# STORING OF BIBLIOMETRIC INDICATORS IN CERIF DATA MODEL

Siniša Nikolić, Valentin Penca, Dragan Ivanović
{ sinisa_nikolic, valentin_penca, chenejac}@uns.ac.rs
***University of Novi Sad/Faculty of Technical Sciences, Novi Sad, Serbia***

**Abstract** – *This paper presents an example for classification of bibliometric indicators. The basics of the CERIF model and one solution of bibliometric indicator mapping to CERIF model are presented. Mapping was accomplished by using the CERIF semantic layer and applying appropriate classification scheme.*

## 1. INTRODUCTION

Evaluation in scientific-research domain is necessary as in all other areas. Today, we are witnesses of many efforts, aimed at objective and efficient evaluation of scientific-research results. Reasons for evaluation of scientific-research might be:

- hiring and promotion of the adequate staff in research institution
- allocation of material resources and/or awards to researcher/organization
- funding of scientific research

Evaluation in scientific-research domain [1] is a process based on critical analysis of information and data which leads to a judgment of merit.

The objectivity of the evaluation process involves:

- predefined criteria and methodologies
- evaluators who were not involved in the creation process of entities that have to be evaluated.

The following entities are subject of evaluation in scientific-research domain: publication, researcher, projects, patents and institutions. Since there is no global standard for evaluation of research data, institutions/organizations are forced to define their own processes for evaluation (rulebooks). In particular, process of evaluation distinguishes three characteristic approaches [2]:

1. Expert Group (Commission), which evaluates the results based on predefined rules. Commission expresses its result in form of quantity score for each entity that is a subject of evaluation.
2. Use of bibliometric indicator (impact factor, h-index, citation)
3. Combination of the previous two, i.e. the expert group which in its final assessment takes into account the value of the bibliometric indicators.

The third approach in the evaluation is introduced to overcome drawbacks of the first and the second approach (expert group and bibliometric indicator). Disadvantages of first approach are possible due to the lack of transparency in evaluation rules and in case when objectivity of Commission is taken into question. An example of evaluation, based on the first approach is RAE (Research assessment exercise) [3] framework, upon which the experts group evaluates the scientific results of the research institutions in the UK. Experts groups are formed to be competent for the assessment in particular science field. In the second approach, bibliometric indicators can be a subject of various manipulations. An example of impact factor manipulation can be a situation in which the editors "suggest" to authors to cite certain publications.

A combination of an expert group and the bibliometric indicators is probably the best approach for evaluation of research results. Advantages of this approach is fully discussed in the following literature [4], [5], [6], [7]. In the third approach, commission work could be much more efficient by using the adequate bibliometric indicators. [8].

Bibliometric indicators are especially important for the researchers and organizations [4], [9], as these measurements are often used in funding decisions, appointments, and promotions of researchers.

Electronic sources of bibliometric data are Web of Science [10], Scopus [11], and Google Scholar [12]. These databases have advantages and drawbacks and it is necessary to be cautious in accepting their data [13], [14], [15]. The ideal case would be a combination of data from multiple sources (databases).

In paper [16] is pointed out that for an effective evaluation of the scientific data, it is necessary to construct an information system which contains scientific-research results data and enables storage and computation of bibliometric indicators. Such system has to use a standard (unified) representation of the scientific-research data and the bibliometric indicators. Start-up point for building such system might be the CRIS system based of CERIF, for example CRIS UNS [17]. CERIF [18] model allows presentation and exchange of scientific-research data. Integration of bibliometric data in CRIS system is an essential precondition for evaluation of scientific-research data.

In following text are presented the basics of CERIF model, one possible classification of bibliometric indicators, and a part of CERIF model which enables representation of bibliometric indicators. Future research guidelines are discussed at the end of paper.

### 1.1. CERIF

Hereinafter we will present the main entities of the CERIF data model version 1.5.

- **Base Entities** - represent the core (basic) model entities. There are only three basic entities *cfPerson*, *cfOrganizationUnit* and *cfProject*.
- **Result entities** - A group of entities which includes results from scientific research. Representatives of this group are: *cfResultPublication*, *cfResultProduct* and *cfResultPatent*.
- **Infrastructure Entities** - represent a set of infrastructure entities that are relevant for scientific research. The entities which belong in this group are: *cfFacility*, *cfEquipment* and *cfService*.
- **Geographic bounding box** – This group of entities is used to define the exact geographical location of some entity instances in the CERIF model.
- **Indicator and Measure Entities** – Entities (*cfIndicator*, *cfMeasurement*) used for defining quantitative measurements for various CERIF entities.
- **2nd Level Entities** - Entities which further describe the Base Entities and Result Entities.
- **Link Entities** - are used to link entities from different groups. Typical entities of this group are: *cfOrganizationUnit_OrganizationUnit*, *cfOrganizationUnit_ResultPublication* and *cfResultPublication_DublinCore*.
- **Multiple Language Entities** - These entities provide multilingualism for CERIF data.
- **Semantic Layer Entities** - Provide different kinds of semantics in CERIF model. The entities in this group are *cfClassificationSheme* and *cfClassification*.

