An extension of the CERIF data model for description of scientific projects in the Republic of Serbia

Dejan Mijić[1], Dragan Ivanović[1], Zora Konjović[1], Dušan Surla[2]
[1]*University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia*
[2] *University of Novi Sad, Faculty of Sciences, Novi Sad, Serbia*

Abstract – *This paper presents an extension of the CERIF data model, necessary for storing information about scientific projects funded by state authorities in the Republic of Serbia. Based on the set of metadata identified in these projects, a CERIF compatible conceptual model is designed which will be deployed in the common research information system of the University Novi Sad – CRIS-UNS.*

1. INTRODUCTION

In the early eighties, leading research organizations launched project IDEAS. Goal of this project was to enable interoperability between various national systems for managing information about scientific research activity. Further development of IDEAS project, accompanied by evolution of information technologies, led to creation of the *Common European Research Information Format* (CERIF - www.eurocris.org/Index.php?page=CERIFintroduction&t=1). Today, CERIF is considered to be the most prominent standard for exchange and representation of scientific research data.

This paper presents an extension of the CERIF data model aimed at enabling storage of information about scientific projects in an arbitrary information system of scientific research activities based on CERIF data model. Presented extension will be applied in information system of scientific research activities at University of Novi Sad – CRIS-UNS (http://cris.uns.ac.rs/) which is under development since 2008. The system relies on the CERIF compatible data model based on MARC 21 format [1]. In the paper [2] it has been shown that MARC 21 format has capability of storing metadata declared by Dublin Core and EDT-MS standards. More details about CRIS-UNS implementation can be found in papers [3] and [4]. Automated extraction of metadata from published results in this system is described in paper [5]. Papers [6] and [7] describe extension of CERIF data model for the purpose of evaluation of published scientific results. This extension is based on CERIF's semantic layer, which, when applied properly enables classification of entities and their relationships according to different classification schemes. Process of integrating CRIS-UNS repository of dissertations with ETD-MS compatible repository is described in [8]. Data migration from previously existing repository named DIGLIB UNS is described in [9]. Dissertation repository search is available at: http://www.cris.uns.ac.rs/searchDissertations.jsf.

In addition to this introductory part, this paper contains four sections. Section 2 describes CERIF data model structure and organization. Special emphasis was put on the group of entities containing metadata about scientific projects. Analysis of typical scientific projects funded by Government of Republic of Serbia and Government of Autonomous Province of Vojvodina is topic of the Section 3. Goal of this analysis is to acquire necessary metadata about various projects in order to be able to represent them within information systems of scientific research activities. Result of this analysis is the CERIF model extension proposal given in the Section 4, while Section 5 contains conclusions and further work directions.

2. CERIF DATA MODEL

Common European Research Information Format (*CERIF*) is a conceptual data model that enables interoperability between various information systems of scientific research activity and management of information such as those about researchers, projects, publications, etc. First version of CERIF was published in 1991. Current version of CERIF is CERIF 2008 v1.3 [10, 11], and further analysis will be based on it.

The structure of the CERIF data model consists of 10 interrelated groups of entities listed in Table 2.1.

Entity group name	Entity group represents
Base	Projects, persons and organizational units
Results	Publications, patents and products
2nd level entities	Research context (e.g. country, language, event, funding)
Link	Links between CERIF entities
Infrastructural	Facilities, equipment and services
Multiple Language Features	Names, titles, descriptions, keywords, abstracts, and even semantics
Semantic	Relationship kinds, application views, subject classifications
Geographic	Geographic binding
Measurement	Quantitative measurements
Additional features	Such as Dublin Core Element set mapping and *PersonName* entity

Table 2.1 CERIF entities groups

This section describes structure of basic CERIF elements and their relationships, with special emphasis put on entities containing metadata about scientific projects.

2.1. Base entities

This group is consisted from three entities: *cfProj*, *cfPers* and *cfOrgUnit*. These entities represent projects, persons and organizational units involved in scientific activities. Relationships between these entities are shown in Figure 1.

