University IS Architecture for the Research Evaluation Support

Laila Niedrīte, Darja Solodovņikova
University of Latvia, Faculty of Computing. Address: Raina blvd 19, Riga, LV 1586, Latvia.

***Abstract.** The measuring of research results can be used in different ways e.g. for assignment of research grants and afterwards for evaluation of project's results. It can be used also for recruiting or promoting research institutions' staff. Because of a wide usage of such measurement, the selection of appropriate measures is important. At the same time there does not exist a common view which metrics should be used in this field, moreover many existing metrics that are widely used are often misleading due to different reasons, e.g. computed from incomplete or faulty data, the metric's computation formula may be invalid or the computation results can be interpreted wrongly. To produce a good framework for research evaluation, the mentioned problems must be solved in the best possible way by integrating data from different sources to get comprehensive view of academic institutions' research activities and to solve data quality problems. We will present a data integration system that integrates university information system with library information system and with data that are gathered through API from Scopus and Web of Science databases. Data integration problems and data quality problems that we have faced are described and possible solutions are presented. Metrics that are defined and computed over these integrated data and their analysis possibilities are also discussed.*

***Keywords:** research evaluation, research metrics, data integration, information system, data quality.*

I. INTRODUCTION

Evaluation in science is necessary as in other fields. From peer performed unique review process, evaluation of research results is turning into a routine work based on metrics [1]. The number of different metrics is increasing rapidly, so it is very significant that they should be correctly chosen, computed and applied by organizations that are implementing their own evaluation framework. Therefore, the good praxis examples and guidelines that would allow to avoid different traps in such metrics based evaluation should be provided.

For the science evaluation, different methods including quantitative ones are applied, also the evaluation results are used for different purposes. Research indicators can be used starting from state level political decisions till individual researchers' decisions in his everyday work.

The usage of research indicators can be classified into five main groups [2]:

- General science policy group's typical activity is setting state level goals, for example, stating how many universities should be among top universities in the world.
- Funding allocation describes activities that use indicators in different calculations to compute the amount of funding.
- Organization and management is the group of activities that use indicators, for example, in Human Resource Management for career development or recruiting new research staff. The potential candidates to apply should have certain number of publications indexed in Web of Science.
- Content management and decisions refer mostly to individual researchers' activities, for example, the choice of journal where to publish is based on indicators.
- Consumer information, for example, attracting new students can be based on different university rankings that use also science indicators among others.

This classification [2] takes into account the usage of research performance indicators, but input indicators, such as number of researchers are out of the scope.

Later in this paper the data integration architecture oriented toward the collection and retrieval of bibliometric indicators is proposed. Therefore, let us take a closer look at this type of indicators. Bibliometric indicators can be divided into three main groups [3]:

- Quantity indicators or productivity indicators, for example, number of publications.
- Performance indicators or quality indicators, for example, h-index.
- Structural indicators allow to evaluate connections, for example, co-authors from different fields, institutions or countries.

The principles characterizing the best practice in metrics-based research assessment are given in the

ISSN 1691-5402

http://dx.doi.org/10.17770/ etr2017vol2.2528

"Leiden manifesto" [1], where 10 principles with explanations and examples are described.

Some of these principles [1] should be considered when designing a data integration architecture to support later effective research evaluation process, for example:

- Keep data collection and analytical processes open, transparent and simple.
- Allow to verify data and analysis by those, who are evaluated.
- Account for variation by field in publication and citation practices.
- Recognize the systemic effects of assessment and indicators
- Scrutinize indicators regularly and update them.

The authors of principles [1] state that not only journal publications, but also books for historians, conference proceedings for computer scientists, and national-language literature for social scientists should be considered.

When different sources are used to provide the needed data for different fields, the problem arises, are the results comparable. The authors [1] discuss this question also and argue that normalized indicators should be used, for example, the ones based on percentiles that are computed according to the citation distribution within the respective field.

