	PRJ/EPOS
Proposed Metadata Stack for EPOS	**Keith G Jeffery 20111019**

Introduction

Metadata – it is commonly said – is **data about data**. In fact what is metadata to one application may be data to another so the distinction is not valid. It is all data but some may be used as metadata related to others.

Metadata has been divided into Descriptive, Structural and Administrative used respectively for ***discovery***, ***understanding*** components of the described object and for ***reading it*** technically.

An alternative view considers Schema, Navigational and Associative metadata, the latter divided into Descriptive, Restrictive and Supportive. Schema metadata assures integrity, "navigational" metadata locates the object, "descriptive" and "restrictive" allow for discovery and manage rights and/or charges respectively while "supportive" provides assistance in utilisation through dictionaries, thesauri and domain ontologies.

OAIS (Open Archival Information System, e.g. http://en.wikipedia.org/wiki/Open_Archival_Information_System) considers packages of metadata for the purposes of digital preservation.

One problem with metadata is that there are many 'standards' commonly erected by a group of enthusiasts in a given domain. Metadata 'standards' can be specific to an experiment or a scientific mission, or may be universally applicable. *For the purposes of EPOS we require metadata that provides **unambiguous**, **secure legal access to open public data** such that it can be utilised for the envisaged purpose of the end-user.*

The principles are based on experience of best practice over many years by the project team. The key principles are:

1. **The metadata should comply with more basic standards** wherever possible in the priority International de jure (ISO), International broad consensus (e.g. W3C), International de facto (usually commercial e.g. PDF), International within a domain (e.g. International Geophysical Year Standards, SEED, …), Continental or National but adopted widely (e.g. ANSI);

2. The metadata must have a **formal syntax** and declared semantics so that it is **machine-understandable** as well **as machine readable / duplicable**;

3. The metadata must be fit for purpose for the following functions related to a digital object:

    a. discovery,

    b. reading,

    c. ingesting,

    d. combining,

    e. processing,

f. outputting as a modified version,

g. citing,

h. preserving, (all related to organisations, projects, persons, other datasets, publications etc and located in space and time);

There are many metadata 'standards', each has a niche 'market'. **A key requirement of any metadata stack is interoperability with other metadata standards to allow homogeneous access over heterogeneous data, software and services.** The proposed stack is capable of interoperating with at least DC (Dublin Core, see http://dublincore.org/), MARC, MODS, e-GMS, US International Dataset Standard, INSPIRE, CSMD all of which are more-or-less relevant to research (scientific) datasets and associated metadata.

The Proposed Stack

Following (a) the principles and (b) analysis of the candidate metadata standards it is clear that for the purpose of EPOS we need:

1. A simple 'flat' metadata standard for *discovery*; (flat metadata means it is a single record with attributes rather than a group of linked records each with attributes and with relationships between the records)

2. A structured (linked entity) standard for context (relating the dataset to provenance, purpose, environment in which generated etc);

3. Detailed metadata standards for each kind of data to be co-processed;

It is expected that (1) can be generated from (2).

An example indicating how DC is a proper subset of Formalised DC (as implemented within CERIF follows:

DC	**Formalised DC**
	<UNIQUEID> RAL92-003 </ UNIQUEID >
<TITLE> A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent, Heterogeneous Information Systems </TITLE>	<TITLE> <language> en </language> <title> A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent, Heterogeneous Information Systems </title> </TITLE>

