**CERIF – úvod ke společnému evropskému formátu pro informace o výzkumu**

Dvořák, Jan
2011

*Seminar to access of grey literature 2011: 4th year of the seminar focused on storage and making the grey literature accessible, 25. 10. 2011 [online]. Praha: National Technical Library, 2011. Available at WWW: http://nusl.techlib.cz/Sborniky. ISSN 1803-6015.*

# CERIF - COMMON EUROPEAN RESEARCH INFORMATION FORMAT: AN INTRODUCTION

JAN DVOŘÁK

jan.dvorak@infoscience.cz

InfoScience Praha, s.r.o.

**Abstract**
CERIF (Common European Research Information Format) is the model of the research domain. It includes the base entities of Publication, Project, Funding, Person, Organization Unit, and many others. Every entity instance can have any number of type or subject classifications. The facts are kept in temporally bound, role-based M:N relationships between entities in the model. Both roles and types are kept in the Semantic Layer that allows for formal syntax and declared semantics. Text attributes can have values in multiple languages.

CERIF has been recommended by the European Commission as the standard for building CRISs (Current Research Information Systems). The stewardship of the data model has been entrusted to euroCRIS, the not-for-profit non-government association of research information management professionals.

**Keywords**
CERIF, research information, Current Research Information System, euroCRIS

## 1. OVERVIEW

CERIF (the Common European Research Information Format) is a standard for representation and interchange of information about research: projects, people, organizations, publications, patents, products, funding, equipment, facilities, etc. and the relationships between them.

CERIF is:

1. a model on the conceptual level,

2. a logical entity-relationship model,

3. a set of database creation scripts (for selected database engines).

The European Commission recommends using CERIF to build Current Research Information Systems.[1]

[1] Current = of current interest, relevance. Not only the present state: in fact, CERIF seamlessly incorporates historical information.

## 2. SCOPE OF CERIF

CERIF includes the following five principal entities:

- ResultPublication - can represent an individual article, an issue of a journal, or the whole journal. Alternatively, it can represent a whole book or a book chapter.
- Project - can represent the whole project, a work package of a project, or any other activity that is relevant to research.
- Funding - a whole funding programme, a call, or an individual grant.
- Person - a human being.
- Organization Unit - the whole organization, its organization units (such as departments or research groups), also committees, networks, clusters, …

Along with their multilingual attribute entities, their classifications, and the links among them (see below) these entities are considered the Core of the CERIF model.

Other types of results can also be represented: patents and products. The latter can represent many different types of outputs from research: datasets, software, physical samples, production technologies, breeds, treatment procedures, etc.

Apart from the core and the result entities, additional entities (termed the "2nd level entities") are present in CERIF:

- Event - a congress, symposium, conference, workshop, seminar, get-together.
- Facility, Equipment, Service (research-oriented) - the infrastructure for research.
- PostalAddress, ElectronicAddress - contact or location information.
- CV, Expertise/Skill, Qualification, Prize - information about a Person.

## 3. STRUCTURE OF CERIF

All of the basic CERIF entities have the following two attributes:

1. An identifier (the primary key). While any mechanism for constructing identifiers can be used, the CERIF TG recommends using UUIDs, as they are guaranteed to be unique on the global scale.
2. A URI of a resource on the Internet that gives more detail about an object.

Apart from these, just a few attributes are present. Where meaningful, one finds an acronym or the dates of the start or end of existence.

CERIF puts a strong emphasis on supporting multiple languages. All free text attributes, such as titles, descriptions, abstracts, and keywords, are represented as dependent entities with an additional distinction by language and translation mode (original value, human-translated, machine-translated).

CERIF entities do not include many attributes. In fact, the main strength of CERIF is in representing information by semantically meaningful links and classifications. As an example, the publisher of a book is often represented by an attribute or a set of attributes in many models. That leads to bad normalization in the data and consequently, to data quality problems. In CERIF, the book publisher is listed as an Organization Unit entity instance (with all its attributes). This instance is linked to the book with the role of publisher (i.e., Publication was published by OrgUnit, OrgUnit was the publisher of Publication).

Such relationships are stored in the Linking Entities of CERIF. These entities represent M:N relationships with specified roles and temporal intervals of validity. This allows to express uniformly the present state of the problem domain, the past, and also the bits of the future that are foreseen.

E.g. the fact that person X is employed by organization Y since D1 is expressed as the following tuple in the cfPers_OrgUnit table: (cfPersId=id_X, cfOrgUnitId=id_Y, cfStartDate=D1, cfEndDate=+∞, cfClassSchemeId=id_CERIF_semantics_2008-1.2, cfClassId=id_Employee). Here +∞ means "until things change" and is conveniently represented by a date that is sufficiently far in the future (the constant depends on the database engine used). When the employment ends on D2, the tuple is updated: cfEndDate := D2.

Most queries should specify the time range in which they want facts to be considered. Example: Give me the list of employees from June 2011.

Some of the linking relationships relate instances of the same entity: they are termed recursive linking entities. Most notably, they represent a hierarchic structure using the isPartOf/hasPart role.

