Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Publication metadata in CERIF: Inspiration by FRBR

Jan Dvořák[a]*, Barbora Drobíková[a], Andrea Bollini[b]

[a]*Institute of Information Studies and Librarianship, Faculty of Arts, Charles University in Prague, U Kříže 8, Praha, CZ-15800, Czech Republic*
[b]*CINECA, Via dei Tizi 6/B, IT-00185 Roma, Italy*

Abstract

The Functional Requirements for Bibliographic Records (FRBR) and its Scholarly Works Application Profile (SWAP) are used to inspire the representation of complex real world situations in the publication part of the Common European Research Information Format (CERIF), the model for Current Research Information Systems (CRIS). CERIF is found to have room for different approaches to representing metadata of scholarly publications, which could hamper the interoperability of CRIS. To lessen that risk, we propose guidelines for representing scholarly publication metadata in CERIF; our design goal is to enhance the utility of CRIS in supporting the functions of scientific communication. The guidelines are formulated using the notions of Scholarly Work, Expression and Manifestation from FRBR/SWAP.

Peer-review under responsibility of euroCRIS.

Keywords: CRIS; interoperability; CERIF; functions of scientific communication; publication metadata; FRBR; SWAP; scholarly work; expression; manifestation; research evaluation

1. Introduction

The Functional Requirements for Bibliographic Records study[1] by the International Federation of Library Associations and Institutions (IFLA) provides a conceptual model of the bibliographic universe: the set of all recorded knowledge. The Scholarly Works Application Profile (SWAP)[2,3] covers scholarly works, an important subset of the bibliographic universe. These are state-of-the-art models from library science.

The Common European Research Information Format (CERIF)[4] by euroCRIS is the primary model for Current Research Information Systems (CRIS); it covers research information – including research outputs, their creators and contributors.

* Corresponding author. *E-mail address:* jan.dvorak@ff.cuni.cz

Peer-review under responsibility of euroCRIS.

We find that CERIF – with its formal syntax and declared semantics – offers a great deal of flexibility in representing scholarly publication metadata. Although that is generally regarded as an advantage that allows all possible cases to be represented, it may also lead to different approaches to representing one single publication. This would create misunderstandings that could effectively hamper the CERIF-based interoperability of CRIS. We address this risk by introducing a convention: we propose guidelines for representing scholarly publication metadata in CERIF. Our design goal was to enhance the utility of CRIS in supporting the functions of scientific communication. The guidelines are formulated using the notions from FRBR/SWAP.

2. FRBR and SWAP

The FRBR model is designed as a completely general model independent on cataloguing rules or other indexing rules in different databases. For specific implementations the model shall be adapted to local needs. Entities, attributes and relationships are defined in the model.

Entities in FRBR are grouped in three groups: (i) group one comprises the work, expression, manifestation, and item entities (see below) that will be central to our discussion; (ii) group two contains agents (persons and corporate bodies) that are responsible for objects of entities in group one; (iii) group three contains entities that represent subjects. In the article we limit ourselves to the first and second group of entities – especially to work, expression and manifestation and person and corporate body. Definitions in this section are extracts from[1].

> ***Work*** *is a distinct intellectual or artistic creation. A work is an abstract entity; there is no single material object one can point to as the work. We recognize the work through individual realizations or expressions of the work, but the work itself exists only in the commonality of content between and among the various expressions of the work.*

SWAP[3] defines **Scholarly Work** as *a distinct intellectual or artistic scholarly creation*.

Variant texts with revisions, translations, abridgements or enlargements of a text are all considered the same work. They are represented as distinct expressions of a work:

> *An* ***Expression*** *is the specific intellectual or artistic form that a work takes each time it is "realized." Expression encompasses, for example, the specific words, sentences, paragraphs, etc. that result from the realization of a work in the form of a text (not only). The boundaries of the entity expression are defined, however, so as to exclude aspects of physical form, such as typeface and page layout, that are not integral to the intellectual or artistic realization of the work as such.*

Physical properties are represented as manifestations:

> ***Manifestation*** *is the physical embodiment of an expression of a work. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form.*
>
> *If two manifestations embody the same or almost the same intellectual or artistic content, even though the physical embodiment may differ and differing attributes of the manifestations may obscure the fact that the content is similar in both, we can make the common link through the entity defined as expression.*

We skip the last entity in the first group, Item (Copy in SWAP). For the discussion in this article it is marginal.