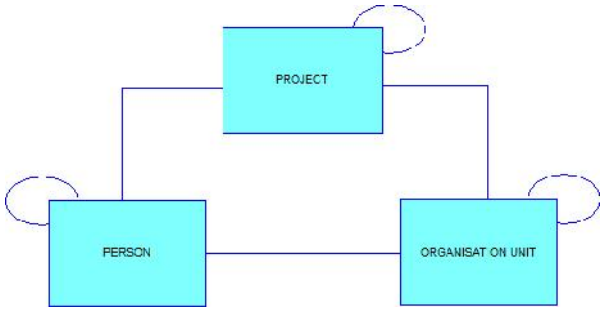

Figure 1. CERIF base entities

Various data about projects is mapped to a CERIF entity ***cfProj*** and group of entities related to it. Table 2.2 shows internal organization of the entity *cfProj*.

Attribute	Semantics
cfProjId	Primary key – ID
cfStartDate	Project start date
cfEndDate	Project end date
cfAcro	Project acronym
cfURI	Project URI

Table 2.2 Entity *cfProj*

Relationships with other entities (e.g. projects, persons, organization units, publications, products, etc.) are established using link entities with associated semantics. Beside aforementioned entities, three multilingual entities are also used to store data about scientific projects in CERIF: *cfProjTitle*, *cfProjAbstr* and *cfProjKeyw*, which are used to represent project title, abstract and keywords respectively.

Entity ***cfPers*** and a group of entities related to it are used to store data about persons involved in various scientific research activities. Table 2.3 shows internal organization of the entity *cfPers*.

Attribute	Semantics
cfPersId	Primary key – ID
cfBirthdate	Person birth date
cfGender	Person gender
cfURI	Person URI

Table 2.3 Entity *cfPers*

Relationships with other entities are again established using link entities with associated semantics. Multilingual entities *cfPersResInt* and *cfPersKeyw* are used to represent research interests and keywords related to a person. Entity *cfPersName*, together with link entity *cfPersName_Pers*, is used to store personal name of a participant in scientific activity.

Entity ***cfOrgUnit*** together with entities related to it is used to represent organization units involved in scientific activities. Table 2.4 shows internal organization of the entity *cfOrgUnit*.

Attribute	Semantics
cfOrgUnitId	Primary key – ID
cfAcro	Organization acronym
cfCurrCode	Organization currency
cfTurn	Organization turnover
cfHead	Organization headcount
cfURI	Organization URI

Table 2.4 Entity *cfOrgUnit*

As with previously mentioned entities, link entities with associated semantics are used for interconnection with other CERIF entities. Name of organization unit, research area and keywords associated to it are represented with multilingual entities *cfOrgUnitName*, *cfOrgUnitResAct* and *cfOrgUnitKeyw*.

2.2. Result entities

Result entities are used to hold data about publications, patents and products arising from a scientific research activity. These entities are *cfResPubl*, *cfResPat* and *cfResProd*. Relationships between them are shown in Figure 2.

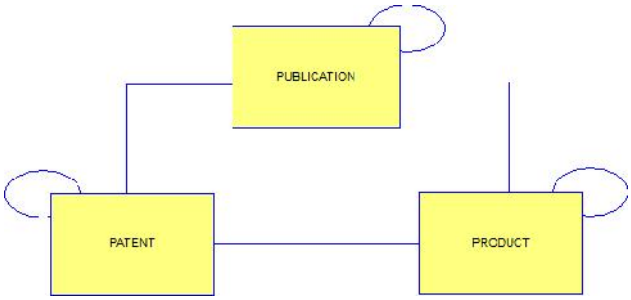

Figure 2. CERIF result entities

In addition to aforementioned entities, various multilingual and link entities with associated semantics are used to hold data about scientific research results.

2.3. Link entities

Link entities are all entities aimed at connecting other entities. Relationships established are always binary (between two entities), usually between base, result, infrastructure, measurement or 2nd level entities.

Name of the link indicates entities that are connected using it. For example, entity *cfProj_Pers* is used to

connect entity *cfProj* with entity *cfPers*. Table 2.5 illustrates a common structure of link entity.