When designing institutions' internal system, it should be taken into account that indicators change the system, therefore instead of one indicator, a set of indicators should be chosen, to avoid different biases.

Today there are many efforts trying to evaluate research results objectively and develop information systems to support these activities. Institutions develop their own or use commercial or non-commercial products to maintain data about research results.

A research information system in Scandinavia [4] is an example of such system that is implemented and used in Denmark, Finland, Norway, and Sweden and mostly contains integrated, high quality bibliometric data. The system is used for performance-based funding. Remarkable, that this system has also its own publication indicator that by weighting the results from different fields allows to compare them.

The requirements for research evaluation in Latvia are formulated in the regulations issued by the government and prescribe how the funding for scientific institutions is calculated [5], [6]. According to these regulations, the productivity of scientific work is evaluated according to the number of publications indexed in Scopus or Web of Science (WoS).

II. MATERIALS AND METHODS

The goal of this research is to develop and implement an architecture for bibliometric data collection for metric-based research evaluation support that takes into account the best practice principles and as a result provides a qualitative and comprehensive data collection.

This paper presents the components of this architecture, discusses the main integration problems that we have faced during implementation and solutions that we have chosen to overcome the shortcomings. This architecture is developed at the University of Latvia (UL) and the main component is implemented as a module of UL information system (LUIS).

A. Types and Choice of Evaluation Indices

Our architecture is discussed in detail in the later sections, but it must be mentioned that one distinguishing feature of it is the usage of external data sources Scopus API [7] and Web of Science API [8] provided by both largest publication citation indices.

Because one of the external data sources is Scopus, the data analysis possibilities directly in Scopus database were evaluated. Scopus provides SciVal tool that is based on some groups of metrics [9], for example: Productivity metrics measure the volume of output, Citation Impact metrics describe the influence of the output, for example, citation counts, Collaboration metrics give information on the research partnerships. The particular metrics that are used in SciVal are Scholarly Output, Journal Count, Journal Category Count, Citation Count, Cited Publications, Citations per Publication, Number of Citing Countries, Field-Weighted Citation Impact, Collaboration, Collaboration Impact, Academic-Corporate Collaboration. It must be mentioned that SciVal tool uses only publications indexed by Scopus as a data source. Metrics e.g. "Scholarly output", "Citation count" and others quantitatively measure different aspects of research activities. These aspects can be associated with the groups of measures e.g. productivity or citation impact.

Scopus API provides all data about UL publications, so all these metrics can be calculated also in LUIS. However, not all publications in LUIS have the same set of data due to different data sources, so not for all publications all these metrics can be calculated. The above-mentioned metrics and also some derived metrics can be calculated at the extent that the data are provided. As an example of derived metrics, the Scopus quartiles can be mentioned, that are calculated based on Scopus percentiles for CiteScore [10].

B. Scenarios of Obtaining Publication Data

Data about publications of the staff and students of UL are stored in the information system of the university (LUIS). On one hand, these data are gathered from multiple other information systems and on the other hand, authors and faculty and library staff have an opportunity to enter publication data directly into the university information system (Figure 1).