We illustrate the notions in an example.

Example 1. In Figure 1 we represent the scientific article

> Gover AR, Somberg P, Soucek V. Yang-Mills Detour Complexes and Conformal Geometry. In: *Communications in Mathematical Physics* March 2008, **278**(2):307-327. DOI: 10.1007/s00220-007-0401-5. ISSN Print 0010-3616. ISSN Online 1432-0916.

That is one scholarly work. The authors are responsible for the work.

This work has two expressions (versions): a preliminary one (before the review by journal's reviewers) and a final one (after the authors implemented comments from the review). Reviewers have contributed to the latter expression. Note that as a rule in most scientific disciplines, the reviewers remain anonymous.

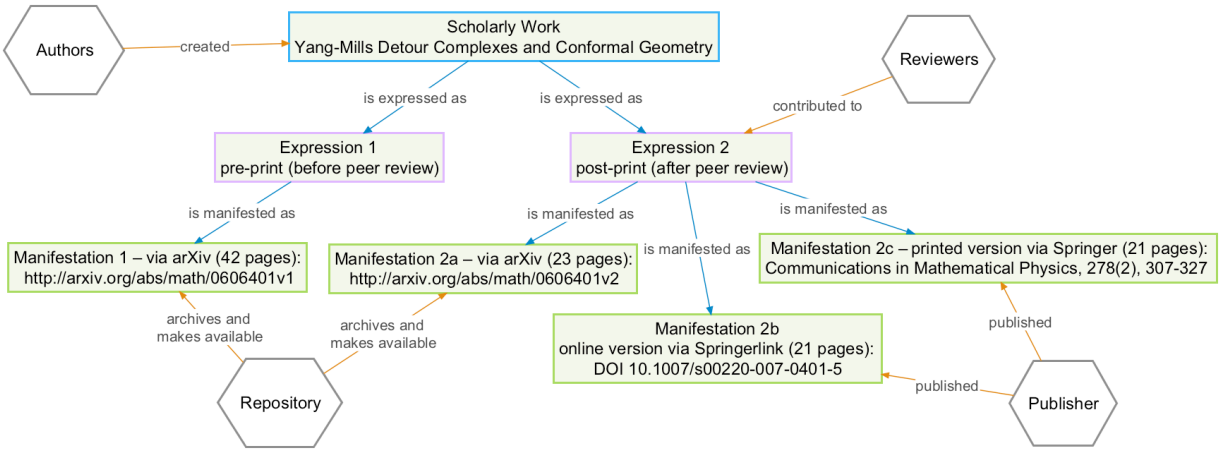

Figure 1. Example 1 represented in FRBR/SWAP.

The first expression (the version before the review process) is manifested as a deposit at arXiv.org, the *e-print service in the fields of physics, mathematics, computer science, […] and statistics*. The second expression (the final version) is manifested as: (i) another deposit at arXiv.org, and (ii) the "official" publication by the publisher (Springer). The latter takes the form of the printed article (with the navigational metadata consisting of: the journal title and ISSN, the volume, the issue, and the page range) and an online version (with the Digital Object Identifier (DOI) giving the location while sharing the printed version attributes). All this makes the total of four manifestations. The additional responsibilities of the publisher and of the repository are recorded at this level.

3. CERIF

Three entities represent outputs of research in CERIF: cfResultPublication, cfResultPatent, and cfResultProduct. Our discussion being concerned with publications, we focus on cfResultPublication. It has the following definition:

> *Collection of information records that, in combination, represent a full and up-to-date history of research or scholarly published outputs resulting from, or related to, the person's research activities.* [*Definition Source: http://dictionary.casrai.org/research-personnel-profile/1.1.0/contributions/outputs/publications*]

CERIF commonly uses additional classifications to express the exact type or other properties of base entity instances. These classifications are stored in the CERIF semantic layer. Classifications are organized in classification schemes. Different classification schemes represent different aspects of the base objects. The following classification schemes from the CERIF standard vocabulary[5] apply to the cfResultPublication entity:

- Output Types – with classifications such as Journal Article, Conference Proceedings, Book, Chapter in Book, etc., but also ones representing types of sources (Journal, Conference Proceedings, Anthology, Encyclopedia, etc.).
- Publication Statuses – with the following classifications: In Preparation, Submitted for Consideration, In Press, Published, Unpublished.