Attribute	Semantics
cfEntityAId	Entity A primary key
cfEntityBId	Entity B primary key
cfClassId	Classifier ID
cfClassSchemeId	Class. scheme ID
cfStartDate	Link start date
cfEndDate	Link end date
cfFraction	Relationship percent

Table 2.5 Structure of link *cfEntityA_EntityB*

All attributes except *cfFraction* are mandatory. In addition to a given set, link entities can have as much additional attributes as necessary.

2.4. 2nd level entities

2nd level entities are all those entities that are used to hold various data about scientific research activities, but can't be classified as base or result entities. Main purpose of these entities is to store additional information such as physical or e-mail addresses, citations, biographies of researchers etc.

Each entity from this group has identifier and URI, with additional set of attributes, depending on entity purpose. The following entities belong to this group: *cfExpSkills*, *cfQual*, *cfPrize*, *cfCV*, *cfCite*, *cfEvent*, *cfMedium*, *cfFund*, *cfEAddr* and *cfPAddr*.

3. SCIENTIFIC PROJECTS

The focus of this section is analysis of metadata appearing in two types of scientific projects in the Republic of Serbia. First part of analysis targets projects funded by the Provincial Secretariat for Science and Technological Development of Autonomous Province of Vojvodina (referred as APV projects), while the second part discusses projects funded by the Serbian Ministry of Education, Science and Technological Development (referred as SMESTD projects). Results of the analysis are used to create an extension of the CERIF data model, in order to create CERIF-compatible conceptual data model.

During the analysis, the following convention has been introduced: attributes of existing CERIF entities that are equivalent to the extracted metadata, will be named between parentheses, together with entity they belong to.

3.1. APV projects

APV project structural analysis has been conducted according to the contract for the project "**Audio library for people with disabilities**", signed between Provincial Secretariat for Science and Technological Development and implementers: Faculty of Technical Sciences, Faculty of Sciences and Medical Faculty from Novi Sad.

Table 3.1 lists APV projects metadata together with their CERIF equivalents.

Metadata	CERIF equivalent
Project id	*cfProj (cfProjId)*
Project title	*cfProjTitle*
Start date	*cfProj (cfStartDate)*
End date	*cfProj (cfEndDate)*
Org. unit research area	*cfOrgUnitResAct*
Org. unit name	*cfOrgUnitName*
Org. unit address	*cfOrgUnit_PAddr*
Org. unit telephone	–
Org. unit e-mail	*cfOrgUnit_Eaddr*
Org. unit account	–
Person name	*cfPersName*
Researcher record id	–
Months per researcher	*cfPers_Proj*
Person title	*cfPers*
Person institution	*cfPers_OrgUnit*
Budget	*cfProjFund*

Table 3.1 Metadata about APV projects

3.2. SMESTD projects

Structural analysis of projects funded by Serbian Ministry of Education, Science and Technological Development has been conducted according to the contract for the project "**Infrastructure for Technology Enhanced Learning in Serbia**", signed between Republic of Serbia and implementers: Faculty of Electronic Engineering Nis, Faculty of Mathematics, Faculty of Mining and Geology, Faculty of Technology and Metallurgy, Faculty of Organizational Sciences and Faculty of Philology from Belgrade, Faculty of Sciences and Faculty of Technical Sciences from Novi Sad and Technical Faculty Cacak.

Table 4.1 lists metadata identified in the contract together with their CERIF equivalents.

Metadata	CERIF equivalent
Project id	*cfProj (cfProjId)*
Project title	*cfProjTitle*
Start date	*cfProj (cfStartDate)*
End date	*cfProj (cfEndDate)*
Org. unit name	*cfOrgUnitName*
Org. unit address	*cfOrgUnit_Paddr*
Taxpayer id. number	–
Identification number	–
Representative	*cfPers_OrgUnit*
Person's id. number	–
Person name	*cfPersName*

Metadata	CERIF equivalent
Person title	*cfPers*
Person institution	*cfPers_OrgUnit*
Months per researcher	*cfPers_Proj*
Equipment	*cfProj_Equip*
Budget	*cfProj_Fund*
Phase id	–
Phase title	–
Activity id	–
Activity title	–
Activity start date	–
Activity end date	–

Table 3.2 Metadata about SMESTD projects

4. CERIF MODEL EXTENSION PROPOSAL

Based on the analysis presented in the previous section, a set of necessary extensions to CERIF data model has been identified. The following list of conclusions has been made according to the set of missing metadata:

- New entities representing activity and telephone number should be added.
- New link entities connecting activity and telephone number entities with corresponding person as well as telephone number with organizational unit should also be added.
- Remaining metadata should be implemented by adding new attributes to existing CERIF entities.