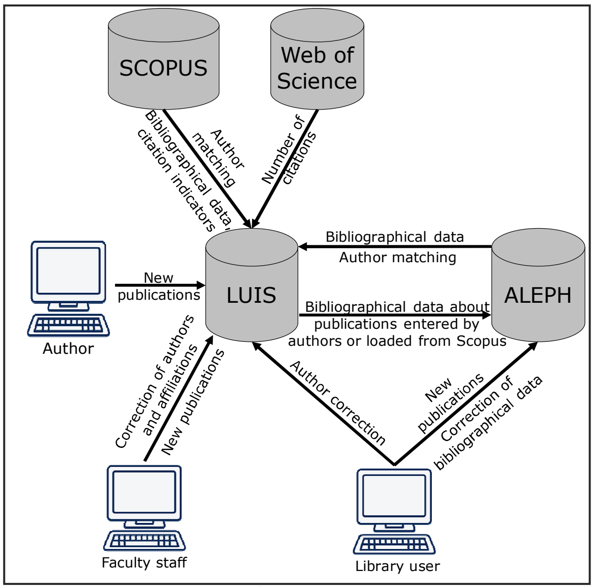


Fig.1. Scenarios of Obtaining Publication Data

C. Publications Added by Authors

The first and the preferred scenario of obtaining publication data is when these data are entered by an author of a publication. University employees, PhD and Master's degree students have a publications section of their profile in the university management information system (LUIS), where all publications are listed.

Before adding a new publication, an author must search for publications authored by him/her in the library information system (ALEPH) and LUIS with the purpose to discover whether the publication that the author planned to add to the system has already been entered in ALEPH or LUIS. In such case the author can just select it from the list and add it to the profile.

If the desired publication is not found, to add it an author must select a type of the publication and then enter bibliographical information, which includes: title of the publication, field, co-authors, affiliations of authors, year and place of publication, publisher, number of pages, ISBN, ISSN, web link, keywords. Besides, an author must indicate the status of the publication: published, submitted for publication, developed or under development, attach publication files (at least a book cover) and indicate whether the publication files may be made public. It is also possible to select databases where the publication is indexed and write any other additional information in comments.

In addition to entering new publications, authors also have an opportunity to unlink publications from their LUIS profiles that were erroneously automatically added to them during the synchronization process, when publication data are loaded from ALEPH or Scopus (see Sections E and F).

D. Publications Added by the Faculty Staff

Each faculty of the university can designate a person responsible for entering and editing information about publications authored by the faculty members: employees and students. Such faculty user can add new publications written by the faculty members by employing the same procedure as publication authors, which was described in the previous section of the paper.

Besides, a faculty user can change authors and author affiliations for the existing publications authored by the faculty members. This is necessary to correct authors that were erroneously automatically assigned to the publication during the synchronization process.

E. Publications Added by the Library Staff

Publications that are not indexed by Scopus can be also added to the library information system ALEPH by the library staff. This is done when a library user comes across a new publication authored by the university members in some journal or conference proceedings or when the information about a new publication is obtained from the list of recently indexed publications in Web of Science database, which is monthly distributed by Web of Science. The bibliographical information about a new publication is entered in ALEPH and during the synchronization process is also loaded in LUIS to ensure that LUIS always stores data about all publications available in ALEPH.

When the new publication is loaded in LUIS from ALEPH, the author detection is conducted, when for each author of a publication, a corresponding person in LUIS is searched for. If the person is found, the publication is added to his/her profile. For author matching, names and surnames of authors are used. Since authors tend to use different spelling versions of their names and surnames, when special characters of the Latvian language are present in their names or surnames, Jaro-Winkler similarity [11] is used to find the most similar name/surname combination of a person. After process testing and evaluation of experiment results, it was discovered that the most appropriate threshold for Jaro-Winkler similarity to perform name and surname matching is 0.93 and this coefficient is currently applied to consider name-surname combination similar.

In addition to entering new publications to ALEPH, library staff are also responsible for correcting and supplementing bibliographical data of publications added by authors and by faculty staff and of publications data imported from Scopus database (see Section F).

F. Publications Loaded from Scopus

Another data source that is used to populate publication data in LUIS is Scopus database. Data about articles published during the last 2 years are loaded from Scopus to LUIS daily and data about all other articles are loaded weekly. The synchronization process uses Scopus API to obtain bibliographical data and citation metrics of articles authored by UL staff and students indexed by Scopus. The following information is extracted from Scopus about each publication: unique identifier, publication title, journal or proceedings title, ISSN, ISBN, DOI, page range, volume, issue, publishing date, type and subtype of a publication, as well as author information: unique author identifier, name, surname, author affiliation, H-index and publication affiliation information: name, city, country. Affiliations are associated with authors as well as with publications directly. In addition to bibliographical information, citation metrics are also obtained that include the following information: number of citations, Source Normalized Impact per Paper (SNIP) [12], the SCImago Journal Rank (SJR) [13], CiteScore. The last 3 metrics are calculated and obtained for the particular joirnal or conference proceedings (not for the particular publication) and for the particular subject areas. Previously, it was possible to obtain Impact per Publication (IPP) metric [14], which is not available from Scopus anymore, so this number is retained for previously loaded publications.