The cfResultPublication entity also has attributes (including multilingual ones) and any number of identifiers (possibly of different types) to express the information about a publication.

Every publication is connected to instances of other entities that represent related objects in the research information domain. For instance, authors of the publication are linked (using the cfResultPublication_Person linking entity with the role Author) to the cfPerson instances that represent the researchers who authored the publication. This way, CERIF records the context of each object.[6]

There is a common practice in CERIF that the source containing an output is also represented as a cfResultPublication instance. The output is linked with the source using a cfResultPublication_ResultPublication link with the Part role.

These are the building blocks to use to represent publications, such as the one in our Example 1, in CERIF. It is readily seen that there is room for several possible approaches.

CERIF's flexibility would certainly allow for a verbatim representation of the FRBR model. One would introduce a FRBR vocabulary into the CERIF semantic layer, with the Scholarly Work, Expression, Manifestation, and Copy terms being possible unary classifications for instances of the cfResultPublication entity. One would also introduce the isExpessedAs, isManifestedAs, isAvailableAs roles that would be used on relationships between Scholarly Works and Expressions, Expressions and Manifestations, and Manifestations and Copies, respectively. A basis of such vocabulary is found in[7].

However, such a structure seems to be rather complex, without much benefit for CRIS applications.

In order to make satisfactory design choices, we have to formulate our design goals. The FRBR and SWAP terms will be helpful in structuring the discussion.

4. Publication metadata in CRIS: Why

Scholarly publications are the vehicles for communication between researchers and from the researchers to all other stakeholders. Scholarly publication metadata therefore is an integral part of the research information domain.[6] Let us have a look at the larger picture of scholarly communication.

Functions in scientific communication are set forth by Rosendaal and Geurts[8]. With reference to C. G. Jung they define the following functions (as summarized in [9]):

1. *Registration*, which allows claims of precedence for a scholarly finding.
2. *Certification*, which establishes the validity of a registered scholarly claim.
3. *Awareness*, which allows actors in the scholarly system to remain aware of new claims and findings.
4. *Archiving*, which preserves the scholarly record over time.

Herbert van de Sompel et al.[9] add yet another function:

5. *Rewarding*, which rewards actors for their performance in the communication system based on metrics derived from that system.

CRIS support all of these functions. For that, the CRIS needs to hold enough information to unambiguously reference the scholarly publication. CRIS often record richer information about the context of a publication than what can be recovered from the publication itself (and indexed by a bibliographic database).

The archiving function is often delegated to Open Access repositories or library information systems. In those cases, the CRIS records locators of the relevant resources in the other systems (the navigational metadata).

The rewarding function often boils down to the need to demonstrate that a researcher, a research team, a research performing organisation or its organisational unit performs (or a funding agency funds) relevant research. All the actors seek to win and defend their space in the global ecosystem of research. They all are under a continuous pressure to demonstrate they are relevant, because they perform or support relevant research.

CRIS have traditionally been the sources of information for such demonstration. A structured way of the research relevance demonstration is participation – willingly or not – in a research evaluation exercise. For CRIS to effectively support research evaluations the scope of what counts as a single output has to stand on a solid foundation. For instance, while the distinction between the print and the online versions of a journal is important for keeping the precise bibliographic records and for managing library collections, it is generally considered irrelevant to research evaluation. However, the information has to be structured enough to avoid double counting.

These functions determine the focus of tracking publications in a CRIS. Making use of the SWAP terms, the focus is on the scholarly work: of the five scholarly communication functions, four – registration, certification, awareness, rewarding – are centered around the scholarly work. Where access to the full text of the publication is required – for archiving, awareness, and certification – concrete manifestations become important. The need for explicit representation of expressions (separate from their manifestations) is less pronounced in CRIS.

