RDA Metadata Principles and their Use	**Keith G Jeffery, Rebecca Koskela 20141114**

Introduction

The four ‘core’ metadata groups of RDA have worked together. The groups are: MSDWG (Metadata Standards Directory Working Group), DICIG (Data In Context Interest Group), RDPIG (Research Data Provenance Interest Group) all coordinated by MIG (Metadata Interest Group). One aspect of that work is to bring forward a set of principles for metadata that the groups believe RDA should adopt and promote.

Metadata Principles

1. The only difference between metadata and data is mode of use
2. Metadata is not just for data, it is also for users, software services, computing resources
3. Metadata is not just for description and discovery; it is also for contextualisation (relevance, quality, restrictions (rights, costs)) and for coupling users, software and computing resources to data (to provide a Virtual Research Environment)
4. Metadata must be machine-understandable as well as human understandable for autonomicity (formalism)
5. Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…)

Illustration and Examples

1. Consider a library catalogue stored electronically. To a researcher it is metadata – using the catalogue finds the book or article. To the librarian it is data: she can count how many books or articles exist on biochemistry compared with clinical medicine.
2. In a VRE (Virtual Research Environment) the amount of work a researcher has to do manually just does not scale. Autonomic services are required. In order to achieve this data, services, users and computing resources need to be described to middleware which manages the scheduling, allocations, connection of the components etc. These descriptions are metadata.
3. Metadata for discovery (followed by manual selection and connection) is already achievable. However the selection of appropriate datasets (or software) is greatly enhanced by using contextual metadata; that is metadata characterising the object of interest. Contextual metadata concerns persons, organisations, projects, funding, outputs (publications, products, patents), facilities and equipment – in short attributes which allow the end user (or software representing the end-user) to assess the relevance and quality of an object (dataset, software) for their current purpose.
4. The mantra is formal syntax and declared semantics. This allows machine processing rather than manual processing.
5. Management metadata links with (3); the contextual metadata can also be used for evaluation of research, policy-making and other management functions at institutional or funding organisation level.

Implications

These principles lead to certain implications:

a) Syntax (metadata standards structures – what they cover)

  a. Objects/entities and properties/attributes

b) Semantics (terms in metadata standards – what they mean)

    a. Relationships between terms including multilinguality

c) Temporal information

    a. Relationships not base information

    b. Provides the temporal interval when the assertion is true

d) Integrity

    a. Referential (represent dependencies)

    b. Functional (all attributes depend uniquely on the unique ID)

e) Represented in some form of first order logic

    a. Allows induction and deduction – saves input and permits brokering

    b. Performance

Illustration and Examples

a) Syntax defines the structure imposed upon the metadata bitstream(s). It allows a computer to know where each entity (object) and attribute (element, property) starts and ends. It has properties of length (or precision for floating point numbers) and type. It is essential for computers to relate the structure to the processing of a software service.

b) Semantics concerns the meaning of terms (lexical values) and the relationships among a term and others such as super- or sub-term, synonyms etc. The relationship between a term in one language and terms in each of other languages allows multilinguality. Semantics may be used in (a) query improvement (making sure the computer system understands what the user means despite what they input); (b) answer explanation; (c) metadata scheme (or schema) matching and mapping; (d) data instance convertors

c) Temporal information concerns the time during which an assertion is true. Example: X is employee of Y between datetime1 and datetime2. Classical temporal data management distinguishes transaction time (when an update occurs) and valid time (when an event happened). In general the temporal information is stored on the instance of the base entity (e.g. TSQL, Richard Snodgrass) but it is advantageous for clearer semantics to associate the temporal interval with a relationship between instances of entities.

d) Integrity is of the greatest importance for a (meta)data representation to reflect the real world. Referential integrity ensures that – in a data structure with one set of attributes dependent on another – it is impossible to delete the set of attributes on which the other is dependent. Functional integrity ensures that every attribute is dependent on the unique identifier of the record. For example <Book ISBN, John Doe> implies that John Doe is author of Book ISBN without declaring it explicitly. In fact John Doe exists independently of Book ISBN (and probably is employed, drives a car, has relatives…) and so his existence is not dependent on the book.

e) First order logic is an area of mathematics/logic that allows great power in computer processing. It can be used to make assertions (i.e. represent an instance of a data structure) but importantly can also be used (a) to define constraints hence improve quality; (b) be used for deduction (facts from rules); be used for induction (rules from facts) and thus, through deduction and induction generate further metadata without the user having to input it.