<CREATOR> Naldi F, Jeffery K G, Bordogna G, Lay J O, Vannini-Parenti I</CREATOR>	<PERSON><role>author</role><person>Naldi F</person></PERSON> <PERSON><role>author</role><person> Jeffery K G</person></PERSON> <PERSON><role>author</role><person>Bordogna G</person></PERSON> <PERSON><role>author</role><person>Lay J O</person></PERSON> <PERSON><role>author</role><person>Vannini-Parenti I</person></PERSON>
<SUBJECT>Current Research Information Systems; legacy; heterogeneous; distributed; protocol; communications; data; exchange</SUBJECT>	<SUBJECT><language>en</language> <scheme> RALClassification </scheme> <subject> Current Research Information Systems </subject> </SUBJECT>
	<KEYWORDS> <language> en </language> <scheme> UKThesaurus </scheme> <keywords> legacy; heterogeneous; distributed; protocol; communications; data; exchange </keywords> </KEYWORDS>
<DESCRIPTION>A system named EXIRPTS has been built which demonstrates access over distributed multilingual information systems of R&D projects. The system resolves problems of resource location and utilises a catalog technique for metadata which allows the end-user to have a homogenous view over heterogeneous information</DESCRIPTION>	<DESCRIPTION> <language> en </language> <description> A system named EXIRPTS has been built which demonstrates access over distributed multilingual information systems of R&D projects. The system resolves problems of resource location and utilises a catalog technique for metadata which allows the end-user to have a homogenous view over heterogeneous information </description> </DESCRIPTION>
<PUBLISHER>Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, OX11 0QX UK </PUBLISHER>	<ORGUNIT><role>publisher</role><orgunit> Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, OX11 0QX UK </orgunit> </ORGUNIT>
< CONTRIBUTOR> Wright, L, Daniels,T </CONTRIBUTOR>	<PERSON> <role> contributor </role> <person> Wright, L </person> <role> proofreader </role> <person> Daniels, T </person> </PERSON>
<DATE>1992</DATE>	<COVERAGE TEMPORAL> <project> 1988-1991 </project> <publication> 1992 </publication> </COVERAGE TEMPORAL>
<TYPE>Technical Report</TYPE>	<RESOURCETYPE> <scheme> RALLibrary </scheme> <language> en </language> <resourcetype> Technical Report </resourcetype> </RESOURCETYPE>
<FORMAT>Word2</FORMAT>	(note handled by conventional MIME typing)
<IDENTIFIER>RAL 92-003</IDENTIFIER>	<RESOURCEIDENTIFIER> <scheme> RALLibrary </scheme> <resourceidentifier> RAL92-003 </resourceidentifier> <scheme> referencelist </scheme> <resourceidentifier> [NaJeBoLaVa92] </resourceidentifier> </RESOURCEIDENTIFIER>
<SOURCE > [null]	Note: done using relationships between resources referenced by UniqueId

<RELATION> [JeLaMiZaNaVa89] </RELATION>	<uniqueid> <RAL92-003> </uniqueid> <role> preliminary investigation </role> <uniqueid> [JeLaMiZaNaVa89] </uniqueid>
<COVERAGE> Europe,1983-1991 </COVERAGE>	<COVERAGE SPATIAL> <scheme> LatLong </scheme> <coordinates>10W35N-30E80N </coordinates> <precision> 5degrees </precision> </COVERAGE SPATIAL> <COVERAGE TEMPORAL> <scheme> years </scheme> <constraints> [1983<x>1991] </constraints> </COVERAGE TEMPORAL>
<RIGHTS> Copyright Rutherford Appleton Laboratory 1992 </RIGHTS>	(note handled separately with access, privacy security etc)

The recommended way forward is:

1. Discovery: **DC**
2. Contextual: **CERIF** (Common European Research Information Format, http://en.wikipedia.org/wiki/CERIF)
3. Detailed: **Individual standards depending on type of dataset**; for research datasets from large-scale facilities CSMD (e.g., http://www.ijdc.net/index.php/ijdc/article/view/149; see also PaNData, http://www.pan-data.eu/PANDATA_–_Photon_and_Neutron_Data_Infrastructure), for geospatial datasets INSPIRE[1] (http://inspire.jrc.ec.europa.eu/, http://en.wikipedia.org/wiki/INSPIRE as in ENVRI).