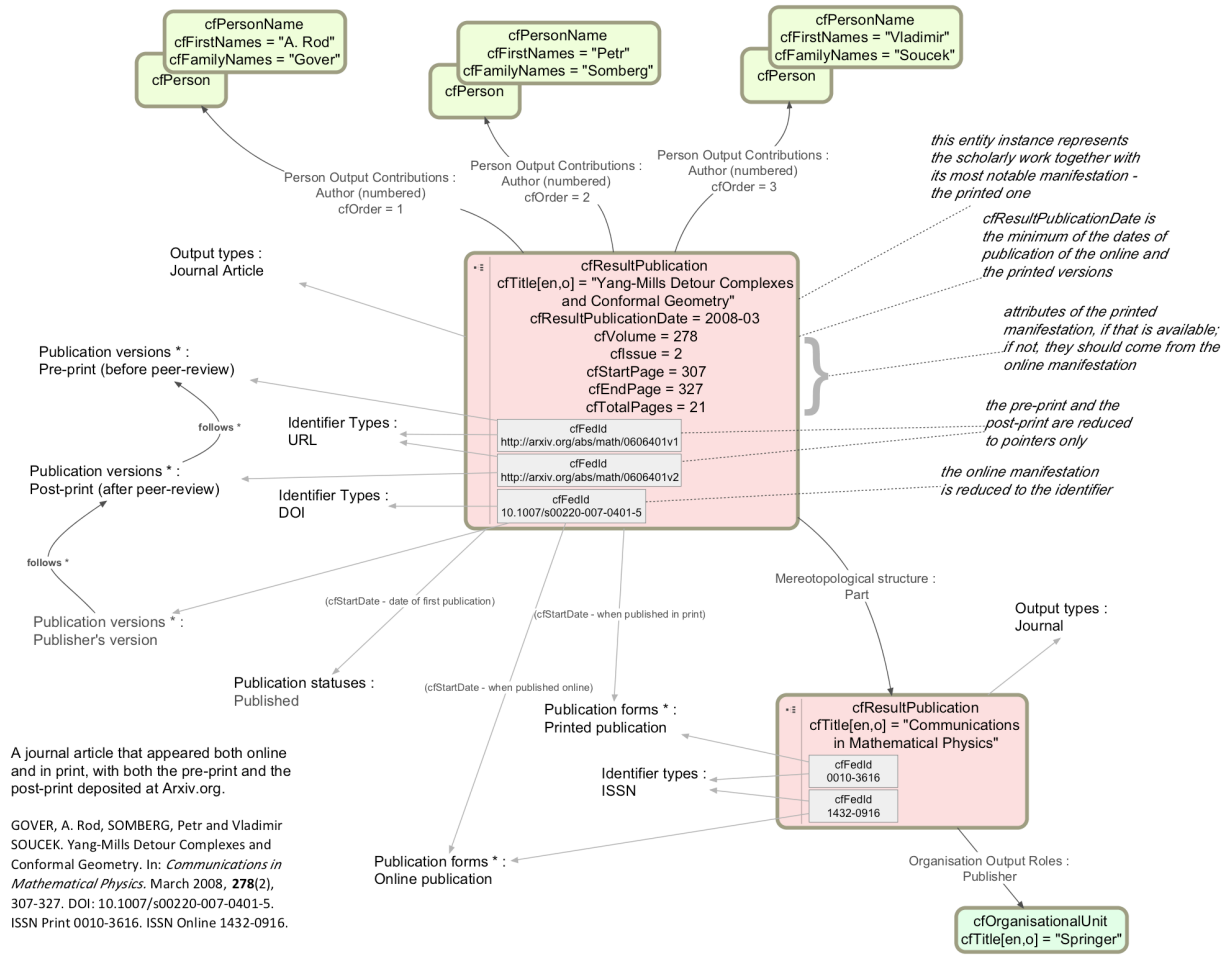

Figure 2. Example 1 represented using the proposed CERIF constructs.

5. Publication metadata in CRIS: How

Typically, we want to use a single cfResultPublication to express the publication allowing a simple and effective use of the information by the CRIS system. To get that effect, we should not introduce cfResultPublication instances beyond the absolutely necessary ones. We thus propose to reduce the online manifestations into just cfFederatedIdentfiers, and merge the scholarly work with the printed manifestation (if that is available) in the cfResultPublication instance. An example of this approach is in Figure 2.

In cases where the same work has several printed manifestations, such as in Figure 3, we allow for several cfResultPublications to represent the same work. These cfResultPublication instances should then be linked using a specific relationship with the role “Same Contents”.

Only when the need arises to collect additional information (such as separate metrics, e.g. citation counts) for specific manifestation, the CRIS should move to a more expanded representation (which would use FRBR).