Interoperability

A major goal of RDA is sharing of research datasets. For this to scale beyond one researcher sending a dataset to another, interoperability is required using computer systems to discover, contextualise, select, access, transmit or process datasets. Interoperability means essentially that a user accessing the world through a local / institutional / national portal sees not only local datasets and software but also all datasets and software known to RDA organisations and members *as if they were local*. This is achieved through the use of metadata characterising the objects (datasets, software, users, computing resources) and techniques to match and map those descriptions leading to generation of convertors for the underlying data instances. There are two main techniques for this: Convergence to common metadata model and Interoperation among many metadata models.

Convergence to a common metadata model implies matching/mapping all existing and planned metadata schemes noting the intersections and differences of entities and attributes and finally generating (a) a superset scheme; (b) the mappings from the originating schemes. This technique has the disadvantage that it cannot be done globally even in any one domain (the scale of the task is too large) unless that domain is vanishingly small. However, it is possible at a certain level of abstraction over one or more domains of research, particularly at discovery level (since only a relatively few attributes are required) and contextual level (where more attributes are required but they tend to be relatively common across research domains).

Interoperation among many metadata models preserves the richness of the original schemes but uses techniques to establish relationships between attributes in the different schemes (matching and mapping). To achieve this the original schemes need to be encoded as rich metadata (see principles and implications above) where brokering technology can be used to achieve the matching and mapping (although usually requiring some human assistance for the first attempt leaving a broker and common scheme that can then be reused thereafter). Once the mapping is done convertors for data instances under the schemes being interoperated can be generated by hand or (semi-)automatically.

It is clear from the above that each technique establishes a common superset scheme over the original schemes and the mappings between them. The common superset scheme may only be populated sparsely from the original schemes depending on the degree of intersection of attributes.

However, the advantage of a common scheme is that it reduces a potential n*(n-1) i.e. almost n**2 set of mappings between every original scheme and convertors to n, i.e. one for each original scheme to the common scheme.

The common schemes can be generalised across multiple domains for each required purpose such as discovery, contextualisation or connection of a dataset to a software service. It is proposed that these general, canonical, common schemes are called 'metadata packages'.

We shall achieve these 'packages' by:

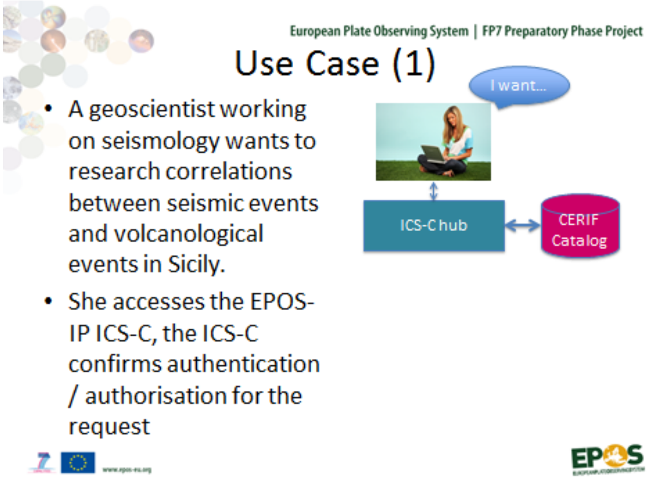
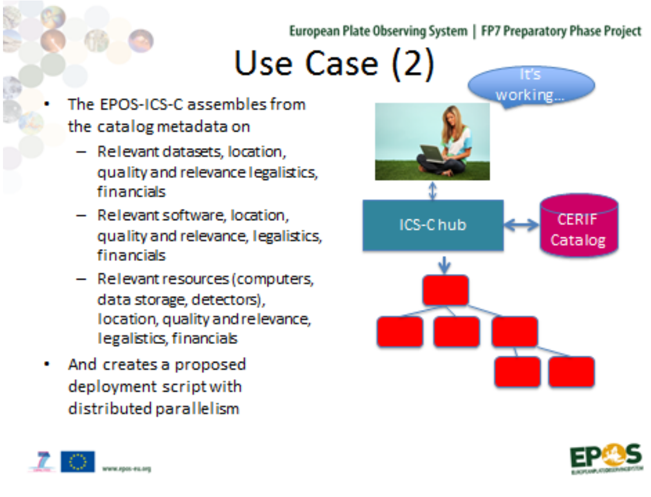
6. Analysing existing metadata standards used in various communities and generally stored in the metadata standards directory working group directory;
7. Analysing the use cases provided from the RDA community for requirements of metadata, i.e. the elements or attributes required for their processing intent;
8. Recording the most commonly used elements allowing for terminology differences including multilinguality;
9. Proposing 'packages' based on these elements for the various purposes and following the principles outlined above.