Figure 1 shows some of Base, Result, Link, Semantic and Measurement Entities which are relevant for the mapping proposed in this paper.

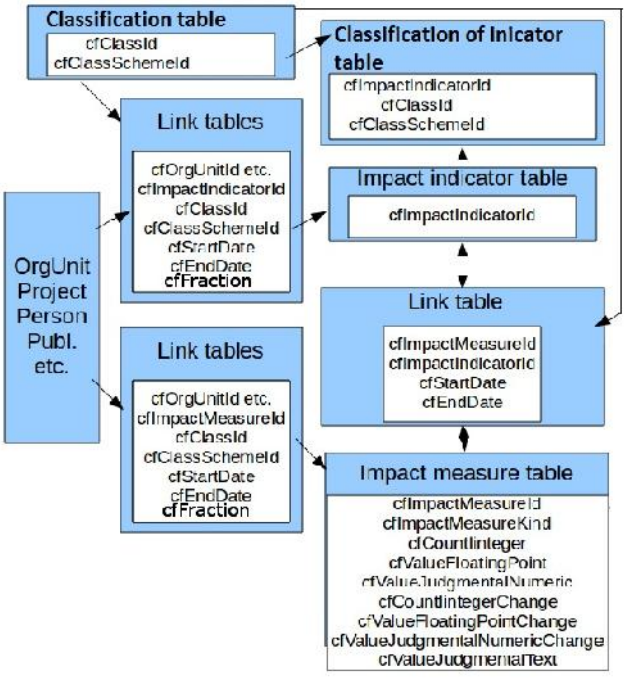


Figure 1 –entities of CERIF model

## 2. CLASSIFICATION OF BIBLIOMETRIC INDICATOR

This section provides an overview of the widely adopted bibliometric indicators.

There are three types (main classification) of bibliometric indicators [8]:

1. Quantity indicators that measure the productivity of a particular researcher or research group.
2. Performance indicators which measure the quality of a journal, researcher, or research group.
3. Structural indicators which measure connections between publications, authors, or research fields.

This paper will focus on quantity and performance indicators only (Figure 2).

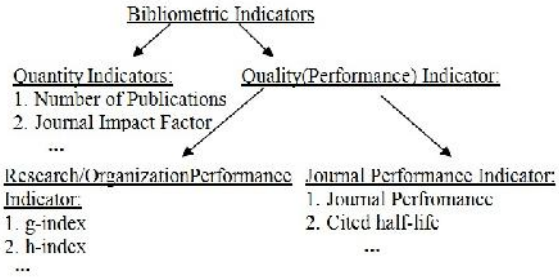


Figure 2 –Bibliometric classification

### 2.1. Quantity indicator

Quantity indicators are intended to measure the productivity of a researcher or a group. Relevant representatives from this group are: number of publications, number of citations, citation per publications, number of Publications in Top-ranked Journals, number of publication in ISI journals, number of publications per researcher for organization, number of citation per researcher for organization. From the name of each quantitative indicator, semantic (meaning) can be acknowledged. Thus, e.g. *Number of Publications* counts the number of articles published by a particular author or research group during a certain time period. The number of publications reflects the productivity of an author or a research group, but does not address the quality of the articles. In a case of groups comparing, one must bear in mind, that the number of publications is also influenced by the group's size. Details for all quantitative indicators are presented in [19].

### 2.2. Performance indicator

Beyond productivity, there are also additional criteria which is useful to consider in the bibliometric domain. Performance indicators help to identify the level of quality of an author or an organization (group of researcher) research. How often an article, an author, or a journal is cited by others, is an indication of performance. The citation number can be used in a different context (e.g. when calculating the average citations per year, a

citations number is divided by the number of years in which the citations were counted). Performance indicators are further classified into two subclasses; the first present indicators used to measure the quality of journals, while the second presents indicators used to measure the performance of researchers (individuals and organization).