The first step of the Scopus synchronization process is publication recognition phase, when publications obtained from Scopus are mapped with the existing publications in LUIS to avoid creation of duplicates and detect new previously non-existing publications. The recognition is firstly based on the Document Object Identifier (DOI) which is unique for every publication. If the matching publication with the same DOI is not found in LUIS the recognition based on the title and publication year is applied, i.e. for each publication obtained from Scopus for the first time, the process searches for a publication with the same year and similar title in LUIS. Jaro-Winkler similarity is used again to detect the existing publication in LUIS with the most similar title, because variations of title spelling as well as data quality issues are sometimes present in data. To perform title matching, we use the same threshold for Jaro-Winkler similarity (0.93) and this coefficient is currently applied to consider titles similar.

If the matching publication record is found in LUIS, its citation metrics are updated and the link between this publication and Scopus record is established. If the publication is new, it is added to LUIS with all its bibliographical information and citation metrics. In case of a new publication, author matching is also performed, when for each author of a publication affiliated with the University of Latvia, a corresponding person in LUIS is searched for. If the person is found, the publication is added to his/her profile. For author matching, firstly author Scopus identifier is used, which allows to find authors that were previously loaded from Scopus. If a corresponding person is not found by author Scopus identifier, names and surnames of authors are used. Since authors tend to use different spelling versions of their names and surnames, when special characters of the Latvian language are present in the name or surname, Jaro-Winkler similarity with the threshold of 0.93 is used to find the most similar name-surname combination of a person. If a corresponding person is found by his/her name and surname, author Scopus identifier is saved for the person for matching future publications.

After a new publication record is loaded from Scopus to LUIS, it is also automatically added to ALEPH and later checked by the library staff, the bibliographical information is supplemented and possible errors are corrected.

G. Publications Loaded from Web of Science

We are also using Web of Science web services as an additional source of information about publications. The information obtained from WoS includes: unique identifier, title, issue, pages, publication date, journal or proceedings title, volume, book series title, DOI, ISSN, ISBN, number of citations. The information about authors includes just author names, surnames and in some cases also Researcher identifier in the web services version, which is available to the University of Latvia. Since the affiliation of authors is not available, we have discovered that author matching process for Web of Science data produces too many incorrectly identified authors, therefore, it was decided to add new Web of Science publications manually.

However, we match Web of Science data with existing publications in LUIS, loaded from Scopus or entered previously by authors, faculty staff or library employees. Just as for publications loaded from Scopus, we use DOI as the primary data unit for matching and title and year of publication as the secondary data unit for searching for publications that are not found by DOI. For all matched publications, we update Web of Science citation number.

H. Internal UL Index for Publication Evaluation

Due to different types and levels of publications in addition to the ones indexed in Scopus or WoS some system that at institutions level systematizes publications can be introduced.

In 2013 the University of Latvia introduced their own internal index [15] for evaluation of publications. This index is calculated from all publications in LUIS system. Index can be calculated at the individual researcher's level or the faculty level. Index considers a publication type, publication level and number of authors. According to the type and level, points are calculated and divided with the number of authors.

