Research Information Management: the CERIF approach

Keith Jeffery
Keith G Jeffery Consultants
10 Claypits Lane, Shrivenham, SN68AH, UK
Email: keith.jeffery@keithgjefferyconsultants.co.uk

Nikos Houssos
National Documentation Centre / National Hellenic Research Foundation
48 Vassileos Konstantinou Avenue, GR-11635, Athens, Greece
Email: nhoussos@ekt.gr

Brigitte Jörg
JeiBee Ltd
London, UK
Email: brigitte.joerg@gmail.com

Anne Asserson
Anne Asserson
University of Bergen
5020 Bergen, Norway
Email: anne.asserson@fa.uib.no

Abstract: In the context of the wide research environment we introduce the CERIF (Common European Research Information Format) data model which (a) has a richer structure than the usual metadata standards used in research information; (b) separates base entities from link entities thus providing flexibility in expressing role based temporal relationships; (c) defines a distinct semantic layer so that relationship roles in link entities and controlled value lists in base entities are separately managed and multiple vocabularies can be used and related to each other; (d) can generate the common metadata formats used in research information. CERIF is used widely and is an EU Recommendation to Member States. At the request of the European Commission, CERIF is maintained, developed and promoted by euroCRIS.

Keywords: Research Information Systems, research information management, CRIS, CERIF, data infrastructures for e-science, metadata, ontologies, semantics

Corresponding author: Keith Jeffery

Biographical notes: Keith Jeffery is a consultant and past Director International IT Strategy at STFC (Science and Technology Facilities Council) based at Rutherford Appleton Laboratory. Keith previously had strategic and operational responsibility for ICT with 360,000 users, 1100 servers and 140 staff. Keith holds 3 honorary visiting professorships, is a Fellow of the Geological Society of London and the British Computer Society, is a Chartered Engineer and Chartered IT Professional and an Honorary Fellow of the Irish Computer Society. Keith is President of ERCIM and past President of

euroCRIS, and serves on international expert groups, conference boards and assessment panels. He chaired the EC Expert Groups on GRIDs and on CLOUD Computing.
Nikos Houssos works (since 2006) for the Hellenic National Documentation Centre/NHRF as a unit head and technical architect of national scale scholarly communications infrastructures. He is a euroCRIS Board member (since 2009), an active contributor to the development of the CERIF model and is involved in various EU funded projects in the areas of research infrastructures and digital libraries. Nikos holds a B.Sc. and a Ph.D. (2004) in Computer Science from the National and Kapodistrian University of Athens and has served as an adjunct lecturer at the Technical University of Crete (2004-2007).

Dr Brigitte Jörg is Director and Founder of JeiBee Ltd. to provide consultancy with growing demands for standardising Research Infrastructures. She is a euroCRIS Board member since 2004 and more recently also active as a Coordinator with CASRAI. Before setting up JeiBee Ltd., Brigitte was engaged in the Jisc-funded CERIF Support Project in the role of CERIF National Coordinator at UKOLN, University of Bath, UK. From 2001-2012 she worked with DFKI – the German Research Center for Artificial Intelligence in the Language Technology Lab. Prior to her academic work she was more active in industry.

Anne Asserson holds a Master's degree in Information Science from UiB (University of Bergen) and has worked there since 1992. She played a major role in CRIS developmental work: locally, nationally and internationally with special emphasis on data modelling. She was a lead in the development of Norwegian national CRIS systems from FORSKDOC through FDOK to FRIDA and latterly CRIStin. She participated in the CORDIS funded project on 'Best Practice' in 1997 and in the specification of CERIF in 2000. She joined euroCRIS in the early nineties and was a Board member from 2003 to 2012.

1 Introduction

Technological advances related to data collection, networking, storage and management have created a shift towards the paradigm of data-intensive science that is changing the way research is conducted worldwide (Hey, Tansley, and Tolle, 2009). Increasingly important in various aspects of research activities are data infrastructures for e-science and the management of research information, including metadata that describes research output and context (Hey and Trefethen, 2005). The present article defines a holistic approach for research information management with CERIF (Common European Research Information Format) as the metadata model central to the architecture. CERIF is currently being used in numerous systems in production across Europe (e.g. national or institutional research information systems), as well as in European FP7 e-infrastructure projects such as OpenAIREplus, EuroRIs-Net+ and ENGAGE.

We establish the research context (Section 2) and the need for metadata leading to a 3-layer model to accommodate the requirements (Section 3). We focus on the metadata model describing research information (Section 4). The CERIF modelling approach and the CERIF Model are covered in Section 5 while an architecture for research information based on CERIF is described (Section 6). The article concludes with a summary section.

2 A Model of the Research Environment

The research environment should provide a complete information management system for all the stakeholders. This ranges from an e-infrastructure of detectors, data processing, simulation and scholarly communication through to cooperative virtual working environments and mechanisms for publicising research. The key is research management information since it ties together all the other

ICT systems, their information, processes and resources, into one contextual whole. The researcher should be able to access research management information, complete research proposal documents, generate scholarly publications, collect data from detectors, perform statistical analysis, do simulations, produce output reports and visualisations all within a workflow and manage the research group within one environment. This is the thinking behind e-Science. This leads to an emerging general model for the research environment (Figure 1) where the components in the model depend upon metadata for their (inter-)operation. In this article, we focus on a model for the metadata.