The proposed representation also meshes well with research evaluation exercises. The present exercises usually make different branches based on the type of the output (i.e., of the specific manifestation). Our proposed representation would allow the research evaluation exercises to consider the scholarly work as such, or perhaps select one of its manifestations. This would open the road to a more precise evaluation of research activities.

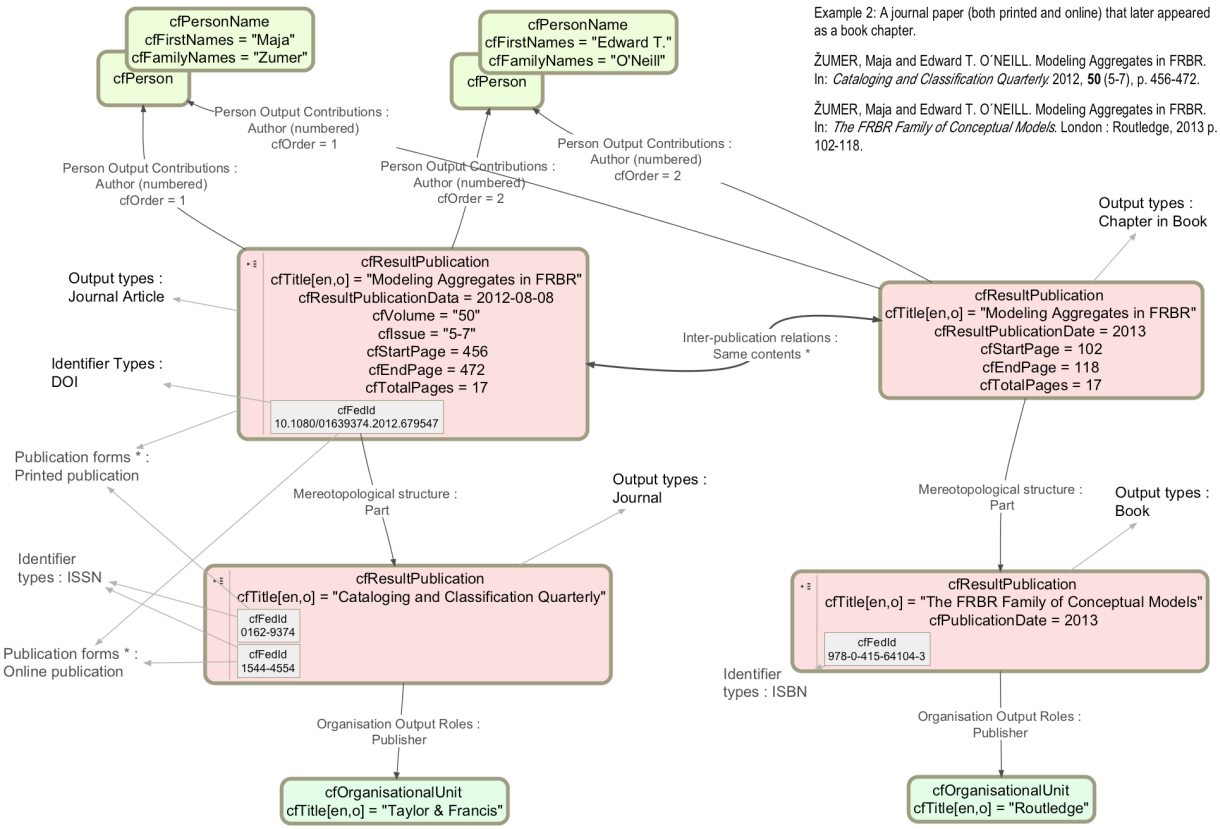

Figure 3. Example 2 represented using the proposed CERIF constructs.

Following the same reasoning process we are convinced that tracking journal issues in a CRIS is generally not useful; the exception is to record that a researcher was a guest editor of a special issue, as a kind of merit. In any case, all journal articles should contain the values of the attributes that specify the journal issue (de-normalized).