#### 2.2.1. Journal Performance Indicators

The representatives of Journal performance indicators are: Journal Impact Factor (JIF), Journal to field impact score, Immediacy index, Cited half-life, Journal self-citations, Eigenfactor score and Article influence score.

Nowadays, the Journal Impact factor is probably the most used bibliometric indicator. JIF [20] is a measure of the frequency by which the average article in a journal has been cited in a particular year or period. It is one of the evaluation tools provided by Thomson Reuters [21] *Journal Citation Reports*® (*JCR*®). It represents the number of citations in the current year to any items published in a journal in the previous 2 years denominated with the number of substantive articles (source items) published in the same 2 years.

Obviously that scientific research progress is not equal in all science fields (e.g. theoretical sciences like astronomy and physics have lower progress than computer science or economy). Thus, in addition to the existing two year IF, the five-year IF was introduced.

Eigenfactor [22] is created in order to measure the importance (to emphasize the prestige and influence) of a journal in the scientific community. The theory behind Eigenfactor Metrics is that a single citation from a high-quality journal may hold more value than multiple citations from more peripheral publications. This rich and complex set of journal-to-journal relationships is used in an iterative calculation of citation exchange. The result is ranking of journals that not only reflects citation count, but embeds information about the JCR citation network and provides a complementary measure to citation ranking. One of the Eingfactor specialization can be the *Article influence score*, where Eingfactor is divided with the number of published articles.

#### 2.2.2. Researcher/Organization Performance Indicators

This category of indicators is designed to evaluate the performance of either individuals or research groups (organization). Most representative indicator of the *Researcher/Organization Performance* are crown Indicator [23], h-index [24], g-index [25]. H-index and g-index are indicators of researcher efficiency which take into account the scientific productivity (number of publication) of researchers and influence measured by the number of citations. The g-index and h-index have an ability to make a difference between „types“ of scientists (low producers, big producers, selective scientists and top scientists). Crown indicator measures the scientific impact of a researcher or a research group by dividing the average number of received citations (from a researcher or a research group) with the average number that could be expected for publications of the same type, during the same year, and published in journals within the same field.

## 3. CERIF model for storing bibliometric indicators

CERIF model initially provides storage of various metrics. Version CERIF_2008_1.0, for the first time, includes *cfMetrics* entity that provides storage of some type of metrics. The limitations of *cfMetrics* entity arise from the fact that he can be linked only with the entity cfResultPublication, which disable possible metrics for other scientific-research entities like people, organizations, projects, etc. euroCRIS [26] organization realized this and in recent versions (since version 1.3) of CERIF, included entities (*cfIndicator* and *cfMeasurement*) to overcome aforementioned limitation. Therefore, *cfIndicator* and *cfMeasurement* can represent any metrics that is not only related to publication, but also to other entities from scientific-research domain (Figure 1). The aforementioned entities are the start-up point for the construction of various quantitative measures. *cfIndicator* is a number or ratio (value on scale), which is derived from a series of observed facts. *cfMeasurement* is a dimension, quantity or capacity established in the measurement process. It is possible to make a connection between the *cfIndicator* and the cfMeasurement entity (e.g. all measurements that are necessary for specific indicators).

CERIF model allows making connection between measurement entities (*cfIndicator*, *cfMeasurement*) and with other entities of CERIF model (publications, persons, organizations) by using CERIF Link entities. For all link and measurement entities can be assign a specific semantics, defined by CERIF semantic layer.

The CERIF data model has the CERIF Semantic layer entities that enable classification of entities and relations between entities in accordance with some classification scheme. Those entities are *cfClassification* and *cfClassificationSheme,* which describe classes and classification schemes, respectively. A relation between a class and a classification scheme is established by using the primary key (*cfClassificationSchemeIdentifier*) from *cfClassificationSheme* as a foreign key in *cfClassification* entity. Also, in Semantic Layer it is possible to form a relation between two classes (with *cfClassification_Classification* entity) and relation between two schemes (with *cfClassificationSheme_ClassificationSheme* entity). Both relation entities belong to the group of *Link* entities, which are classified with appropriate classification scheme and contain identifiers of instances which are forming the relation. In addition to the list of entities that belong to the Semantic Layer, CERIF prescribes a specific vocabulary of possible classification values.

The following sections describe the mapping of classified bibliometric indicator from Section 2 to the entities of CERIF model and concrete example for storing bibliometric indicator values in CRIS UNS data model.