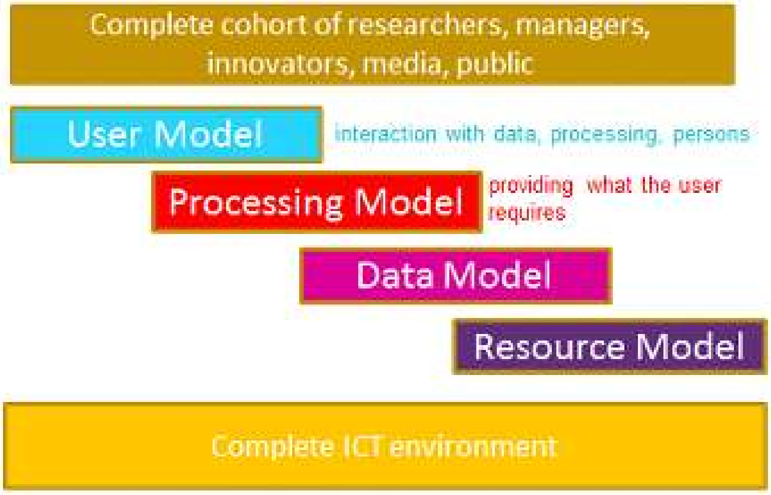

Complete cohort of researchers, managers, innovators, media, public
User Model
interaction with data, processing, persons
Processing Model
providing what the user requires
Data Model
Resource Model
Complete ICT environment

Figure 1. Emerging e-Infrastructure.

3 Metadata Model

3.1 The role of metadata

Metadata is the key to the e-infrastructure. Metadata is used to describe users, services (software processes), data and resources. Metadata is a parsimonious description of the system components, describing content, context and structure (Gill et al., 2008 and Haslhofer and Klas, 2010). We suggest metadata is used at three levels (a) for discovering system components of relevance (e.g. to find relevant datasets); (b) to understand the component in context (how does it relate to projects, persons, organisations, funding, facilities, equipment and what are the related outputs); (c) to point to more detailed metadata which is domain- or even component-specific and is used to connect services to datasets or resources taking into account both functional (what is to be done) and non-functional (under what conditions is it to be done) aspects.

It should be noted that (a) and (b) above are general and apply over all components thus allowing comparison, integration and a homogeneous view over the heterogeneous components; (c) is specific at domain or even component level.

Metadata is also of paramount importance in the e-infrastructure for open Public Sector Information (PSI) data. Increasingly, governments wish to make publicly-funded information available to the citizen and to innovators who may use it for wealth creation and improvement in the quality of life. A need for linking and interoperation between PSI and research information is emerging, since at present the data made available is dominantly from government departments (such as census data, information on environment, healthcare, education) but already there are demands to make available the research data (and services) which led to / generated this summary data for the purposes of deeper investigation. Furthermore, PSI data is potentially very useful for researchers for reuse in their activities, something particularly relevant in the Social Sciences, but applying to all disciplines. A further capability is that researchers can curate and process the open governmental data to make it higher quality and more suited to research purposes. It is very important for metadata to be able to capture these relationships with accurate semantics (e.g. Dataset A derived from Dataset B after curation by Researcher X in the frame of Project Y).

Unfortunately, current PSI open data sites are characterised by (a) basic descriptive metadata; (b) very limited contextual metadata; (c) usually no detailed metadata. In fact they usually provide no more than a simple metadata description of the dataset or file and a URI pointing to the file itself leaving the user to download the file and do what they can with it without further assistance.

Semantic web technology i.e. adding semantics (meaning) to datasets is improving the descriptive metadata and LOD (linked open data) technology (Berners-Lee, 2006) (Bizer et al. 2009) is allowing users to indicate that one dataset is related to another. Unfortunately, LOD does not easily allow the linking relationship to be characterised in any detail either thematically (what the link provides in terms of subject area) or in spatio-temporal relevance (where and when the link is relevant or valid). This commonly leads to publishing of research and/or PSI information as LOD, "out of context", which does not allow for example the recording of adequate provenance information that is critical for the data user to correctly interpret, trust the data, evaluate its reuse potential and eventually reuse it (Bechhofer et al., 2013). Furthermore, inconsistent information difficult to identify and repair is plaguing the LOD cloud (Jain et al., 2010).

However, there remains a need to integrate semantic web / LOD information since (a) some research information is only available in semantic web/LOD form using RDF and there is a need to interoperate with it in order for the end-user to gain a complete picture; (b) there are demands to present research information in LOD form for ease of navigation, linking and use. The 3-layer metadata model presented in paragraph 3.2 provides the structure for this integration.

3.2 A 3-layer metadata model

The task of representing, in a coherent way, research information, including datasets, in the research context is very challenging, given the heterogeneity of the research data produced in different scientific disciplines. The approach proposed for handling metadata is therefore based on a three-level scheme (depicted in Figure 2) with gradual increase in complexity and granularity. The figure shows the 3 layers and their function and also relates them to the PSI environment (semantic web, linked open data) mentioned above.

Varying degrees of detail and the need to address different requirements are reflected in these three discrete levels, as described in the following:

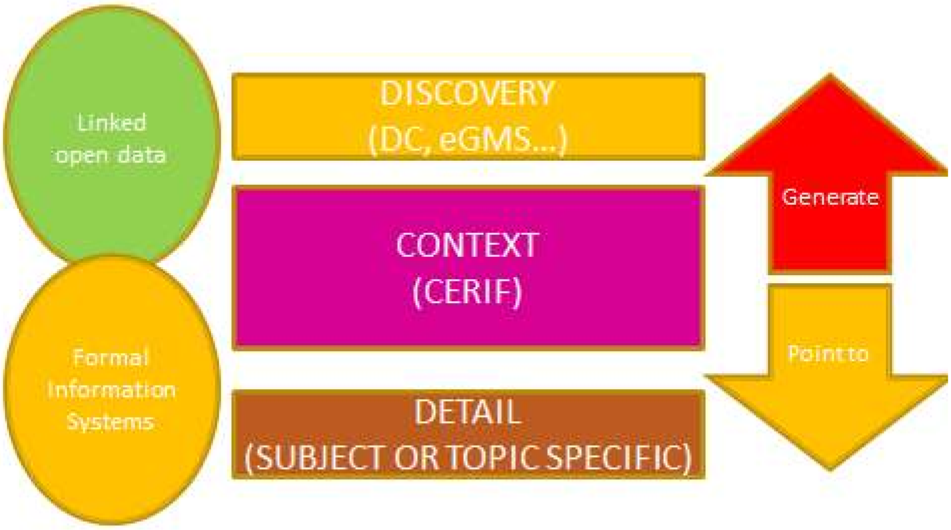


Figure 2: 3-Layer Metadata Model

Level-1. Discovery Metadata. Simple, 'flat' metadata schemata, analogous to legacy Dublin Core. This level of information is useful to assist non-sophisticated users to perform basic searches and find data sets using a very limited and easy to learn vocabulary ("metadata pidgin") (Baker, 2000). Examples of such schemata are legacy Dublin Core (plain and qualified) and data models used by software platforms like CKAN. In schemata of this type, the typical case is that only research output items are primary entities or objects. Therefore, there are records for publications and/or data sets but no separate complex metadata items for autonomous entities external to the publication/data set like persons, organisations, projects, facilities. The resulting flat records are relatively easy to populate and simple to understand and use for basic discovery services. However, they do not represent entities in the research environment in a balanced, fully structured way (e.g. publication records are structured, composite data elements containing distinct fields while persons and organisations are represented by plain strings, for example their name) and do not capture adequately semantic interrelations among entities. Only limited, rigid facilities (e.g. references to authority files) exist for linking to separate external entities. Essentially, they lead to catalogues with one "catalogue card" describing each publication or data set and constitute a common denominator for information about them, leading to a loss of semantic information when integrating data from heterogeneous sources. For example, if a number of persons are related with a publication with the role "author", the author names are included in a flat metadata record as values of a particular field (e.g. dcterms.creator) - providing no structured representation of data about persons (e.g. affiliations, contact details).

Level-2. Contextual Metadata. A structured, linked entity model for contextual metadata, able to treat any type of entity within the domain of Research Information (e.g. research outputs, persons, organisations, projects, funding programmes, facilities, etc.) as first-class citizens and capture the semantic relationships of entities with each other as well as entity classifications (i.e. roles). This enables the representation and reuse of semantically well-defined information about a data set's provenance, purpose, coverage, etc. Such a level of metadata allows functions and advanced services to be provided over data sets, like search and discovery, visualisation, navigation and browsing, mining, analytics and reporting to be available for more detailed analysis by end users. A formal metadata model is needed for this level, able to represent the concepts and relationships of interest to applications, so that integration does not lead to loss of information and semantic ambiguity.

Level-3. Domain Metadata. Detailed metadata standards for data sets of particular types or domains (e.g. CSMD for scientific data sets (Matthews et. al. 2009), SDMX for statistical data, INSPIRE for geospatial data, the Data Documentation Initiative (DDI) for social science data). These can be used for advanced domain-specific services and tools that can be provided for particular categories of data sets.

It should be noted that the contextual layer both generates congruent descriptive metadata to drive less rich interfaces but at the same time allows to point to more detailed metadata for composition of services and datasets. This means that the contextual metadata has to have a formal syntax (for efficient and effective processing) and declared semantics (for meaning understanding). Furthermore it should relate objects of interest together (linking) with semantics (meaning). Setting the stage for an international environment it must support multi-linguality and because of the needs of linking together datasets of different provenance it must support mapping of semantic terms across subject domains using vocabularies, dictionaries and ontologies.

4 Research information

Research information can be defined as any information that describes the research output as well as the context in which research is being conducted.
A diagrammatic illustration of the main entities in the research information domain is provided by Figure 4 (CERIF Entities). The main elements of research information can be described as follows:

a. Research output, including various kinds of text-based scientific publications, data sets, patents, software, devices, designs, artistic works and performances and a wide array of other types.
b. Information on the processes, workflows and methods utilised during the research process, e.g. observations, experiments, simulations with their associated measurements. (Matthews et al., 2009)
c. The people involved in research activities, including various categories of R&D personnel, i.e. researchers, research administrators/managers, technical and support staff participating in research projects (OECD, 2002, pp. 89-105).
d. Organisations involved in research activities , for example research performing organisations and research funders of various sectors, and their internal structure (e.g. schools, departments, institutes) (OECD, 2002, pp. 23 and 51-73).
e. Research projects, which refer to planned research activities aiming at the accomplishment of specific tasks under particular resource and time constraints. Projects typically, but not necessarily, rely on some sort of funding support.
f. The research funding environment that supports research, for example structured funding programmes with competitive allocation procedures executed by national and international public bodies or non-governmental organisations or direct state grants to research performing organisations, covering for example salaries of permanent personnel and basic operational costs.
g. Facilities and equipment that are utilised for research purposes. Facilities include research infrastructures which can be physical (e.g. buildings, synchrotrons, telescopes, vessels, supercomputers) or virtual (e.g. software systems), single-sited or distributed (Beckers, Jägerhorn, and Höllrigl, 2012).

h. Services related to research activities and/or provided through research infrastructuresor by organisations using facilities or equipment. Services can be targeted to other researchers, for example making facilities and equipment available for experiments, or to third-parties like industry (e.g. identification of materials through spectroscopic methods).
i. Events related to research activities, such as scientific conferences and workshops or periods of observation or experiment.
j. Measurements and Indicators concerning research activities, covering research outputs, outcomes and impacts and on the input side research funding. (Gartner, Cox, and Jeffery, 2013).