5.1. Guidelines for representing publication metadata in CERIF

We propose the following guidelines for representing publication metadata in CERIF:

1. Let cfResultPublication represent the scholarly work together with its most prominent ("official") manifestation. The manifestation forms are given the following priorities (highest first): Printed publication. Online publication. Publication on a memory medium.
2. Other manifestations of the same work that can be reduced to a simple pointer (resource identifier or locator) are represented as cfFederatedIdentifiers of the cfResultPublication. These identifiers have sufficient classifications to convey their semantics.
3. Manifestations that cannot be reduced to a simple pointer are represented as separate cfResultPublications. The cfResultPublications that embody expressions of the same scholarly work are interlinked using the "Same Contents" role.
4. Different expressions are not represented as separate objects. The types of expressions are merged in the classification terms that are used to classify the individual manifestations.
5. In the cases where it is applicable to CRIS, full texts are stored in the cfMedium entity that is linked to the relevant cfResultPublication.

5.2. Proposed vocabularies and/or terms

Following the proposed guidelines for representing publication metadata in CERIF, we also propose the following new vocabularies:

- Publication versions (to classify different versions (FRBR/SWAP expression) of the scholarly work) with the following terms: Pre-print (before peer review), Post-print (after peer review), Publisher's version. The terms can be linked (within the CERIF semantic layer) using a specific relationship to represent the sequence of versions in the scholarly publishing process.
- Publication forms (to classify different FRBR/SWAP manifestations) with the following terms: Printed publication, Online publication, Publication distributed on memory media.

We also propose to extend existing classification schemes with the following terms:

- Inter-publication relations: Same contents.

6. Conclusions

Given the functions of scientific communication and the purpose of CRIS, we conclude the FRBR/SWAP model is very detailed. This level of detail brings a lot of added complexity, which mainly contributes to areas of the usage of bibliographic information that are not in the scope of CRIS.

This motivated our simplified approach. We propose guidelines for representing publication metadata in CERIF together with the necessary vocabulary extensions.

Nonetheless, the clear definitions of the core FRBR/SWAP entities, specifically the scholary work / expression / manifestation / copy hierarchy, are very useful. We propose to use them indirectly in the CERIF vocabulary term descriptions to make them more clear and specific when referring to these base concepts.

Note

This article documents an exploratory effort. Its findings and proposed recommendations will be submitted to the euroCRIS CERIF Task Group for consideration.

Acknowledgements

Jan Dvořák's work on this article was in part supported by the Ministry of Education, Youth and Sports of the Czech Republic through grant no. LG14007, and in part by the Faculty of Arts of the Charles University in Prague through the *Bibliometry and Scientometry* internal development project.

References

1. IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records : Final Report*. IFLA, 1997, as amended and corrected in February 2009 [cit. 2014-04-18]. Available from: http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf
2. Allinson J. Describing Scholarly Works with Dublin Core: A Functional Approach. *Library Trends* 2008;**57**(2):221-243 [cit. 2014-04-18]. ISSN 0024-2594. Available from ProQuest Central.
3. Jörg B, Jeffery K, Dvořák J, Houssos N, Asserson A, van Grootel G et al. (editors). *CERIF 1.3 Full Data Model (FDM) : Introduction and Specification*. EuroCRIS, 2012 [cit. 2014-04-22]. Available from: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf
4. Žumer M, Zeng ML, Salaba A. FRBR: A Generalized Approach to Dublin Core Application Profiles. In: *International Conference on Dublin Core and Metadata Applications*. Pittsburgh: Dublin Core Metadata Initiative; 2010 [cit. 2014-04-22], p. 21-30. ISSN 1939-1366. Available from: http://dcpapers.dublincore.org/pubs/article/view/1024
5. CERIF 1.5 Vocabulary. EuroCRIS, 2012 [cit. 2014-04-22]. Available from: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF1.5_Semantics.xls
6. Jeffery K, Houssos N, Jörg B, Asserson A. Research Information management: the CERIF approach. In: *International Journal of Metadata, Semantics and Ontologies* 2014; **9**(1):5-14.
7. Davis I, Newman A. Expression of Core FRBR Concepts in RDF. 2005 [cit. 2014-04-22]. Available from http://vocab.org/frbr/core.html

8. Roosendaal HE, Geurts PATM. Forces and functions in scientific communication: an analysis of their interplay. In: *Cooperative Research Information Systems in Physics*, August 31—September 4 1997, Oldenburg, Germany [cit. 2014-04-22]. Available from: http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html
9. Van de Sompel H, Payette S, Erickson, J, Lagoze C, Warner, S. Rethinking scholarly communication. *D-Lib Magazine* 2004;**10**(9) [cit. 2014-04-22], ISSN 1082-9873. Available from: http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html