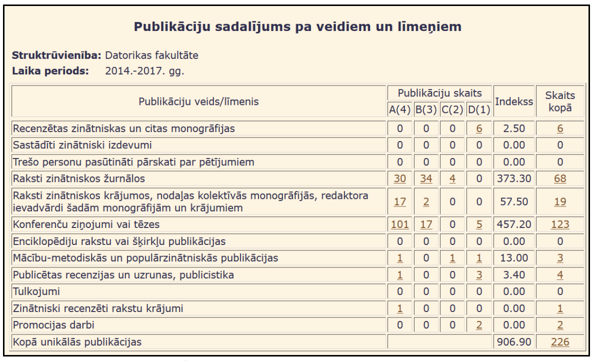
For example, a publication type can be "Journal publication", and within this type publications are classified according to their significance. So for this type some level examples are "Indexed in WoS Q1 or Q2" or "Indexed in WoS or Scopus".

On the one hand this index considers all publications, differentiates their significance according to their type, but there are also some controversial issues, e.g. division with the author count, that need to be discussed and improved. At the moment, this index is calculated, the LUIS system provides also the interface for analysis of this index, but in praxis this index is not used yet for evaluation of the research results.

Alternative ways how to evaluate the research output are being searched due to several reasons. Despite the fact that the calculation of state budget financing for the institution depends on the publications count indexed in Scopus or WoS, these indexes show uneven distribution among different fields. According to the UL's publication count for time period 2012 -2015, physics, natural sciences and engineering are prevailing [5]. However, it does not mean that researchers from other disciplines do not work or their results are not significant.

I.Analysis Tools in LUIS for Research Evaluation

Analysis tools in LUIS provide the possibility to evaluate an individual researcher or a faculty. For the faculties, the internal UL index can be also calculated (see Figure 1). The publication registration module allows to gain an insight about the quantity and quality of publications of the faculties' researchers.



Publikāciju sadalījums pa veidiem un līmeņiem

Struktrūvienība: Datorikas fakultāte
Laika periods: 2014.-2017. gg.

Publikāciju veids/līmenis	Publikāciju skaits				Indekss	Skaits kopā
	A(4)	B(3)	C(2)	D(1)		
Recenzētas zinātniskas un citas monogrāfijas	0	0	0	6	2.50	6
Sastādīti zinātniski izdevumi	0	0	0	0	0.00	0
Trešo personu pasūtināti pārskati par pētījumiem	0	0	0	0	0.00	0
Raksti zinātniskos žurnālos	30	34	4	0	373.30	68
Raksti zinātniskos krājumos, nodaļas kolektīvās monogrāfijās, redaktora ievadvārdi šadām monogrāfijām un krājumiem	17	2	0	0	57.50	19
Konferenču ziņojumi vai tēzes	101	17	0	5	457.20	123
Enciklopēdiju rakstu vai šķirkļu publikācijas	0	0	0	0	0.00	0
Mācību-metodiskās un populārzinātniskās publikācijas	1	0	1	1	13.00	3
Publicētas recenzijas un uzrunas, publicistika	1	0	0	3	3.40	4
Tulkojumi	0	0	0	0	0.00	0
Zinātniski recenzēti rakstu krājumi	1	0	0	0	0.00	1
Promocijas darbi	0	0	0	2	0.00	2
Kopā unikālās publikācijas					906.90	226

Fig.2. Report about faculties' results in LUIS publications module

Another useful tool provides the possibility to select and extract detailed information about publications, that is collected and integrated from all information sources that are included into previously described architecture including WoS and Scopus. In most cases the information added to the publication records is the actual citation count, the information about the Journal or book series e.g. SJR, SNIP, IPP, and Cite Score. Originally Cite Score provides percentiles that are used in LUIS publication module to compute Scopus quartiles.

III. RESULTS AND DISCUSSION

To demonstrate the volume of data collection used for research evaluation at the University of Latvia, we are will provide several statistical indicators. The total number of publications by UL members for the last 30 years is 42417. 6967 publications out of them are indexed in Scopus and 7764 publications are indexed in WoS.