A significant part of research information is contained in the ***relationships*** among the aforementioned information entities, which should have clear ***semantics***, indicating the ***roles*** of each entity instance in a relationship. More details on the significance of relationships for research information and their representation in CERIF are provided in Section 5.2.
Research information is used increasingly for several purposes (CIBER, 2010)(Kroll and Forsman, 2010)(McColl and Jubb, 2011):

1. ***Researchers*** use the information to review relevant work, to discover competitors and potential collaborators and to gain ideas for future research. They also use research outputs such as datasets and software to cross-check the work of others, verify the corresponding results and possibly reuse them in their own research.
2. ***Research managers in research institutions*** use the information just as a commercial company would use business intelligence: to manage effectively resources, to plan future research, to monitor income and expenditure, to manage intellectual property and to perform benchmarking against competitors.
3. ***Research managers in funding organisations*** also use it in a similar way: to justify the funding expended in particular to evaluate the outputs of the research, outcomes from the research and – usually some time later – the impact of the research.
4. ***Policy and decision makers*** use the information to monitor research activities, identify strong and weak points, define strategic priorities and make informed decisions on funding allocation at the macro level.
5. ***Innovators*** use the information to pick up research prototypes, designs and ideas for commercial exploitation or improvement in the quality of life.
6. ***The media*** use the information to substantiate research 'stories' so that the citizen learns of the value of research.
7. Furthermore, making research outputs openly and freely available encourages participation by ***citizens*** in so-called 'citizen science' – well-known examples are found in astronomy and biomedical science.

Systems where various types of research information are maintained and interlinked are called Current Research Information Systems (CRIS).

5 CERIF: A data model for research information

5.1 Background

CERIF was developed in two phases by nationally-nominated experts convened by the European Commission. The first phase produced a data/metadata standard (CERIF91) that is not unlike typical discovery metadata of today (DC, CKAN, eGMS). However, as predicted by some of the experts and borne out by experience between 1991 and 1997 this was insufficient for automation of homogeneous access over heterogeneous resources. The second phase produced the CERIF2000 standard, formalised by Jeffery and Asserson and documented in the expert group report (Asserson, Jeffery, and Lopatenko, 2002). In 2002, a loose group of experts in research information which had been cooperating since 1991 formed euroCRIS: a not-for-profit organisation registered in the Netherlands and the EC mandated euroCRIS to maintain, develop and promote CERIF as an EU recommendation to Member States. euroCRIS now has more than 100 institutional members in approximately 40 countries and there are hundreds of implementations of CERIF, including by several commercial ICT suppliers.

5.2 The CERIF Modelling Approach

The entities in CERIF are selected and designed according to fundamental data modelling principles which also can be applied to ontologies thus providing within CERIF a very rich representation of both data (used as data or as metadata) and controlled vocabularies / ontologies to describe the semantics.

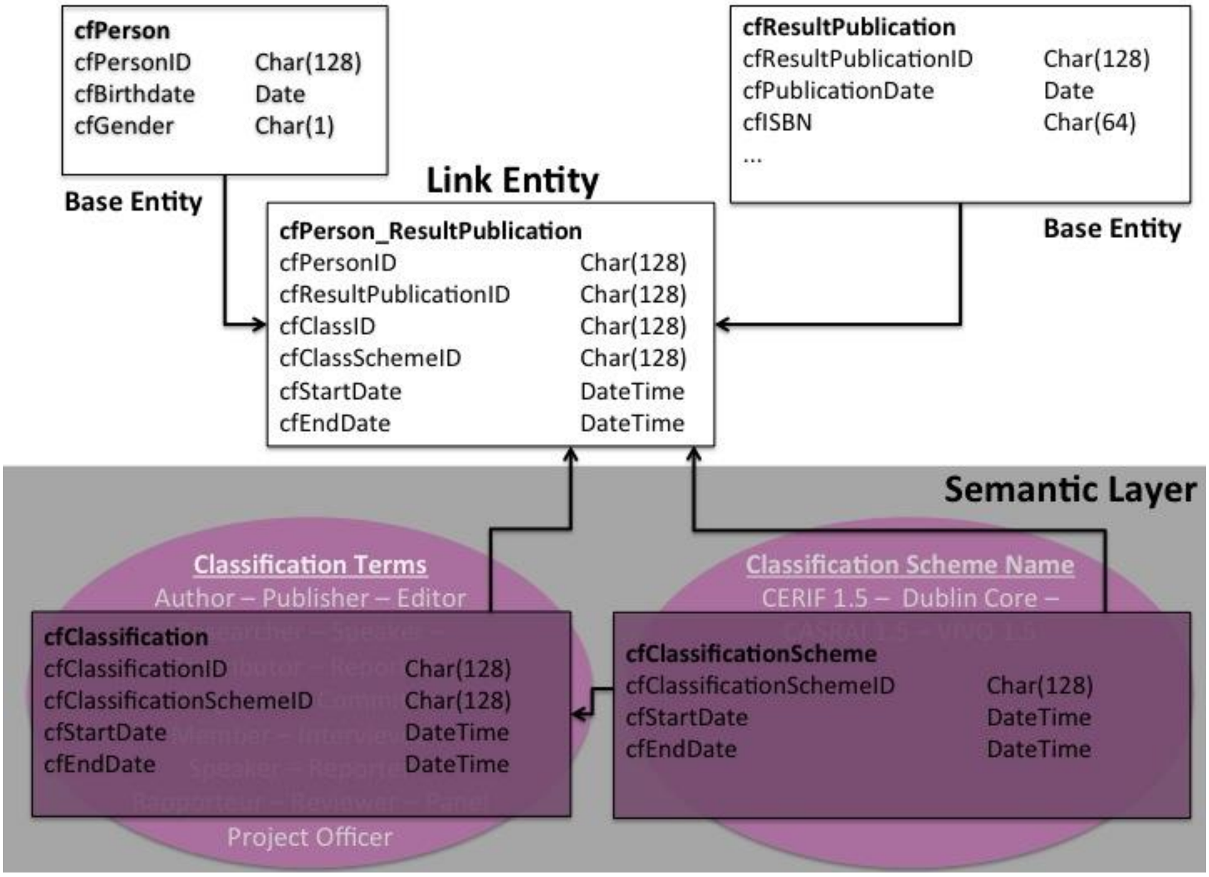
- Base (non-link) entities in CERIF represent real-world objects in the research domain such as persons, organisations, publications. They are distinct from link entities which describe relationships. In contradistinction to base entities, link entities represent asserted temporally-bound role-based relationships somewhat like but different from the potentially non-permanent, accidental relationships among entity individuals (Sowa, 1984). Instances of base entities linked in the relationship can enter and leave the relationship without losing their identity or existence in a similar way to (Guarino, 1992). This approach is aligned with the principle that relationships shall be represented as separate entities (although of different kind than base entities) and have their own attributes, instead of being hidden in attributes/fields of base entities. This follows the approaches of (Boella and Steimann, 2008 and Rumbaugh, 1987).