DC for discovery can be generated from CERIF so assuring consistency especially in semantics (DC is notably imprecise in semantics despite the introduction of qualified DC).

CERIF can provide the contextual data surrounding links (URLs) to individual detailed metadata standards for particular domains of science with their data, software and services in particular data-centres.

[1] INSPIRE is based on the infrastructures for spatial information established and operated by the 27 Member States of the European Union. The Directive addresses 34 spatial data themes needed for environmental applications, with key components specified through technical implementing rules. This makes INSPIRE a unique example of a legislative “regional” approach.

Metadata Standards Recommended

DC and CERIF are now described.

DC (Dublin Core)

Reference

http://dublincore.org/documents/dces/

Introduction

DC is widely used as a simple metadata standard to catalogue web pages or web resources. It has developed to a 15-element set from the original 13 and more recently utilised namespaces to avoid ambiguity of lexical terms across elements.

The Elements

Term Name: contributor	
URI:	http://purl.org/dc/elements/1.1/contributor
Label:	Contributor
Definition:	An entity responsible for making contributions to the resource.
Comment:	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity. For seismology: Mario Rossi, INGV, mario.rossi@ingv.it, head of the MN network. Is this related to who is taking care of the curation of the resource – I am assuming that the resource is the digital object.
Term Name: coverage	
URI:	http://purl.org/dc/elements/1.1/coverage
Label:	Coverage
Definition:	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.

Comment:	Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges. For seismology: should we put, for example, the time of operation of the network (from 1989 to present) and the location of the individual stations ????
References:	[TGN] http://www.getty.edu/research/tools/vocabulary/tgn/index.html
Term Name: creator	
URI:	http://purl.org/dc/elements/1.1/creator
Label:	Creator
Definition:	An entity primarily responsible for making the resource.
Comment:	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity. For seismology: the institution running the network ??
Term Name: date	
URI:	http://purl.org/dc/elements/1.1/date
Label:	Date
Definition:	A point or period of time associated with an event in the lifecycle of the resource.
Comment:	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF]. For Seismology: would this be an event within the continuous data window. For example 10 minutes of data from… to ….
References:	[W3CDTF] http://www.w3.org/TR/NOTE-datetime
Term Name: description	
URI:	http://purl.org/dc/elements/1.1/description
Label:	Description
Definition:	An account of the resource.
Comment:	Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource. For Seismolgy: Would it be OK a plot ?
Term Name: format	
URI:	http://purl.org/dc/elements/1.1/format
Label:	Format
Definition:	The file format, physical medium, or dimensions of the resource.

Comment:	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME]. For Seismology: we use various formats and the most common is SEED for continuous data and SAC for windowed data ready for the analysis. The forst is a standard whereas the second can be substituted with many others.
References:	[MIME] http://www.iana.org/assignments/media-types/
Term Name: identifier	
URI:	http://purl.org/dc/elements/1.1/identifier
Label:	Identifier
Definition:	An unambiguous reference to the resource within a given context. For Seismology: could it be the URI we create when the windowed data are extracted from the continuous and provided to the user requesting them ?
Comment:	Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
Term Name: language	
URI:	http://purl.org/dc/elements/1.1/language
Label:	Language
Definition:	A language of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as RFC 4646 [RFC4646]. For seismology: we do not have language (…but the Earth talks to us using waves ☺)
References:	[RFC4646] http://www.ietf.org/rfc/rfc4646.txt
Term Name: publisher	
URI:	http://purl.org/dc/elements/1.1/publisher
Label:	Publisher
Definition:	An entity responsible for making the resource available.
Comment:	Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity. For seismology: the data center or the network operator ?
Term Name: relation	
URI:	http://purl.org/dc/elements/1.1/relation
Label:	Relation
Definition:	A related resource.
Comment:	Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system. Not clear what this is.
Term Name: rights	
URI:	http://purl.org/dc/elements/1.1/rights
Label:	Rights
Definition:	Information about rights held in and over the resource.