List of all necessary extensions is given in Table 4.1.

Metadata	APV project	SMESTD project
Org. unit telephone	+	–
Org. unit account	+	–
Researcher record id	+	–
Taxpayer id. number	–	+
Org. unit registration code	–	+
Person id. number (JMBG)	–	+
Phase sequence number	–	+
Phase title	–	+
Activity sequence number	–	+
Activity title	–	+
Activity start date	–	+
Activity end date	–	+

Table 4.1 CERIF model extensions overview

Conceptual diagram of the extended CERIF compatible model is shown in Figure 3. For the sake of simplifying the presentation, the existing CERIF entities and relationships between them are omitted. Expansion in terms of new attributes is presented for the base entities – *cfProj*, *cfPers* and *cfOrgUnit*.

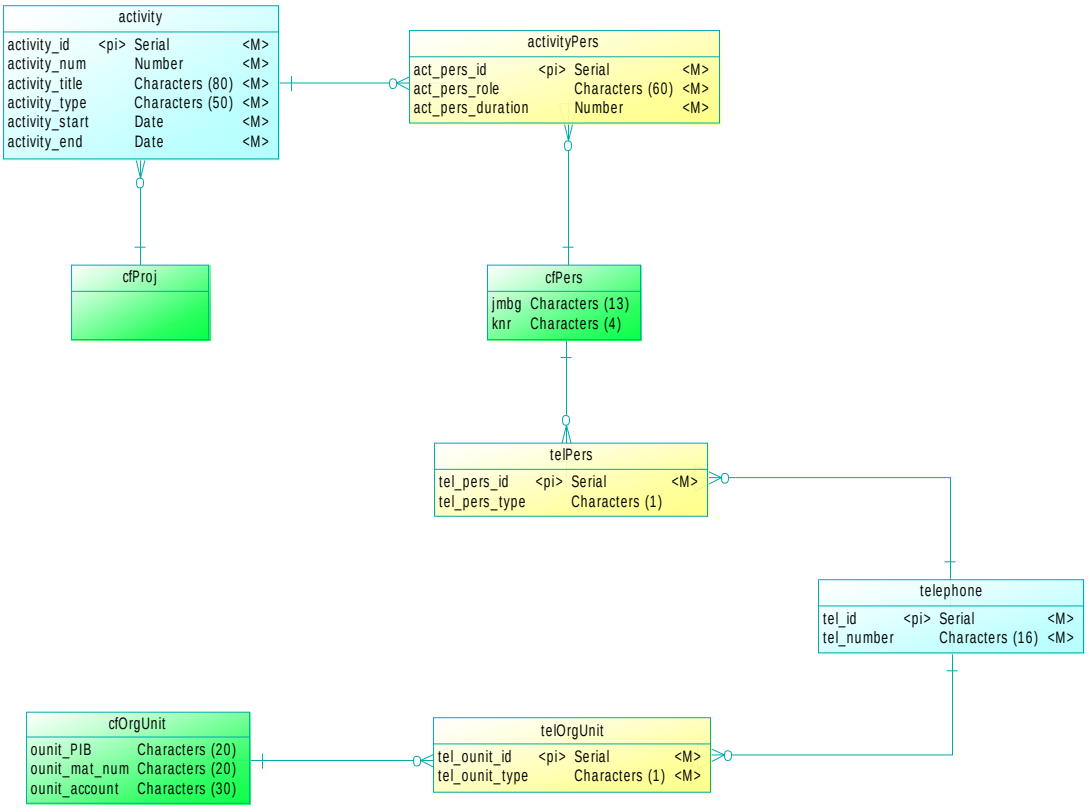

Figure 3. CERIF model extension

5. CONCLUSION

This paper describes the structure of the CERIF data model with the emphasis on groups of entities aimed for representation of scientific research projects and achieved results.