- One of the attributes of every link entity is ***role*** which has declared semantics specified within the semantic layer with the CERIF link entity attribute 'role'. This avoids the need for the proliferation of rigidly defined data fields in entity definitions. For instance, properties of a dataset like creator and maintainer can be modelled as relationships of product with persons and/or organizations, while the creation date can be captured as temporal information in the "creator" relationship. This approach enables representing state information as the status of an entity at a specific time point or transition events, instead of explicitly representing states of entities (using fixed sets of values, or boolean flags).

An innovative key design choice of CERIF was to separate base entities from link entities. In particular, the CERIF link entities allow for a generic classification mechanism to define their semantics while taking into account temporal aspects that enable time intervals during which a relationship is valid. This feature distinguishes CERIF from conventional relational database theory relating instances of entities through primary and foreign keys without inherent semantics, from native XML which equally indicates a linkage between instances of entities via IDs with limited semantics. Thus, instead of having a foreign key of authorID in a publication entity providing the linkage between a publication and its author, CERIF has a link entity bringing together the authorID, the publicationID, the role (in this case 'author' within a certain namespace (classification scheme)) and the date/time start and end. In fact, CERIF does not use author ID but personID since author in fact is a role, and the same person may have many roles with respect to publication (author, editor, reviewer, illustrator, contributor) and with respect to other entities such as project (manager, participant.)

A second major innovation in CERIF is that the values of 'role' in link entities are not stored directly as attribute values but as references to the so-called CERIF semantic layer. In this layer, one or more ontologies are modelled using the Classification and Classification scheme entities as base entities and having link entities indicating relationships between terms in the same scheme and in different schemes. igure 3 explains how the link entity 'role' attribute is replaced by Class (the term capturing the role semantics) and Class Scheme (the vocabulary the term belongs to). The combination of these two features enables flexible representation of relationships that is not easy to support in other data representation models. For example, in hypertext / hypermedia systems also linkage is separated from base data, however the links lack temporal scope information. Linkages between instances of entities can be also provided in Semantic Web systems (e.g. RDF/OWL), possibly with some semantics through its URIs. However, additional relationship attributes such as temporal information are theoretically possible but not straightforward to implement in practice; although some experimental approaches exist (Perry, Jain, and Sheth, 2011 and Gutierrez, Hurtado, and Vaisman 2007), these are not standardized and not supported by common triplestore implementations. Topics maps is another standard data representation model where relationships are first-class citizens, however temporal information on relationships is not included in the standard, although being investigated as a research topic (Teichmann and Maicher, 2009).

An important advantage of CERIF on the practical side of things, which has contributed greatly to its wide adoption, is the fact that besides the conceptual Entity-Relationship CERIF model a logical relational model is published and forms part of the CERIF standard. This enables the standard implementation of CERIF on relational databases which are the most common data management technology. Thus, the flexibility of CERIF, for example the ability to load and use multiple vocabularies and thesauri in a system for relationship semantics without altering the underlying storage schema (i.e. relational schema in relational database implementations) comes naturally with the use of standard, mature and robust relational database tools and straightforward support for non-functional requirements (e.g. performance/scalability, security) is available. While tools and platforms for developing applications with other technologies (e.g. Semantic Web / RDF, topic maps) are constantly improving, their maturity is far from reaching the level of their counterparts in relational databases.



igure 3. CERIF Base Entities (Person and Publication) connected via a Link Entity. The role / type of the link (e.g. Author, Publisher) is maintained in the Semantic Layer, where each classification term gets its own classification identifier cfClassificationID within a namespace or so-called classification scheme cfClassificationScheme that equally maintains its own identifier. The term and the namespace (context) identifiers as well as the Base Entity identifiers are re-used in the link table (e.g. Person-Publication) to build, along with time values, the composite key of the link entity.

However, the semantic layer can be used for base entities in addition to link entities. Within base entities a particular attribute may have a list of allowed values. Examples are the ISO standard country 2 and 3 character codes. In CERIF these are so-called Localisation entities. These entities are themselves linked with a Classification scheme (similar in concept to a namespace) and Classification (the actual terms belonging to the codes in a scheme). Similarly within a link entity, the role attribute may have a list of allowed values such as author, editor, illustrator etc. managed in the same way, however being bidirectional, i.e. allowing for either navigating direction. Grouping these vocabularies (expressed in terms and relationships between them i.e. terminology ontologies) together ensures semantic consistency within one implementation.

CERIF has a defined formal syntax (structure) – the physical names of the tables (e.g. cfPerson; cfPerson_ResultPublication); the names of their attributes (e.g. cfPersonIdentifier). Additionally it has declared semantics in the so-called semantic layer (cfClassification; cfClassification_Classification); cfClassificationScheme;cfClassificationScheme_ClassificationScheme). This means that within CERIF one can define what is meant by a specific term, for example 'assistant professor'. As indicated in Figure 3 above, an "Assistant Professor" would be a term with an underlying identifier specified in a namespace, i.e. following a particular classification scheme. Furthermore, one can relate 'assistant professor' in one dataset to, for instance, 'senior lecturer' in another dataset, via the recursion table cfClassification_Classification. Essentially, recursive relationships for classification schemes and classes enable the representation of any vocabulary structure (e.g. taxonomy, thesaurus) and the association and mapping among terms in different vocabularies. While a CERIF-based system is extensible to include any vocabularies, a set of common vocabularies is published as a separate component of the CERIF standard.