Comment:	Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights. For seismology: open access data with a few constraints on their use for private companies.
Term Name: source	
URI:	http://purl.org/dc/elements/1.1/source
Label:	Source
Definition:	A related resource from which the described resource is derived.
Comment:	The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the identification system. For seismology: this could refer to the original continuous data stored at the data center ????
Term Name: subject	
URI:	http://purl.org/dc/elements/1.1/subject
Label:	Subject
Definition:	The topic of the resource.
Comment:	Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element. For seismology: ???
Term Name: title	
URI:	http://purl.org/dc/elements/1.1/title
Label:	Title
Definition:	A name given to the resource.
Comment:	Typically, a Title will be a name by which the resource is formally known.
Term Name: type	
URI:	http://purl.org/dc/elements/1.1/type
Label:	Type
Definition:	The nature or genre of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element. For seismology: ????

Overall evaluation and suitability for EPOS

DC has been criticised [Je99] because of its lack of formality. Put simply it breaks functional referential integrity in that elements like creator or contributor do not depend functionally on the primary key (ID) of the resource being described. Worse, it is hard to consistently use DC and there are wide variations in interpretation of what value should be encoded under contributor or creator; similar problems occur with source and reference. DC is inappropriate to provide the detail of metadata required for automated or semi-automated access to –and utilisation of – datasets. However, its widespread use (albeit in many

different semantic interpretations and even syntactic structures) means that we should utilise the option to generate qualified DC from the CERIF metadata standard for EPOS to ensure consistent semantics within a formalised syntax.

CERIF (Common European Research Information Format)

