FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 43 (2018) No. 2
ISSN 0867-6356
DE GRUYTER OPEN DOI: 10.1515/fcds-2018-0008 e-ISSN 2300-3405

Architecture Enabling Adaptation of Data Integration Processes for a Research Information System

Darja Solodovnikova*, Laila Niedrite*, Aivars Niedritis*

Abstract. Today, many efforts have been made to implement information systems for supporting research evaluation activities. To produce a good framework for research evaluation, the selection of appropriate measures is important. Quality aspects of the systems' implementation should also not be overlooked. Incomplete or faulty data should not be used and metric computation formulas should be discussed and valid. Correctly integrated data from different information sources provide a complete picture of the scientific activity of an institution. Knowledge from the data integration field can be adapted in research information management. In this paper, we propose a research information system for bibliometric indicator analysis that is incorporated into the adaptive integration architecture based on ideas from the data warehousing framework for change support. A data model of the integrated dataset is also presented. This paper also provides a change management solution as a part of the data integration framework to keep the data integration process up to date. This framework is applied for the implementation of a publication data integration system for excellence-based research analysis at the University of Latvia.

Keywords: integration architecture, adaptation, research evaluation, research metrics, data quality, data model

1. Introduction

Evaluation in science is increasingly turning into routine work based on metrics [10]. Research metrics can be used for different purposes [14]: science policy-making at state level, e.g. distribution of research funding, organisation and management activities; in human resource management for recruiting or promoting employees involved in research, content management and decisions at individual researchers' level; where and what to publish, and providing consumer information; and university rankings that include science indicators. All these examples belong to the research performance indicators. Besides, other indicator types may also be used e.g. input indicators, such as number of researchers.

* University of Latvia, Faculty of Computing, Riga, Latvia, e-mail: {darja.solodovnikova, laila.niedrite, aivars.niedritis}@lu.lv

It is obvious that there is a wide range of different metrics available that are also intensively applied. Organisations use different measures to align their activities to strategic goals [21].

At research institution level, quantitative metrics such as number of scientific papers, amount of funding, number of scientific staff and many others are commonly used for research evaluation, whereas the strategy of the institution can be set to achieve ambitious scientific goals. Therefore, the question arises as to how more quality-oriented aspects of the research outcomes can be measured.

Quality-oriented research evaluation is already performed, usually at state level, for example, to allocate funds to excellent institutions. Different measurement methods can be used: peer review-based models, publication count-based models and citation-based models [1]. These methods can also be adapted at institutional level.

One specific type of indicators that characterize the research output are the bibliometric indicators. These can also be classified as quantity and quality indicators [20]. For example, the publications count measures the research output quantitatively, but an example of a quality indicator is h-index.

Research information management has many similarities with data the integration field: many data sources with inconsistent data models, heterogeneity, and many involved stakeholders with diverse goals [22]. When building a data collection and integration system for effective science evaluation, some principles are important [10], for example, data collection should be transparent, the institutions and persons that are evaluated can verify the data provided for evaluation, the indicator values should be updated regularly.

To supply an appropriate dataset for evaluation of quantitative and qualitative aspects of research outputs, we provide a framework that ensures the use of an appropriate, qualitative and regularly updated dataset, and that has the following features: metric computation formulas are discussed and are valid, the university employees are involved to verify the process, data from various available sources are integrated to achieve an overall view of the scientific activity of an institution.

One of the solutions that can be used for storing indicators is a data warehouse. Data warehouse system is "a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management decisions" [11]. The multidimensional data model that is typical for the data warehouses must be implemented in alignment with the information requirements [27] of institutions. These requirements express the performance indicators of an organisation. Data warehouses are designed for querying and analysing data to evaluate the progress of institutions.

A data warehouse should provide accurate and historically correct information to users to support the decision-making. This means that the data warehouse must reflect all changes that occur in the analysed institution’s processes. In addition, data sources of the data warehouse can also change during this time. All changes can invalidate the data warehouse infrastructure, e.g. data warehouse database schema and data collection processes. To avoid the loss of history and ensure correct information, different solutions exist, e.g. data warehouse schema versioning [7].

Many information sources have been used at the University of Latvia (LU) for a while to gain an insight into the actual situation with research outcomes, but the research evaluation process needed improving by providing integrated information oriented towards scientific excellence. The requirements for research evaluation in Latvia are declared in the regulations issued by the government and prescribe how the funding for scientific institutions is calculated [2]. As stated in these regulations, the productivity of scientific work is evaluated according to the number of publications indexed in Scopus or Web of Science (WoS). These quantitative data must be extended with data necessary for computation of qualitative indicators.

In this paper, we will propose a research information system for bibliometric indicator analysis that is incorporated into the adaptive integration architecture based on ideas from a data warehousing framework for change support. However, the proposed solution is not a traditional data warehouse architecture. The main reasons for developing a new data integration solution, based on ideas from the data warehousing field, are the data quality issues, such as, data accuracy, timeliness, identification and others. Data integration and solutions for data quality issues are also typical for data warehousing, however, keeping the users involved in the data quality controlling, providing the possibility to update the collected information by the users, and also sending back updated information to the data sources are outside of the typical data warehouse architecture.