5.3 *Overview of the CERIF model*

The Common European Research Information Format (CERIF) is a model of the research domain, typically applied in Common Research Information Systems (CRIS). Technically, it is a core conceptual model - sometimes termed "property-centric ontology" (Doerr, 2003) or "enterprise model" (Calvanese, 1998) - that is able to represent the concepts of interest to research information applications and, importantly, their relationships in a semantically clear way (Calvanese, 2009 and Doerr, 2003). CERIF captures research results as well as entities involved in the research lifecycle and constituting the research context, like persons, organizations, projects, funding programmes, facilities, equipment, services, events, indicators and measurements and enables their linkage with geolocation information.

The main entities in CERIF, excluding link entities that represent relationships, are the following (Jörg et al., 2013) as depicted in Figure 4:

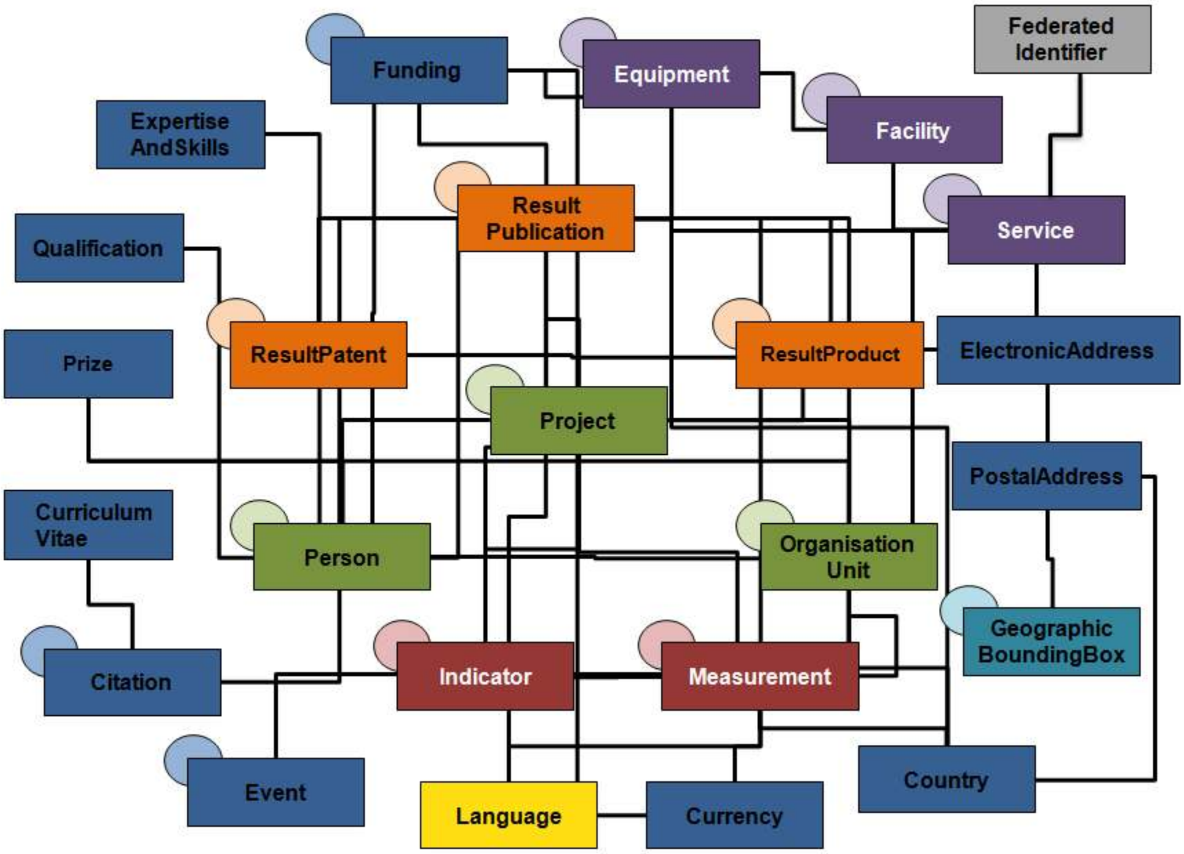


Figure 4. CERIF entities. Orange colour indicates entities representing research results, green fundamental entities of the research environment, purple research infrastructure entities, brown indicator and measurement entities, blue 2nd-level supporting entities.

- ***Base entities***: Project, Person, OrganisationUnit.

- ***Result entities***: Publication, Patent, Product. Product covers datasets, software, devices and other types of research output.
- ***Infrastructure entities***: Facility, Equipment, Service.
- ***Indicator and measurement entities***: Indicator and Measurement.
- ***2nd level entities***: Many entities of supporting role in research information, the most frequently used of them are Funding, Event, Medium. Funding refers to an amount of money or an inkind equivalent value allocated to a purpose (e.g. a funding programme). Medium refers to a means for storing information, essentially digital files.
- Geographic Bounding Box, enabling specification of geographic areas through the specific coordinates of their boundaries.

As elaborated in Section 5.2, the selection of entities in CERIF is in accordance with fundamental ontology design and data modeling principles and a key feature of CERIF (through the CERIF Semantic Layer) is the ability to represent semantic binary relationships among entities (e.g. person-publication, organization-project, project-funding programme). These can also be recursive links, that is relationships among two instances of the same type, for example two person instances (e.g. Person A is manager of Person B) or two organisation unit instances (e.g. Organisation Unit A is part of Organisation Unit B).

Further important characteristics of CERIF are the following:

- ***Entity instance identification***. Every entity instance in CERIF is identified by an internal system identifier – known as the primary key from traditional database technology. In addition, each base entity is associated with a URI and allows for federated identifiers (Jörg et al. 2012a). With a recent release (Jörg et al 2013), the CERIF model allows for linkage to external systems via a newly introduced federated identifier entity. This not only enables the linkage of CRIS with external systems, but provides a generic means to assign federated identifiers to system-internal records, i.e. identifiers assigned to the entity by external sources (e.g. different ids assigned to researchers by national authorities, commercial providers, etc.) This facilitates the generation of Linked Open Data from CERIF (Jörg et al, 2012b).
- ***Multilinguality***. All text fields in CERIF employ multiple language variants attributes and it can be declared whether the textual translation is human or machine. This holds for free-text descriptive metadata fields attached to core entities (e.g. title, name, abstract, keywords)

6 Research information interoperation

6.1 An architecture for research information interoperation

CERIF is a formal data model for implementing CRIS and thus the canonical way of providing interoperation (e.g. homogeneous query over heterogeneous CRIS). However, the interoperation is more subtle; since vocabularies are declared the interoperation can include transformations from one vocabulary to another via the canonical CERIF vocabularies. Thus, the end user works in their own semantic domain whatever the semantic domains of the other CRIS involved in the interoperation.

