Smart Descriptions & Smarter Vocabularies (SDSVoc) Workshop

20161130-1201 CWI Amsterdam

Position Paper: Why CERIF?

Keith G Jeffery (ERCIM), Anne Asserson (euroCRIS)

The Question

The VRE4EIC project uses CERIF as its metadata catalog. It is reasonable to question why.

Background

The connection of applications to datastores has been an enduring challenge. Historically, the vast majority of solutions have been specific to a particular pairing, commonly with a close (programmatic) coupling. Database technology provided a possible solution – with metadata represented by schemata - only to be partially-closed by software theorists demanding compile-time typing. More recently scripting languages have broken out of this restriction allowing execute-time typing thus allowing any software to be coupled with any data by means of some mediating description of the data: metadata.

Metadata is usually associated with data but in fact is equally important in dealing with digital representations of software, organisations, persons, and many other entities – it is all data. In fact, what is metadata to one application may be data to another: a library catalog record is metadata for a researcher finding a digital object but data to a librarian counting digital objects related to geochemistry. Metadata is thus defined by its purpose. Metadata is used for three major purposes: (a) discovery (finding digital objects); (b) contextualisation (assessing relevance and quality and preparing for); (c) action (coupling the entities together to achieve the (business) objective). This relates directly to the FAIR (Findable, Accessible, Interoperable, Re-usable) principles.

In a world of virtualisation (e.g. GRIDs and CLOUDs) it becomes necessary to describe all the relevant components by metadata to allow the development and execution environments to provide the end-user with: (a) a simplified representation of complexity; (b) homogeneous access over heterogeneous computing platforms; (c) middleware to structure and partition the application to optimal deployment. This is an e-I (e-Infrastructure). In an e-RI (e-Research Environment) additionally to virtualising the platforms as indicated above, it is necessary to virtualise the data, software, persons and organisations - in their various roles at various times - products, services and many other entities. In a VRE (Virtual Research Environment) it is necessary to virtualise across the multiple virtualisations of e-RIs themselves using e-Is. This virtualisation (or representation of the world of interest) requires metadata with computer-processable syntax and semantics.

The Problem

Most metadata standards (de jure or de facto) fail to provide the necessary formality and richness to enable their use for the requirements outlined above. DC (Dublin Core), DCAT (Data Catalog Vocabulary) and CKAN (Comprehensive Knowledge Archive Network) are widely used. Other standards such as ISO 19115/INSPIRE for geospatial data are based on DC and share the same problems. In their basic forms (using text or html) they lack referential integrity (there can be multiple occurrences of an attribute related to one unique identifier) and functional integrity (there exist attributes linked to one unique identifier which do not depend functionally upon it). Thus the concept of <creator> may have multiple instances within one metadata record (lacking referential integrity)

and the <creator> does not depend for existence on the digital object identified by the unique identifier. Put simply: several persons may be authors of a publication but with different roles and each person exists (and has other relationships to other digital objects) whether or not they are the author of a particular publication. This information is not represented accurately by simple DC.

The proponents of those metadata standards subsequently realised the problems and progressively introduced (a) qualifiers (which provide a level of sub-structure under each attribute with constraints as to permissible values thus linking to semantics) and usually encoded in XML; (b) RDF (Resource description framework) which provides logical assertions as triples. However, recently a survey indicated >95% of DC is still in text or HTML form thus not demonstrating these advantages. The use of qualified DC is now deprecated (2012) in favour of RDF. DCAT is most commonly encoded in XML while CKAN is usually encoded as RDF. Nonetheless, the problems of referential integrity and functional integrity remain leading to ambiguity in interpretation.

Proposed Solution

In 1987-90 the EC convened a group of national experts to propose a standard for interoperation of information about research. CERIF91 emerged as a simple record structure with attributes, not unlike DC a decade later. Jeffery raised objections but the majority favoured this approach. In 1997 (i.e. around the time of DC publication) the expert group was reconvened following the failure in implementations of CERIF91 and a solution was proposed by Jeffery & Asserson; to use – at conceptual level - extended entity-relationship modelling involving the concept of base entities (like persons, publications, organisations) and linking entities (to represent the relationships between the base entities with role and temporal semantics). The logical and physical implementations could follow any technology supporting the conceptualisation. This proposal became CERIF2000 and is the basis for subsequent developments managed by euroCRIS.

The main characteristics of CERIF are:

(a) it separates clearly base entities from relationships between them and thus represents the more flexible fully-connected graph rather than a hierarchy;

(b) it has generalised base entities with instances specialised by role (for example <person> rather than <author>), the role specialisation is in the linking entities;

(c) it handles multilinguality by design (multiple language representations are linked with role (e.g. machine or human-translated) and temporal information (so representing versions of the translation) to the appropriate attribute treated as an entity – example <title> linked to <publication>;

(d) temporal information is in the link entities not the base entities (example employment between two dates is in the linking relation between <person> and <organisation> and not an attribute of either of the base entities;

(e) the temporal information in linking entities provides provenance and versioning recording e.g. versions of datasets and – in the associated role attribute – the method of update or change;

(f) CERIF separates the semantics into a special 'layer' which is referenced from CERIF instances. The sematic layer includes permissible values for roles in any linking entity (e.g. <person> <author|editor|illustrator|reviewer…> <publication>) and also permissible values for controlled values of attributes in base entities e.g. ISO country codes). Thus semantic terms are stored once and referenced many times (preserving integrity). The semantic layer – like the syntactic layer - consists of base entities (e.g. the valid values for an attribute or valid roles for a linking entity) and linking entities thus allowing relationships between vocabularies and relationships between individual terms to be represented i.e. an ontology. Thus CERIF provides formal syntax and declared (multilingual) semantics.

CERIF has a formal review and update process controlled by euroCRIS and so can evolve. However, it is designed to evolve by accretion (there is a method of adding additional entities and appropriate linking entities to existing base entities) so that the core of CERIF remains constant for interoperability among CERIF installations.

Interoperability

CERIF installations can interoperate naturally. A representation of CERIF in character encoding (CERIF-XML) is defined and used widely in production environments. Since XML represents only a linearised hierarchy, the base entities and linking entities are transmitted as entities and reassembled to the fully connected graph at the receiving installation using the unique identifiers (keys). Over the years euroCRIS members have mapped other metadata standards to CERIF. The experience generally is that the other standards are a subset of CERIF entities and attributes. Thus CERIF has been used as a 'switchboard' for converting from one metadata standard to another thus permitting interoperability between sources with different metadata standards. In the ENGAGE project CERIF-RDF interconversion was demonstrated. It became clear that to represent one CERIF linking entity instance required between 5 and 13 RDF instances, depending on the richness of the CERIF instance. This has performance implications.

Utilisation

CERIF is used widely for representing research activity in universities, funding organisations and related industry. It has been used in >43 countries and on all continents except Antarctica. CERIF is being used outside of the research activity information domain; notably as the virtualisation metadata for one ESFRI research infrastructure (which itself has >400 research facilities) and it is being considered for another. It is being used in VRE4EIC as the metadata for the e-VRE (enhanced virtual research environment). CERIF has stimulated discussion in the RDA (Research Data Alliance) community where its flexibility and formality are of interest in defining a canonical metadata element set – with syntax and semantics – for widespread interoperability.

Conclusion

The flexibility and formality of CERIF provides: (a) a canonical superset metadata standard that can be used to interoperate between other metadata standards; (b) a reliable representation of the world of interest for use in virtualising technologies.