### 3.1. Mapping Schema

For the purpose of interoperability, the authors carried out mapping without changing the existing CERIF model, which means that it is not added, deleted, or changed neither of entities from the CERIF 1.5 model. It was necessary to construct new classification scheme to support the classification of and relations with bibliometric indicators from Section 2.

Mapping was started by defining new classification scheme (instance of *cfClassificationScheme*) with mnemonic name *BibliometricsIndicators.* Since bibliometric indicators are divided into quantitative and qualitative, two new classes are created (instance of cfClassification). Mnemonic names (*cfClassficationTerm*) of created classes are *BibliometricsQuantityIndicators* and *BibliometricsQualityIndicators*. Since it is possible further to divide the qualitative bibliometric indicators into two subgroups, two new classes (*JournalPerformanceIndicators* and *ResearcherAndOrganizationPerformanceIndicators*) are created. The hierarchy between the classes within the scheme *BibliometricsIndicators* is achieved by creating instances of *cfClassification_Classification* entities (semantic relationship between two existing instances of *cfClassification*). Each instance of *cfClassification_Classification*, in addition to identifiers (cfClassificationIdentifier1 and cfClassificationIdentifier2) of classes that participate in a relationship, also contains a description of their relationships. So in order to classify the relationship between the class *BibliometricsQualityIndicators* and her specializations *JournalPerformanceIndicators* and *ResearcherAndOrganizationPerformanceIndicators*, a new classification scheme *BibliometricsIndicatorsRelation* and appropriate class *IsDividedIntoSubgroups* is created.

In order to represent existing types of quantitative bibliometric indicators the following instances of *cfIndicator* are created: *NumberOfPublicaitons, NumberOfCitations, CitationPerPublications, NumberOfPublicationsInTopRankedJournals, NumberOfPublicationInISIIndices, NumberOfPublicaitonsPerResearcher, NumberOfCitationPerResearcher*. Every instance of *cfIndicator* that represents quantity indicator is classified with classification scheme *BibliometricsIndicators* and with their class *BibliometricsQuantityIndicators*.

In order to specify the journal quality bibliometric indicators, the following instance of *cfIndicator* entity are created: *JournalImpactFactor, 5YearJournalImpactFactor, ImmededіacyIndex, CitedHalf-Life, journalSelf-Citations, EingfactorScore, ArticleInfluenceScore*. In addition to mentioned instances, it was necessary to create multiple instances of *Journal Impact Field To Score* bibliometric indicator, so that each of them can represent a specific area of JCR [21]. Therefore, for each of the scientific field (215 of them) is created an instance in a form *JournalToFieldXXXImpactScore* where XXX represents the name of the appropriate of scientific field. Every instance of *cfIndicator* that represents journal quality indicator is classified with classification scheme *BibliometricsIndicators* and with their class *JournalPerformanceIndicators.*

The research/organization qualitative bibliometric indicators are represented with following instances of *cfIndicators* entity: *CrowIndicator, H-index and G-index.* Every instance of *cfIndicator* that represents researcher/organization quality indicator is classfied with classification scheme *BibliometricsIndicators* and with their class *ResearcherAndOrganizationPerformanceIndicators.*

When defining a concrete value of bibliometric indicator for particular researcher, organization or journal, a connection (Link entity) between bibliometric indicators (cfIndicator instance) with some basic (cfPerson, cfOrganization) or resulting (cfResult_Publication) entity instances is necessary to be established. CERIF Link Entities that are used to achieve the above mentioned links (connection) are *cfResultPublication_Indicator* (for journals), cfPerson_Indicator (for researchers), *cfOraganizationUnit_Indicator* (for the organization). The use of particular CERIF link entity mostly depends on the semantics and the purpose of bibliometric indicators. Thus, the quantitative bibliometric indicators (NumberOfPublication) could be used as a measure of publications for researchers, organizations or journals. Instances of Link entities (e.g. cfPerson_Indicator) are consisted of the entity identifiers that make a link (e.g. cfPersonIdentifier and cfIndicatorIdentifier), validation time period (cfStartDate, cfEndDate) and numerical value of bibliometric indicator (cfFraction).

Classification of the link entities (*cfResultPublication_Indicator, cfPerson_Indicator, cfOraganizationUnit_Indicator)* is necessary to define precisely the "type of relation". Therefore, in proposed model the classification scheme *RelationOfIndicatorWithCerifEntities* which contains the class with name *ValueOfBibliometricIndicator* is created.