In particular, let us consider information integration using a canonical model, that could be termed property-centric ontology or global schema (Doerr, 2003 and Calvanese, 2009) and which is able to express all concepts of interest in the research information domain. Every data source is typically expressed in terms of the canonical conceptual model (in a way somewhat analogous to the Localas-View approach to mapping specification) so that any query formulated with the common ontology can be answered by all sources, with the replies being also represented in terms of the canonical schema. This way, information meaning is not lost for end users despite the heterogeneity of the sources, at the cost of a detailed mapping of the data source schema to the core model. CERIF is the common canonical model for research information that can be used for interoperation following the architecture of Figure 5.

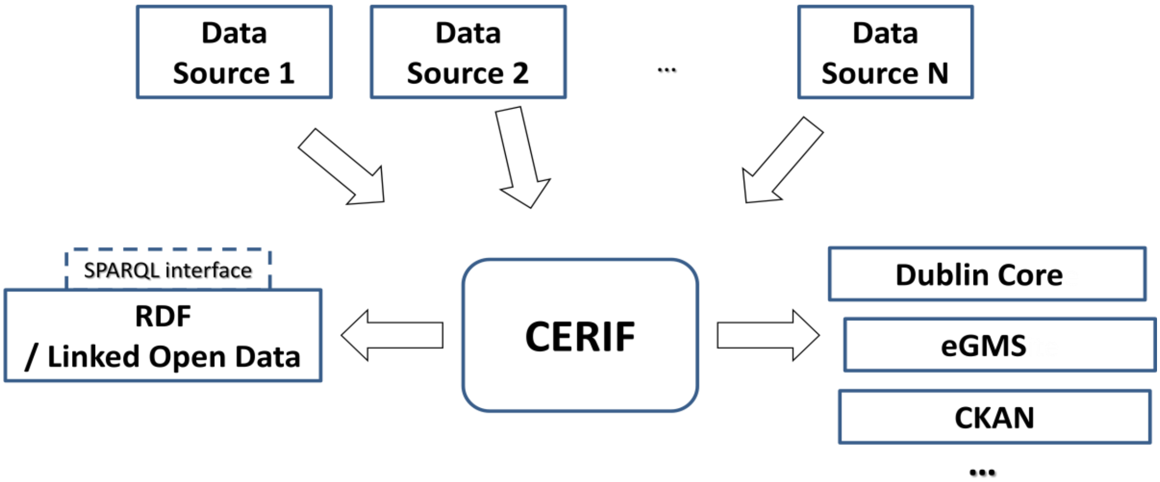


Figure 5. Overview of interoperation architecture

This is in contrast with a "lowest common denominator" approach to information integration, where information in sources is mapped to a central repository which follows a simplistic schema (Level-1 in the terminology of paragraph 3.2) that cannot adequately represent the semantics at the sources – as applied to digital repositories aggregators employing Dublin Core over OAI-PMH like OAISter; although useful for basic discovery services, this does not utilize the full potential of the available information.

Using this approach, level-1 standards (e.g. Dublin core, CKAN) can be fully supported as import and export formats of the systems. That is, any of these formats can be generated by CERIF as a model with greater expressiveness. Furthermore, data from sources using these formats can be included in the CERIF infrastructure, albeit not utilising the full potential of the global canonical schema since they might be ambiguous. A very interesting case is the provision of metadata from the sources in a Level-3 metadata schema, typically containing some contextual, domain-independent information that is mapped to CERIF.

6.2 Support of Linked Open Data

Linked open data is becoming a widely used approach to publish and access data and metadata on the web using semantic web technologies. LOD is based on some simple principles (Berners-Lee, 2006) (Bizer et al. 2009): assigning dereferenceable stable URIs to data sources, publishing metadata in RDF; and providing SPARQL endpoints for access. In order to support the use of LOD within the common infrastructure, we need to: use a stable URI scheme to identify entities within the CERIF model; and develop an exchange layer on the RDBMS implementation of CERIF to allow querying of the metadata using SPARQL and the delivery of the metadata in RDF.

CERIF is compatible with Linked Open Data, since the structure of the link entities and the semantic layer as well as the URI identifier with every research entity enables a straightforward publishing of data from a CERIF database according to the LOD principles. Therefore, the aforementioned architecture inherently supports the generation of research information as LOD. A well-known case where this has been achieved is the VOA3R project, where a CERIF back-end has been exposed as LOD (Jörg et al., 2012b). A standard way of providing CERIF metadata as LOD is being developed by a dedicated euroCRIS Task Group.

7 Summary

We propose CERIF as the key metadata format – at contextual level – for research information and research e-infrastructure. The key features are the separation of base and link entities providing integrity and flexibility and the use of a semantic layer so that terminology can be maintained with integrity and flexibility. The 3-layer model of metadata optimises the commonality among heterogeneous research datasets by describing them at contextual level with CERIF and generating congruent discovery metadata for both general discovery and interoperation with the linked open data / semantic web domain. The maintenance, development and promotion of CERIF are undertaken by the euroCRIS community, thus ensuring its continuous (but backward-compatible) evolution with ever increasing requirements. Important ongoing efforts are the finalisation of the standardisation of providing CERIF metadata as Linked Open Data and the wider application of CERIF as the contextual metadata layer in data-intensive research infrastructures of various disciplines through current initiatives in the frame of the Research Data Alliance.