Further in this section 3 different analysis scenarios for research output evaluation that can be implemented with the new publication module and data integration infrastructure are described.

The following parameters were applied for the data extraction for all research questions: Faculty name "Faculty of Computing" and Time period "2013 – 2016".

For the 1st analysis scenario the following research question was formulated: "How many faculty publications are indexed in Scopus or WoS comparing to all faculty publications?". Figure 3 shows the trend that the whole number of publications decreases and the number of indexed publications increases and in the year 2016 there are only 7 publications that are not indexed.

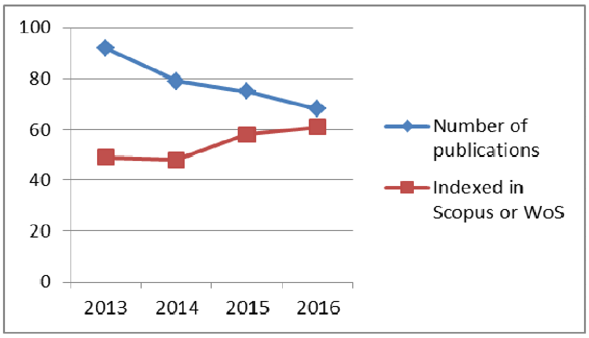


Fig.3. Indexed publications vs. all publications

For the 2nd analysis scenario the following research question was formulated: "How many publications of the faculty are indexed only in WoS and not in Scopus". In Figure 4 two measures to compare are given: the number of publications indexed in WoS and the number of publications that are indexed only in WoS, but not in Scopus. The proportion between both metrics persists over time, which may indicate a relative stability of authors choice where to publish their results and which conferences to attend.

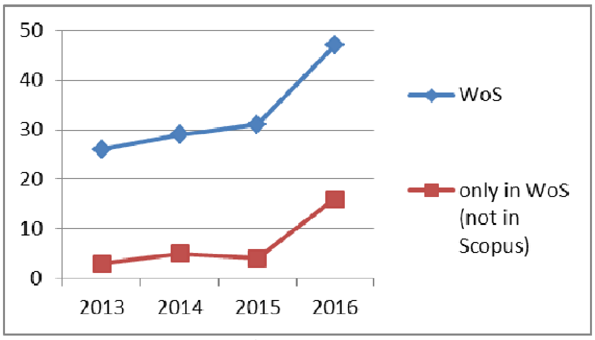


Fig.4. Overlapping of WoS and Scopus publications

For the 3rd analysis scenario the following research question was formulated: "How many publications out of all faculty publications indexed by

Scopus have Scopus quartiles computed according to Cite Score and how many of them have Q1 or Q2?". The results in Figure 5 show that over the time period the number of publications of the faculty indexed by Scopus is getting closer to the number of publications that are published in journals or book series that have CiteScore percentiles, from which we computed quartiles. Among the last ones, the proportion of publications that have Scopus quartiles Q1 or Q2 remains unchanged over the time period.

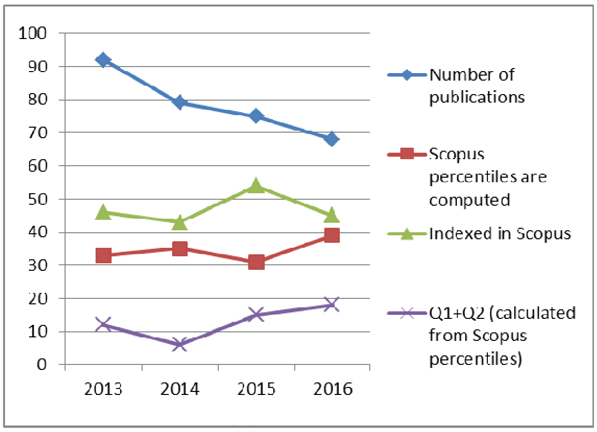


Fig.5. Analysis of faculty's Scopus publications

The analysis results can be used not only for research results evaluation but also as an information for more careful choice of journals or conferences where publications should be submitted with respect to indexing in Scopus or Wos, and also the best possible quartile. This can be useful for young scientists and doctoral students.