While linking entities represent binary relationships between CERIF objects, also unary classification statements about the type, subject or status of a CERIF object are supported. These classifications support the temporal interval of validity, too.

Certain additional attributes are supported in the linking and classification entities: one can specify a fraction to which a statement is true, or its probability.

All the roles, types or other classification terms are kept in a central place: the Semantic Layer - see Fig. 1. The central entity there is Classification. A classification has the term and the description in potentially multiple languages. A classification is a part of exactly one ClassificationScheme. Two classifications can have relationships between themselves, the same is true of two classification schemes.

The classification terms in the CERIF Semantic Layer can represent many structures: flat codelists, hierarchical codelists, thesauri, and even ontologies.

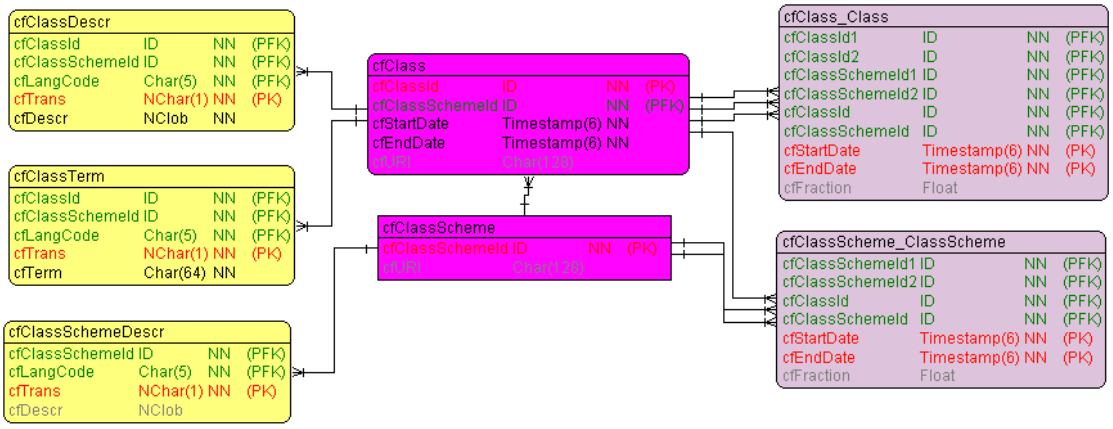


Fig. 1: CERIF Semantic Layer

## 4. TRENDS

### 4.1 INDICATORS AND MEASURES

A recent development in CERIF is the inclusion of a systematic framework for evaluation. Evaluation can be attached to any CERIF entity. Evaluation are usually texts: these are stored in the new Indicator entity. The evaluation statements are supported by evidence: figures that represent a quantity or a change (absolute or relative) of a quantity. These figures are stored in the links between the Indicator and the new Measure entity.

E.g.: A medicine paper (cfResultPublication) led to a new treatment procedure (cfResultProduct) for a disease. Applying the treatment procedure led to a rise in the successful treatment rate (cfMeasure) from 40% (cfIndicator_Measure 2005-2007) to 65% (cfIndicator_Measure 2008-2010). Based on this evidence, the paper is evaluated as having a profound impact on the treatment of the disease (cfIndicator).

### 4.2 LINKING TOWARDS THE EXTERNAL WORLD

As more and more data is brought on-line in the Linked Open Data (LOD) efforts, CERIF is required to relate to this data. While traditionally operating in closed-world databases, allowing to refer to objects that are not brought into this world (not even cached) is a profound change. All the existing integrity assumptions have to be weakened. The effect on data quality have not yet been assessed.

The CERIF TG is currently considering addition of a new "CERIF LOD" entity that would allow to link from an existing object within the CERIF database to an object in the "outside world". This can be a subject term in an external thesaurus, or a representation of another institution, or an external publication which a person in the CERIF database authored.

Another problem is that of publishing the contents of the CERIF database as a Linked Open Data. The highly structured and semantically rich CERIF structures are not straightforward to translate in the environment of the subject-predicate-object triples. Additional architectural decisions include the way of constructing identifiers, as well as many others.

### 4.3 GEOLOCATIONS

An increasing adoption of CERIF for tracking research infrastructure brings up the need to record locations of facilities and equipment that is cannot be represented by postal addresses. Examples include exploratory ships, submarine probes, and satelites. For this purpose, the addition of geolocation data is being considered.

## 5. EUROCRIS

euroCRIS is a not-for-profit non-government association of research information management professionals. It has been appointed the steward of CERIF by the European Commission. In the practice, the CERIF development is done by the CERIF Task Group.

## References

- Jörg, Brigitte; Jeffery, Keith; van Grootel, Geert; Asserson, Anne; Dvorak, Jan; Rasmussen, Henrik: CERIF 2008 - 1.2 Full Data Model (FDM): Introduction and Specification. [online] euroCRIS 2010. Available at WWW: http://www.eurocris.org/.
- Linked Open Data. [online] Available at WWW: http://www.linkedopendata.org/.
- euroCRIS website. Available at WWW: http://www.eurocris.org/.