References

Asserson, A., Jeffery, K., and Lopatenko, A. (2002). CERIF: past, present and future: an Overview. In *Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel* (pp. 33-40).

Baker, T. (2000). A Grammar of Dublin Core. *D-Lib Magazine* 6(10): 3.

Bechhofer, S. et al. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, *29*(2), 599-611.

Beckers, P., Jägerhorn, M., Höllrigl, T. (2012). Advances in Sharing and Managing Knowledge about European Research Infrastructures. *In: Proceedings of the 11th International Conference on Current Research Information Systems* (June 6-9, 2012, Prague, Czech Republic), pp. 129-138.

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved 10 April 2012, http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data – the story so far, *International Journal on Semantic Web and Information Systems.* Vol. 5, No. 3, pp. 1–22.

Boella , G. and Steimann, F, 2008. Roles and relationships in object-oriented programming, multiagent systems and ontologies: report on the 2nd workshop on roles and relationships at ECOOP 2007. *In Proceedings of the 2007 conference on Object-oriented technology, ECOOP'07*, pp. 108-122, Berlin, Heidelberg. Springer-Verlag.

Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., & Rosati, R. (1998, August). Information integration: Conceptual modeling and reasoning support. In *Cooperative Information Systems*, 1998. Proceedings. 3rd IFCIS International Conference on (pp. 280-289). IEEE.

Calvanese, D., G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati (2009). Conceptual modeling for data integration. In *Conceptual Modeling: Foundations and Applications*, LNCS, Springer.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Mag*. 24 (3), 75-92.

Haslhofer, B., and Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys (CSUR)*, *42*(2), 7.

Hornbostel, H. (2006) 'From CRIS to CRIS: integration and interoperability', *Proceedings of the 8th CRIS Conference*, Leuven University Press, pp.29–38.

Ivanovic, D. (2011) 'Data exchange between CRIS UNS, institutional repositories and library information systems', *Proceedings of the 5th International Quality Conference*.

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., and Sheth, A. P. (2010). Linked Data Is Merely More Data. In *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*.

Jeffery, K. G., Lay, J. O., Miquel, J. F., Zardan, S., Naldi, F., and Parenti, I. V. (1989). IDEAS: a system for international data exchange and access for science. *Information processing & management*, *25*(6), 703-711.

Jörg, b., Höllrigl, Th., Sicilia, M.-A. (2012a) Entities and Identities in Research Information Systems. *In: Proceedings of the 11th International Conference on Current Research Information Systems* (June 6-9, 2012, Prague, Czech Republic), pp. 185-194.

Jörg, B., et al. (2012b). 'Connecting Closed World Research Information Systems through the Linked Open Data Web'. *International Journal of Software Engineering and Knowledge Engineering* 22(03):345-364.

CIBER (Centre for Information Behaviour and Evaluation in Research) (2010), Research Support Services in UK Universities. (London: Research Information Network).

Jörg et. al. (2013). CERIF 1.5 Model Introduction and Specification. *to appear*

Gartner, R., Cox, M., & Jeffery, K. (2013). A CERIF-based schema for encoding research impact. *Electronic Library, The*, 31(4), 4-4.

Gill, T., Gilliland-Swetland, A. J., Whalen, M., and Woodley, M. S. (2008) Introduction to metadata. Getty Research Institute. ISBN: ISBN 978-0-89236-896-9.

Guarino, N, 1992. Concepts, attributes and arbitrary relations: some linguistic and ontological criteria for structuring knowledge bases. *Data & Knowledge Engineering*, 8(3), 249-261.

Guarino, N. and Welty, C. (2002) 'Identity and subsumption', *The Semantics of Relationships: An Interdisciplinary Perspective*, Kluwer Academic Publishers, pp.111–126.

Gutierrez, C., C. Hurtado, and A. Vaisman (2007). Introducing time into RDF. *IEEE Transactions on Knowledge and Data Engineering* 19(2), 207 - 218.

Hey, T., Tansley, S., and Tolle, K. (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft, Redmond, WA. ISBN: 978-0-9825442-0-4.

Hey, T. and Trefethen, A.E. (2005). Cyberinfrastructure for e-Science. *Science* 308 (5723):817-821.

Kroll, Susan, and Rick Forsman (2010), A Slice of Research Life: Information Support for Research in the United States. (Dublin, OH: OCLC Research)

MacColl, John and Michael Jubb (2011), Supporting Research: Environments, Administration and Libraries. (Dublin, Ohio: OCLC Research.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., Kleese, K. (2009). Using a Core Scientific Metadata Model in Large-Scale Facilities. *5th International Digital Curation Conference (IDCC 2009)*, London, UK.

Organisation for Economic Co-operation and Development. (2002). Frascati Manual 2002: Proposed Standard Practice for Surveys on Research and Experimental Development. OECD.

Perry, M., P. Jain, and A. P. Sheth (2011). SPARQL-ST: Extending SPARQL to support spatiotemporal queries geospatial semantics and the semantic web. Volume 12 of *Semantic Web and Beyond, Chapter 3*, pp. 61-86. Boston, MA: Springer US.

Rumbaugh, J. (1987). `Relations as semantic constructs in an object-oriented language'. *SIGPLAN Not.* 22(12):466-481.

Sowa, J.F. (1984) Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, New York; 1984.

Steimann, F. (2000) `On the representation of roles in object-oriented and conceptual modelling`. *Data & Knowledge Engineering* 35(1):83-106.

Teichmann, C. and Maicher, L. (2009) Temporal Qualification in Topic Maps. Maicher, L.; Garshol, L. M. (Eds.): *Linked Topic Maps. Fifth International Conference on Topic Maps Research and Applications, TMRA 2009* Leipzig, Germany, November 12–13, 2009. Revised Selected Papers. Leipziger Beiträge zur Informatik. ISBN 978-3-941608-06-1

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer* 25 (3), 38-49.