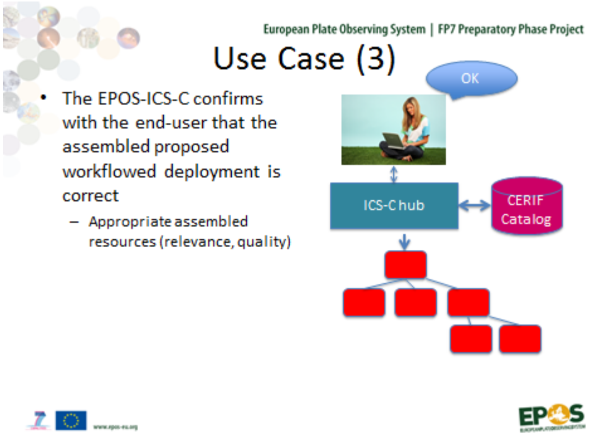
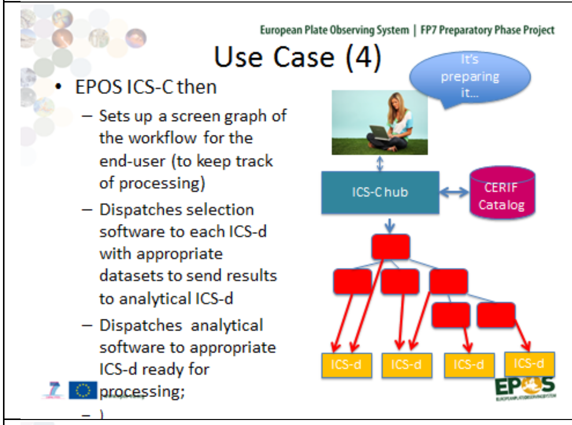
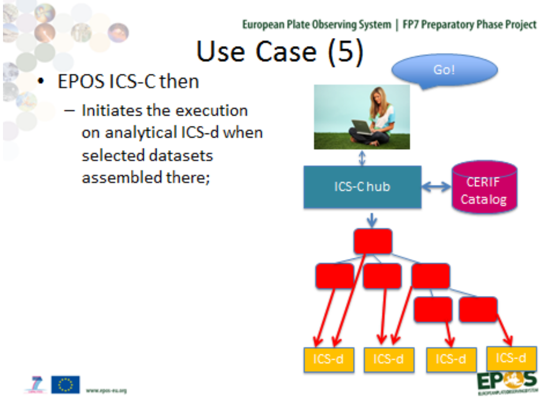
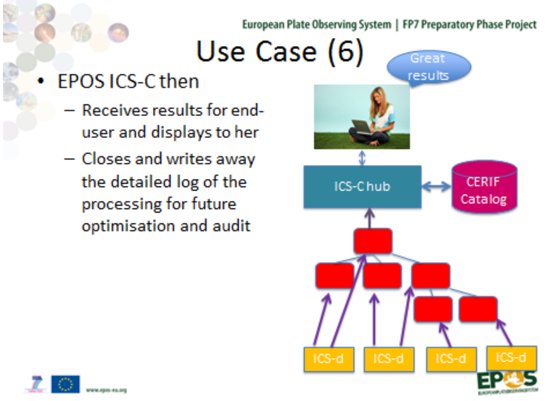
How to achieve interoperability

It is clear from the above that the work plan requires access to RDA repositories (where the work of RDA groups is recorded). The key ones are: (a) the metadata standards directory repository; (b) the use case repository. It also requires access to repositories of data types (data type registry) , data foundations and terminology, PIDs to act as consistency checks.