Reference

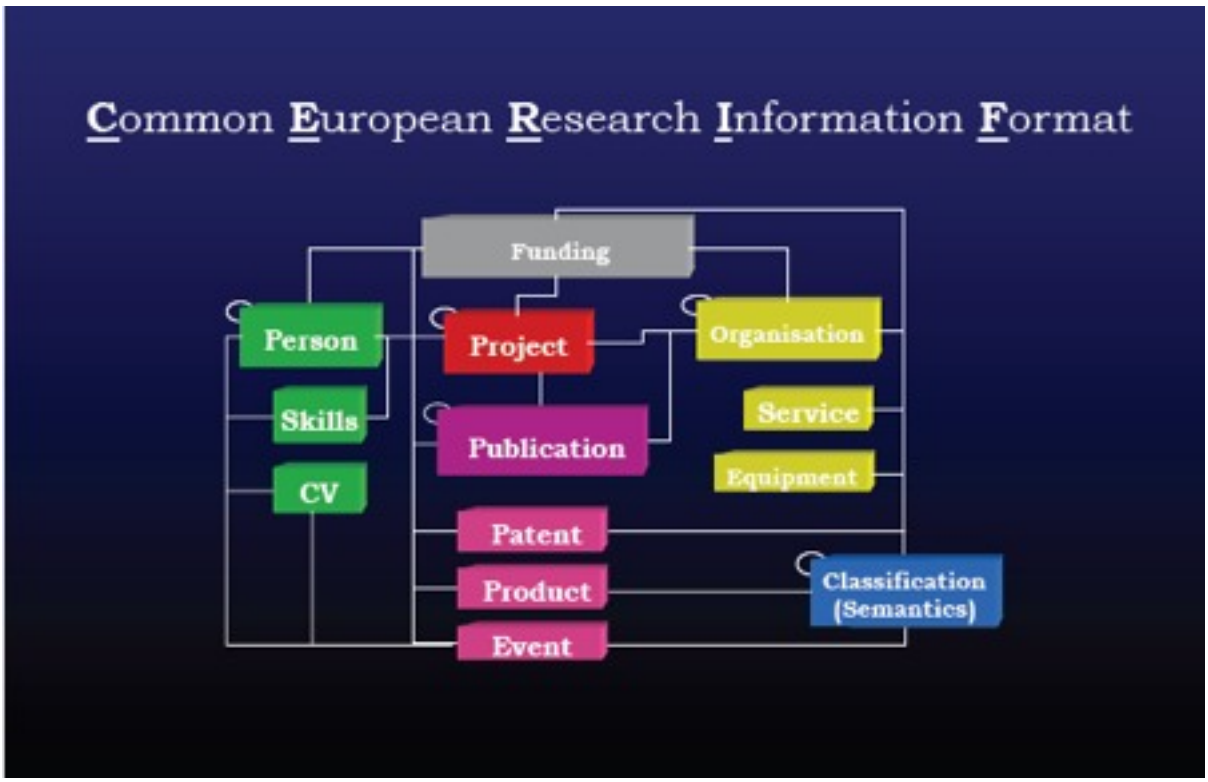
www.eurocris.org

Introduction

CERIF (Common European Research Information Format) was developed under the auspices of the EC and is now maintained, developed and promoted by euroCRIS www.eurocris.org . It is an EU recommendation to member states used widely as the national system for research information in 8 countries and in many organisations in other countries. It is a structured metadata standard covering persons, organsational units, projects, funding, research outputs (publications, products, patents), events, skill and expertise, facilities and equipment, services and other entities.

Elements

It is best visualised as a diagram (because of the relative structural complexity):



Overall evaluation and suitability for EPOS

CERIF is an EU Recommendation to ember states for research information. It is used as the national system for research information management in 8 countries and used widely or

being evaluated in others, including countries outside Europe such as USA, Canada, Korea, India, Australia, Brazil and surrounding countries. There are four commercial companies offering CERIF systems.

CERIF provides highly structured (formal syntax and declared semantics) contextual metadata capturing relationships of entities (e.g. project, persons, organisations, funding, publications, products, patents, facilities, equipment, services, events, etc.). Relationship semantics are clearly defined and there is an explicit exntesible mechanism within the CERIF specifications (CERIF Semantic Layer) to allow the definition of any type of relationships among CERIF entities.

A standard serialisation of CERIF in XML is available, while the euroCRIS CERIF Task Group is currently developing a standard mechanism for providing CERIF data as Linked Data that is scheduled for incorporation into CERIF in 2011.

A concern regarding applicability to EPOS is that CERIF covers only the metadata required to describe research objects of interest (although it has been used for directory services, managing customer relationships and managing e-infrastructure). It does not include the detailed metadata to describe datasets to a level where automated program access is possible. However, the Product entity already available in CERIF can be directly used to represent datasets immediately covering the majority of metadata requirements for datasets by EPOS, as demonstrated by an initial exercise within the ENGAGE project for mapping common government datasets standards to CERIF. Furthermore, any extensions to CERIF that could be useful for representing open datasets can be contributed to the specification y proposing them to euroCRIS.

The Architecture

It is envisaged that the EPOS architecture will consist of

a) **A portal for user access** with querying and browsing capability, also with access (via the catalogue(s) and GRID technology) to datacentres providing the data download, analysis, display, simulation capability;
b) **One centralised or several mirrored catalogues** each with contextual metadata and exposing discovery metadata linked intimately with the portal;
c) **Many distributed datacentres with datasets**, **software** and **services** described by detailed metadata which is referenced in context from the catalogue(s);

The Rationale and Advantage

The metadata stack and architecture proposed gives EPOS the following major advantages:

1. **It can be superimposed** on existing datacentre datasets, software and services;
2. **It can cover all the types of data, software and services** known within EPOS;
3. **The architecture minimises the impact on existing datacentres**;

    a. The datacentres have only to register their data, software and services with metadata in the catalogue(s)
4. The architecture provides **homogeneous access for an end-user to heterogeneous data**, software and services;
5. **The architecture can interoperate with others** if – for example – North America or Asia chooses a different metadata standard;