### 3.2. Mapping example

This section presents the example (Figure 3) of storing the value of JIF for journal *Scientometrics* in 2010 and 2011. The example is based on a concrete data from CRIS UNS system. Data in example is represented in accordance to standard CERIF-XML version 1.5 which is an specific form of XML, for the exchange of research information metadata between research information systems [27]. Main entities in this example are bibliometric indicator JIF (element <cfIndic>) and journal Scientometrics (element <cfResPubl>).

Element <cfIndic> contains identifier of bibliometric indicator (random value), name (<cfName>), description (<cfDescr>) and classification (<cfIndic_Class>). Class *JournalPerformanceIndicators* (in accordance to Section XXX) is used for classification of JIF. In XML, the element <cfIndic> contains sub elements which include values of JIF (<cfFraction> 2.415; 2.44) for defined time period (<cfStartDate> <cfEndDate> 2010; 2011) and

identifier of element <cfResPubl> (<cfResPublId>). All <cfResPubl_Indic> elements are link entities that are classified with *ValueOfBibliometric Indicator class.*
It is obviously that CERIF element <cfResPubl> (Scientometrics) is used for journal description. Mentioned element contains following attributes: creation date of publication (<cfResPublDate>), ISSN (<cfISSN>), uniform resource identifier (<cfURI>), title (<cfTitle>), short description of journal (<cfAbstr>) and classification (<cfResPubl_Class>). Class *Journal* from CERIF 1.5 vocabulary provides necessary classification for observed element <cfResPubl>.

```
<?xml version="1.0" encoding="UTF-8"?>
<CERIF xmlns="urn:xmlns:org:eurocris:cerif-1.5-1"
xsi:schemaLocation="urn:xmlns:org:eurocris:cerif-1.5-1
http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF_1.5_1.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" release="1.5"
date="2013-01-10" sourceDatabase="CRIS UNS">
<!--A indicator record-->
<cfIndic>
    <!--id for IF record-->
    <cfIndicId>5672398</cfIndicId>
    <cfName cfLangCode="EN" cfTrans="o">Journal Impact Factor</cfName>
    <cfDescr cfLangCode="EN" cfTrans="o">The impact factor (IF) of an
    academic journal is a measure reflecting the average number of
    citations to recent articles published in the journal.</cfDescr>
    <!--instance of Link entity to describe the type of indicator-->
    <cfIndic_Class>
        <!--uuid for the term "Journal Performance Indicators"-->
        <cfClassId>458409635193853</cfClassId>
        <!--uuid for the CERIF scheme "Bibliometrics Indicators"-->
        <cfClassSchemeId>460652998251676</cfClassSchemeId>
        <cfStartDate>2000-01-01T00:00:00</cfStartDate>
        <cfEndDate>3000-01-00T00:00:00</cfEndDate>
    </cfIndic_Class>
    <!--instance of Link entity to describe a value of IF-->
    <cfResPubl_Indic>
        <!--id for journal Scientometrics-->
        <cfResPublId>2221909</cfResPublId>
        <!--uuid for class "Value Of Bibliometric Indicator"-->
        <cfClassId>883548128300062</cfClassId>
        <!--uuid for scheme "Relation Of Indicator
        With Cerif Entities"-->
        <cfClassSchemeId>872466912165806</cfClassSchemeId>
        <cfStartDate>2010-01-01T00:00:00</cfStartDate>
        <cfEndDate>2010-12-31T23:59:59</cfEndDate>
        <cfFraction>2.415</cfFraction>
    </cfResPubl_Indic>
    <cfResPubl_Indic>
        <cfResPublId>2221909</cfResPublId>
        <cfClassId>883548128300062</cfClassId>
        <cfClassSchemeId>872466912165806</cfClassSchemeId>
        <cfStartDate>2011-01-01T00:00:00</cfStartDate>
        <cfEndDate>2011-12-31T23:59:59</cfEndDate>
        <cfFraction>2.44</cfFraction>
    </cfResPubl_Indic>
</cfIndic>
<!--A publication (output) record-->
<cfResPubl> ... </cfResPubl>
<!--A classification scheme record-->
<cfClassScheme> ... </cfClassScheme>
<cfClassScheme> ... </cfClassScheme>
<cfClassScheme> ... </cfClassScheme>
</CERIF>
```

*Figure 3- CERIF XML representation of mapping example*

## 4. CONCLUSION

This paper presents the scheme for mapping bibliometric indicators to CERIF model. Mapping was conducted in accordance with the existing bibliometric classification which is presented in Section 2. Model was verified on data from CRIS UNS system. The proposed model provides the following:

1. unambiguous mapping and storing of retrieved bibliometric indicators
2. calculation of existing bibliometric indicators, as well as capability to add (define) a new one.