Use Case

A use case describing utilisation of a VRE (Virtual Research Environment) is presented in steps with information at each step on the metadata required and what is happening. The use case concerns end-user access to the portal of the EPOS (European Plate Observing System) project that is an ESFRI project covering earthquakes and volcanoes in Europe. EPOS has the concept of ICS-C (a portal) and ICS-d (distributed nodes with relevant facilities for EPOS including datasets, computer processing, detector (instrument) arrays, etc). EPOS uses CERIF (Common European Research Information Format) to store its catalog that describes all resources of EPOS (users, datasets, software services, computing services, etc.). CERIF is thus the 'virtualising view of the EPOS world'. The use case is presented as numbered steps.

 European Plate Observing System | FP7 Preparatory Phase Project Use Case (1) • A geoscientist working on seismology wants to research correlations between seismic events and volcanological events in Sicily. • She accesses the EPOS-IP ICS-C, the ICS-C confirms authentication / authorisation for the request I want… ICS-C hub CERIF Catalog www.epos-eu.org	User metadata: user identifier; password; digital signature; organisation and possibly further metadata required for disambiguation and identification. Authentication is done by using associated rights (responsibilities and authorities) metadata.
 European Plate Observing System | FP7 Preparatory Phase Project Use Case (2) • The EPOS-ICS-C assembles from the catalog metadata on – Relevant datasets, location, quality and relevance legalistics, financials – Relevant software, location, quality and relevance, legalistics, financials – Relevant resources (computers, data storage, detectors), location, quality and relevance, legalistics, financials • And creates a proposed deployment script with distributed parallelism It's working… ICS-C hub CERIF Catalog www.epos-eu.org	Discovery metadata: dataset, software or resource: identifier; name; abstract; keywords; characteristics such as size, type; rights information; Contextual metadata such as person; organisation; project; related outputs (publications, products)

 European Plate Observing System | FP7 Preparatory Phase Project Use Case (3) • The EPOS-ICS-C confirms with the end-user that the assembled proposed workflowed deployment is correct – Appropriate assembled resources (relevance, quality) 	Contextual metadata: such as person; organisation; project; related outputs (publications, products)
 European Plate Observing System | FP7 Preparatory Phase Project Use Case (4) • EPOS ICS-C then – Sets up a screen graph of the workflow for the end-user (to keep track of processing) – Dispatches selection software to each ICS-d with appropriate datasets to send results to analytical ICS-d – Dispatches analytical software to appropriate ICS-d ready for processing; –) It's preparing it… ICS-C hub CERIF Catalog ICS-d ICS-d ICS-d ICS-d EPOS	Contextual metadata: describing each ICS-d: the datasets, software, computing resources available
 European Plate Observing System | FP7 Preparatory Phase Project Use Case (5) • EPOS ICS-C then – Initiates the execution on analytical ICS-d when selected datasets assembled there; Go! ICS-C hub CERIF Catalog ICS-d ICS-d ICS-d ICS-d www.epos-eu.org EPOS	The composed workflow (distributed, parallel) of software services with linkages to relevant datasets is executed. The metadata is now bound to the executing software and datasets.
 European Plate Observing System | FP7 Preparatory Phase Project Use Case (6) • EPOS ICS-C then – Receives results for end-user and displays to her – Closes and writes away the detailed log of the processing for future optimisation and audit Great results ICS-C hub CERIF Catalog ICS-d ICS-d ICS-d ICS-d www.epos-eu.org EPOS	Contextual metadata: used for explanation of results if required. Provenance metadata: for later use as contextual metadata (quality, relevance) and for advising on deployment parameters.

Conclusion

The use case demonstrates the requirement for metadata to obey the principles. Furthermore the use case demonstrates that the catalogue used to achieve virtualisation and interoperability must respect the implications of those principles. The use case demonstrates the use of ‘packages’ of metadata elements used for various purposes (discovery, contextualisation, provenance…) although in this use case the ‘packages’ are all stored within the same catalogue.