspace @ Cambridge			√			√	
	Edinburgh DataShare		√					
	IAI Search	√	√					
	WHOAS	√	√	√		√		

3.4 Metadata schema

Table 5 shows the software used to develop the archives as well as the metadata schema used to describe research data. Repositories that usually collect research data in different disciplinary fields along with other research results usually adopt Dublin Core (DC) and are generally developed on broadly diffused software platforms, such as DSpace and E-print.

On the contrary, the majority of data services that are specifically focused on the management and diffusion of research data have developed *ad hoc* software and their own metadata schema. Efforts in the construction of common metadata schema for research data are evident when related documentation is analysed. Dataverse is based on the Dataverse Network an open source application to "publish, share, reference, extract and analyze research data" (Dataverse, 2012). Partners of the Dataverse Network Project belong to the Data Preservation Alliance for Social Sciences (Data-PASS, 2012). Moreover, Dataverse metadata are compliant with the Data Documentation Initiative (DDI, 2012), with Simple Dublin Core, and with Content Standards for Digital Geospacial Metadata (CSDGM, 1998). In addition, IAI Search metadata schema is based on CSDGM, while ADS uses it for the location of archaeological sites. Moreover, metadata of the ARCHSEARCH archive in ADS is based on the UK standard MIDAS Heritage (MIDAS, 2007) developed to support information sharing and preservation on information about historic environments. We should not overlook the metadata model developed in PANGAEA, which at its uppermost level has the class *Project* with its related metadata (acronym, name, project coordinator, institute of coordination or project office, Uri, etc.) (wiki.pangaea, 2012). Besides these characteristics that make this data model comparable with the CERIF model, the PANGAEA consortium is also very active in the promotion of a persistent identifies (DOI) for research data that will permit reliable access and make data citable as publications. Moreover, compliance with other metadata standards (ISO 19115, DIF, DC, etc.) allows interoperability between different systems.

Table 5: Dataset archives distributed by type of software and metadata schema adopted

Type of Archive	Dataset Archive	Software	Metadata
Data Service	ADS	in-house software	ADS metadata template
	Dryad	Dspace	DC
	IFPRI	DRUPAL	Dataverse
	OSTI	DOE Data Explorer (DDE)	N/A
	PANGAEA	Sybase	PANGAEA data model
	Verkehrsmodelle	Eprints	DC
Repository	CEDA	Eprints	DC
	dLIST	Dspace	DC
	Dspace @ Cambridge	Dspace	DC
	Edinburgh DataShare	Dspace	DC
	IAI Search	Mercury	IAI schema based on FGDC-CSDGM standard
	WHOAS	Dspace	DC

4 Discussion and implications for CERIF

The explorative, qualitative analysis carried out on the heterogeneous sample of data archives listed in OpenDOAR can provide some indications of the role that CERIF can play within the context of data archives. A standardized metadata description of projects linked with a new CERIF entity Result_dataset provides the context in which research data are produced and certainly improves its discovery and re-use linking different types of data acquired in the same or related projects.

In our sample of analysis a minority of archives used a specific variable to uniquely identify project ID, title and acronym (CERIF attributes of the core entity Project) and had specific functionality to make this information retrievable. Among them PANGAEA provides the best example, being based on a conceptual model, which foresees at its uppermost level the class Project. Archives that have developed their own metadata schema (ADS, PANGAEA, IAI, IFPRI) have a large holding size and tend to make references to many projects and to have a high correlation indicator. They are usually compliant with other metadata standards. No evident communality can be traced when considering the disciplinary fields they cover. However, even if the number of archives analyzed does not allow us any generalization, data archives (PANGAEA, IAI) dealing with environmental data are a good test bed to verify suitable metadata to describe the connection between research data and project and test CERIFcation and/or its compatibility.

The repositories that manage different digital objects generally have a small holding size, make few references to projects, but have a high correlation indicator. Most of them use qualified DC and do not foresee specific variables to describe projects, so that some DC core elements are not used uniformly and have different semantics. For instance DC core elements, such as dc.publisher, dc.subject dc.right are used to report the project titles, but also the name of the organization responsible for the project. On the contrary a qualified DC element dc.description.sponsorship is used uniformly to provide the name of the funding organization.

Other variables that describe the role of the person and organization involved in projects are not considered either in archives that adopt their own metadata schema or in repositories using DC. In this perspective the rich relationship of the CERIF model together with the well defined semantics can enhance both research data archives and CRIS, making a clear distinction for instance between the data authors and owners, their affiliations, or the role played by an institution as owner of copyright of the data, collaborating organization or funding agency.

References

C4D (2012): C4D project CERIF for datasets, D2.1 *Metadata ontology.* JISC Feb. 2012

CSDGM (1998): Metadata Ad Hoc Working Group. Federal Geographic Data Committee. *Content Standards for Digital Geospacial Metadata*. FGDC-STD-001-1998. URL: http://www.fgdc.gov/metadata.

Dataverse Network (2012): The Dataverse Network Project. Version 2.1.2. URL: http://projects.iq.harvard.edu/thedata/book/learn-about-project

DDI (2012): Data Documentation Initiative. *Specification and Documentation*. URL: http://www.ddialliance.org/

DOE (2012): DOE Data Explorer. Frequently asked questions. URL: http://www.osti.gov/dataexplorer/faq.html

Lynch, Clifford A. (2007): The Shape of the Scientific Article in the Developing Cyberinfrastructure. *CTWatch Quarterly*, Vol 3, No 3, p.5-10. URL: http://www.cni.org/publications/cliffs-pubs/shape-of-the-scientific-article/

Marcial, Laura Haak, Hemminger, Bradley M. (2010): Scientific data repositories on theWeb: an Initial Survey. *Journal of the American Society for Information Science and Technology*, Vol 61, No 10, p. 2029-2048. URL: http://onlinelibrary.wiley.com/doi/10.1002/asi.21339/pdf

MIDAS (2007): MIDAS Heritage. *The UK historic environment data standard*, Part1, 2, 3. URL: http://www.english-heritage.org.uk/professional/archives-and-collections/nmr/heritage-data/midas-heritage/

National Science Foundation (2005): National Science Board. *Long-lived digital data collections: enabling research and education in the 21st century*. Washington: National Science Foundation. URL: http://www.nsf.gov/pubs/2005/nsb0540/start.jsp

National Science Foundation (2010): Press release 10-007. URL: http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

OpenDOAR (2012): The Directory of Open Access Repositories. URL: http://opendoar.org/

PANGAEA (2012): PangaWiki *Data Model* URL: http://wiki.pangaea.de/wiki/Data_model

PANGAEA (2012): PangaWiki *STD-DOI.* URL:http://wiki.pangaea.de/wiki/STD-DOI

Contact Information

Daniela Luzi

National Research Council
Institute for Research on Population and Social Policies
Via Palestro, 32
00185 Rome
Italy

d.luzi@irpps.cnt.it