Proposed model is not an extension of CERIF 1.5, because it adds only the necessary classification scheme for bibliometric indicators. Therefore, the model has a potentially high level of interoperability with existing systems based on the CERIF model. Directions for further research are focused to support other bibliometric indicators that are not included in the proposed classification. In a long term, the intention is to support mapping for metrics from other domains (e.g. economic or social metrics).

## 6. REFERENCES


[1] Committee for the Evaluation of Research - CIVR, "Guidelines for Research Evaluation," Ministry of University and Research (MIUR), Jan. 2006.

[2] S. Nikolić, V. Penca, D. Ivanović, D. Surla, and Z. Konjović, "CRIS service for journals and journal articles evaluation," presented at the 11th International Conference on Current Research Information Systems, Prague, Czech Republic, 2012, pp. 323–332.

[3] RAE (Research assessment exercise) http://www.rae.ac.uk/]

[4] P. Weingart, "Impact of bibliometrics upon the science system: Inadvertent consequences?," Scientometrics, vol. 62, no. 1, pp. 117–131, Jan. 2005.

[5] J. Moravcová, "The Evaluation Methodology and Institutional Funding in Science." Ministry of Education, Youth and Sports Czech Republic, 06-Aug-2012.

[6] B. Meyer, "Research Evaluation for Computer Science An Informatics Europe report," Feb. 2008.

[7] J. K. Vanclay and L. Bornmann, "Metrics to evaluate research performance in academic institutions: a critique of ERA 2010 as applied in forestry and the indirect H2 index as a possible alternative," Scientometrics, vol. 91, no. 3, pp. 751–771, Jan. 2012.

[8] V. Durieux and P. A. Gevenois, "Bibliometric Indicators: Quality Measurements of Scientific Publication," Radiology, vol. 255, no. 2, pp. 342–351, 2010.

[9] J. Neuberger and C. Counsell, "Impact factors: uses and abuses," European journal of gastroenterology & hepatology, vol. 14, no. 3, pp. 209–211, 2002.

[10] Web of Science - http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

[11] Scopus - http://www.scopus.com

[12] Google Scholar - http://scholar.google.com/

[13] M. Franceschet, "A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar," Scientometrics, vol. 83, no. 1, pp. 243–258, Jun. 2009.

[14] P. Jacsò, "Pragmatic issues in calculating and comparing the quantity and quality of research through rating and ranking of researchers based on peer reviews and bibliometric indicators from Web of Science, Scopus and Google Scholar," Online Information Review, vol. 34, no. 6, pp. 972–982, 2010.

[15] N. K. Herther, "Research evaluation and citation analysis: key issues and implications," The Electronic Library, vol. 27, no. 3, pp. 361–375, 2009.

[16] J. Russell and R. Rousseau, "Bibliometrics and institutional evaluation," EncyclopediaofLifeSupport Systems (EOLSS) Part1, 2002.

[17] D. Ivanovic, G. Milosavljevic, B. Milosavljevic, and D. Surla, "A CERIF-compatible research management system based on the MARC 21 format," Program: electronic library and information systems, vol. 44, no. 3, pp. 229–251, 2010.

[18] CERIF FDM model - http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/MInfo.html

[19] C. Rehn and U. Kronman, Bibliometric handbook for Karolinska Institutet. Karolinska Institutet, 2008..

[20] Garfield E . Journal impact factor: a brief review . CMAJ 1999 ; 161 ( 8 ): 979 – 980

[21] Thomson sajt - Thomson Reuters

[22] Eigenfactor - http://www.eigenfactor.org

[23] H. F. Moed, R. E. Bruin, and T. N. Leeuwen, "New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications," Scientometrics, vol. 33, no. 3, pp. 381–422, Jul. 1995.

[24] L. Egghe and R. Rousseau, "An informetric model for the Hirsch-index," Scientometrics, vol. 69, no. 1, pp. 121–129, Apr. 2006.

[25] L. Egghe, "Theory and practise of the g-index," Scientometrics, vol. 69, no. 1, pp. 131–152, Apr. 2006.

[26]EuroCRIS Organization - http://www.eurocris.org

[27] J. Dvořák and B. Jörg, "CERIF 1.5 XML - Data Exchange Format Specification," 2013, p. 16.


## ACKNOWLEDGMENTS


Results presented in this paper are part of the research conducted within the Grant No. III 47003, Ministry of Education, Science and Technological Development of the Republic of Serbia.