IV. CONCLUSIONS

We have developed and implemented the architecture for research evaluation at the University of Latvia. The data collection contains a wide variety of different publication types, fields and sources. The data collection is automatically updated from external data sources Scopus and WoS on regular daily basis.

Not all data quality and identification problems were solved, therefore, special user interface was developed and provided for faculties and scientific institutes to control the automatic data integration and to make corrections if needed. So the users will have more confidence that the evaluation decisions are made over correct data.

However, the different analysis possibilities can be used more intensively, mostly different publication lists, e.g. for study programs accreditation purposes, are produced and used. Also each researcher can see the list of publications in his or her LUIS profile and during the automatic CV generation option in LUIS the actual publication list is added. These, of course, are not the goals why the system was produced, but in the starting point, while all stakeholders are getting familiar with the provided features, also some operational usage is acceptable.

Some examples of the intended usage of the system were also demonstrated in this paper.

REFERENCES

[1] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols, "The Leiden Manifesto for research metrics", in *Nature*, vol. 520(7548), 2015, pp. 429.

[2] J. Kosten, "A classification of the use of research indicators", in *Scientometrics*, vol. 108(1), 2016, pp. 457-464.

[3] S. Nikolić, V. Penca, D. Ivanović, D. Surla, and Z. Konjović, "Storing of Bibliometric Indicators in CERIF Data Model", in *Proceedings of the ICIST 2013 Conference (CD), Kopaonik*, Vol. 3, 2013.

[4] G. Sivertsen,. "Data integration in Scandinavia", in *Scientometrics*, vol. 106 (2), 2016, pp. 849-855.

[5] I. Rampāne, G. Rozenberga, Latvijas Universitātes publikāciju citējamība datubāzēs (2012-2015), 2016. Available: dspace.lu.lv, [Online], [Accessed: 15.03.2017]

[6] "Ministru kabineta noteikumi Nr.1316 Kārtība, kādā aprēķina un piešķir bāzes finansējumu zinātniskajām institūcijām", 2013.gada 12.novembrī, Available: likumi.lv, [Online], [Accessed: 15.03.2017].

[7] Scopus citation database, Available: www.scopus.com [Online], [Accessed: 15.03.2017].

[8] Web of Science citation database, Available: https://webofknowledge.com/ [Online], [Accessed: 15.03.2017].

[9] L. Colledge, and R. Verlinde, "Scival metrics guidebook". *Netherlands: Elsevier*, 2014.

[10] H. Zijlstra, R. McCullough, "CiteScore: a new metric to help you track journal performance and make decisions", December 8, 2016. [Online]. Available: https://www.elsevier.com/editors-update/story/journal-metrics/citescore-a-new-metric-to-help-you-choose-the-right-journal [Accessed: March 15, 2017].

[11] W. Winkler, "The state record linkage and current research problems", Technical report, Statistics of Income Division, Internal Revenue Service Publication, 1999.

[12] H.F. Moed, "Measuring contextual citation impact of scientific journals, Journal of Informetrics", vol. 4(3), pp. 265–277, 2010

[13] B. Gonzalez-Pereira, V. P. Guerrero-Bote, & F. Moya-Anegon, "A new approach to the metric of journals' scientific prestige: The SJR indicator," Journal of Informetrics, vol. 4, pp. 379–391, 2010

[14] J. Hardcastle, "New journal citation metric – Impact per Publication", July 22, 2014. [Online]. Available: http://editorresources.taylorandfrancisgroup.com/new-journal-citation-metric-impact-per-publication/ [Accessed: March 15, 2017]

[15] "LU Rektora rīkojums par publikāciju līmeņiem". Nr.1/278 (09.10.2013), University of Latvia, Internal management document, 2013.