In order to enable data related to characteristic types of scientific projects funded by state authorities in the Republic of Serbia within the CRIS-UNS information system, two types of such projects have been analyzed – projects funded by the Government of Autonomous Province of Vojvodina, and projects funded by the Government of Republic of Serbia.

As a result of this analysis, the set of metadata which is missing in the CERIF data model is identified. Based on this set, the conceptual diagram of CERIF model extension has been constructed.

Further research includes development of software component for collection and management of information related to scientific projects, and integration of such component into the CRIS-UNS information system.

LITERATURE

[1] D. Ivanović, D. Surla and Z. Konjović, "*CERIF compatible data model based on MARC 21 format*", The Electronic Library, Vol. 29, No. 1, pp. 52 – 70, 2011. *DOI:* 10.1108/02640471111111433

[2] L. Ivanović, D. Ivanović and D. Surla, "*A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS*", Online Information Review, Vol. 36, No. 4, pp. 548 – 567, 2012.

[3] D. Ivanović, G. Milosavljević, B. Milosavljević and D. Surla, "*A CERIF-compatible research management system based on the MARC 21 format*", Program: Electronic library and information systems, Vol. 44, No. 1, pp. 229 – 251, 2012. DOI: 10.1108/00330331011064249

[4] G. Milosavljević, D. Ivanović, D. Surla and B. Milosavljević, "*Automated construction of the user interface for a CERIF-compliant research management system*", The Electronic Library, Vol. 29, No. 5, pp. 565 – 588, 2010. DOI: 10.1108/02640471111177035

[5] A. Kovačević, D. Ivanović, B. Milosavljević, Z. Konjović and D. Surla, "*Automatic extraction of metadata from scientific publications for CRIS systems*", Program: electronic library and information systems, Vol. 45, No. 4, pp.376 – 396, 2011. DOI: 10.1108/00330331111182094

[6] D. Ivanović, D. Surla and M. Racković, "*A CERIF data model extension for evaluation and quantitative expression of scientific research results*", Scientometrics, Vol. 86, No. 1, pp. 155 – 172, 2011. DOI: 10.1007/s11192-010-0228-2

[7] L. Ivanović, D. Ivanović and D. Surla, "*Integration of a Research Management System and an OAI-PMH Compatible ETDs Repository at the University of Novi Sad, Republic of Serbia*", Library Resources & Technical Services, Vol. 56, No. 2, pp. 104 – 112, 2012.

[8] D. Ivanović, D. Surla and M. Racković, "*Journal evaluation based on bibliometric indicators and the CERIF data model*", Computer Science and Information Systems, Vol. 9, No. 2, pp. 791 – 811, 2012. DOI: 10.2298/CSIS110801009I

[9] L. Ivanović and D. Surla, "*A software module for import of theses and dissertations to CRISs*", Proceedings of the CRIS 2012 Conference, Prague, June 6-9. 2012, pp. 313 – 322, 2012, available at address: *http://www.eurocris.org/Uploads/Web%20pages/CRIS%202012%20-%20Prague/CRIS2012_34_full_paper.pdf*

[10] D. Ivanović, "*Informacioni sistem naučno – istraživačke delatnosti*", doktorska disertacija, Fakultet Tehničkih nauka u Novom Sadu, 2010.

[11] B. Jorg, K. Jeffery, J. Dvořák, N. Houssos, A. Asserson, G. Grootel, R. Gartner, M. Cox, H. Rasmussen, T. Vestdam, L. Strijbosch, A. Clements, V. Brasse, D. Zendulkova, T. Höllrigl, L. Valkovic, A. Engfer, M. Jägerhorn, M. Mahey, N. Brennan, M. Sicilia, I. Ruiz-Rube, D. Baker, K. Evans, A. Price and M. Zielinski, "*CERIF 1.3 Full Data Model (FDM) Introduction and Specification*", available at address: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf (accessed as of: 12. 12. 2012)

ACKNOWLEDGMENT

The research presented in this paper is partially funded by Serbian Ministry of Education, Science and Technology Development, Grant No. 47003.