This paper discusses data integration flows and data integration problems, including data quality issues. A data model of the integrated dataset is also presented. Based on this data model and integrated data, examples of quality-oriented metrics and analysis results of them are provided. The present paper is an extended version of our paper [19]. This paper also provides a change management solution as part of a data integration framework to keep the data integration process up to date.

2. Related Work

Various methods for implementation of data integration processes can also be applied in research information systems. One of the recent approaches in the field of data integration focuses on the mappings between models of different systems. This mapping definition approach provides specifications for data integration processes and is used in a research information system developed in Germany [22].

The research that was conducted in China [3] proposes a data integration framework and technology based on metadata. This framework was used for the implementation of the research management system for the Chinese Academy of Sciences. The main components of this metadata-based resource integration method support the creation of a metadata standard, definition of ETL (extract, transform, load) operations based on metadata, providing a resource catalogue and others.

An academic data warehouse that supports not only the academic processes but also the research management and evaluation can be a solution for research data integration at university level [4]. The authors present the architecture of a Business Intelligence system for universities, where the integration process is defined according to a methodology that is based on ontology for the data source integration.

Many research papers describe implementations of research information systems in different countries, data models of these systems and data sources. These papers do not usually describe specific data integration problems and novel solutions to data integration problems, however, they characterise the variety of data needed for research evaluation, and different research data analysis applications.

A research information system in Scandinavia [23] is an example of such a system that has been 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. It is remarkable that this system also has its own publication indicator that allows comparison of the results from different fields by weighting them.

The Polish performance-research funding system allows the evaluation of 65 parameters. 962 research units provided data about more than a million research outcomes for the 4-year period. The data collection process was performed through submission of the questionnaire via the Information System on Higher Education in Poland. The study [15] was performed to find the most important metrics to facilitate the transition to a more targeted system to meet the excellence requirements, where only the most important metrics are reported. The research showed that many of the existing metrics are not significant for the evaluation.

The Italian experience [6] shows the implementation of a research information system in Italy, where 66 Italian institutions introduced IRIS, that is a system based on DSpace [5] and customised for the Italian environment. Entities, attributes and relations in this system are compliant with the CERIF ontology [12], [13]. The huge amount of data collected allowed the comprehension of the whole situation in research and, for example, to develop new publication strategies. The authors of the study also mention the problems with the data quality, when not all institutions control the data collection process and do not implement data validation processes of data provided by researchers.

In addition to the typical data integration problems, e.g. data model mapping, metadata definitions that are used for developing a semi-automatic data collection, data quality problems etc., our solution for the data integration problem of research information also considers the change problem of data source models and analysis requirements.

The solution to handling data source evolution problems in the integration field was proposed in the paper [17]. The presented approach is closest to ours and could be applied when data from sources are obtained using wrappers, for example, by means of Rest API in JSON format. The authors propose the use of big data integration ontology to define the integrated schema, source schemata, their versions and local-as-view mappings between them. When a change at a data source occurs, the system steward supplements the ontology with a new release and a new wrapper that allows unchanged and new attributes to be added to the changed source. Our approach differs in that our architecture is capable of handling not only changes in data sources, but also requirements and schema of the integrated system.

3. Motivating Example

The evolution that occurred during the operation of the research information system motivated us to propose a solution for automatic processing of such changes to save time and resources of developers necessary to handle changes. Below is a brief overview of the system and the history of changes that occurred during its operation. The detailed description of the system, as well as solutions used to propagate changes, is provided in section 5.

The research information system was first implemented at the University of Latvia in 2013. The initial version of the system integrated data available in the university information system with data from the library information system.

After successful implementation of the initial version of the research information system at the University of Latvia, new requirements regarding the scientific activity of institutions were issued by the government and the new requirements for reporting on publication activity became topical. To satisfy such requirements, the research information system had to be adapted to include additional data about publications authored by LU employees indexed in the citation databases Scopus and Web of Science. This information can be obtained automatically, therefore, new data sources had to be added to the integration architecture of the research information system. To automate adaptation of the research information system, we introduced the adaptive integration architecture that is described in the following section. After introduction of the architecture, further changes to the research information system were successfully handled semi-automatically. Examples of such changes, together with their solutions, are illustrated in section 5.3.

4. Adaptive Integration Architecture

To support adaptation of data integration processes to changes in underlying data sources and new data requirements, we propose an adaptive integration architecture depicted in Figure 1. The architecture is an adapted version of the data warehouse evolution framework presented in [24]. The main goal of the data warehouse evolution framework is to propagate changes in a data warehouse schema by means of creation of new data warehouse schema versions. The goal of the integration architecture is to automatically obtain the most complete up-to-date valid data from multiple data sources and accumulate them in the central data repository. In case of the adaptive integration architecture, we propose that the central data repository is used, instead of a data warehouse, due to the necessity to validate, correct and supplement data by users, as well as share data with other systems.

4.1. Components of the Architecture

The architecture is composed of five layers: source layer, middleware, integration layer, adaptation layer and access layer. Components of the architecture are included in layers according to their operation, however, all the metadata created and used by the components in different layers are stored in the metadata repository and the integrated data are stored in the central data repository, which is implemented as a relational database. The functionality of each component is described in the following subsections.

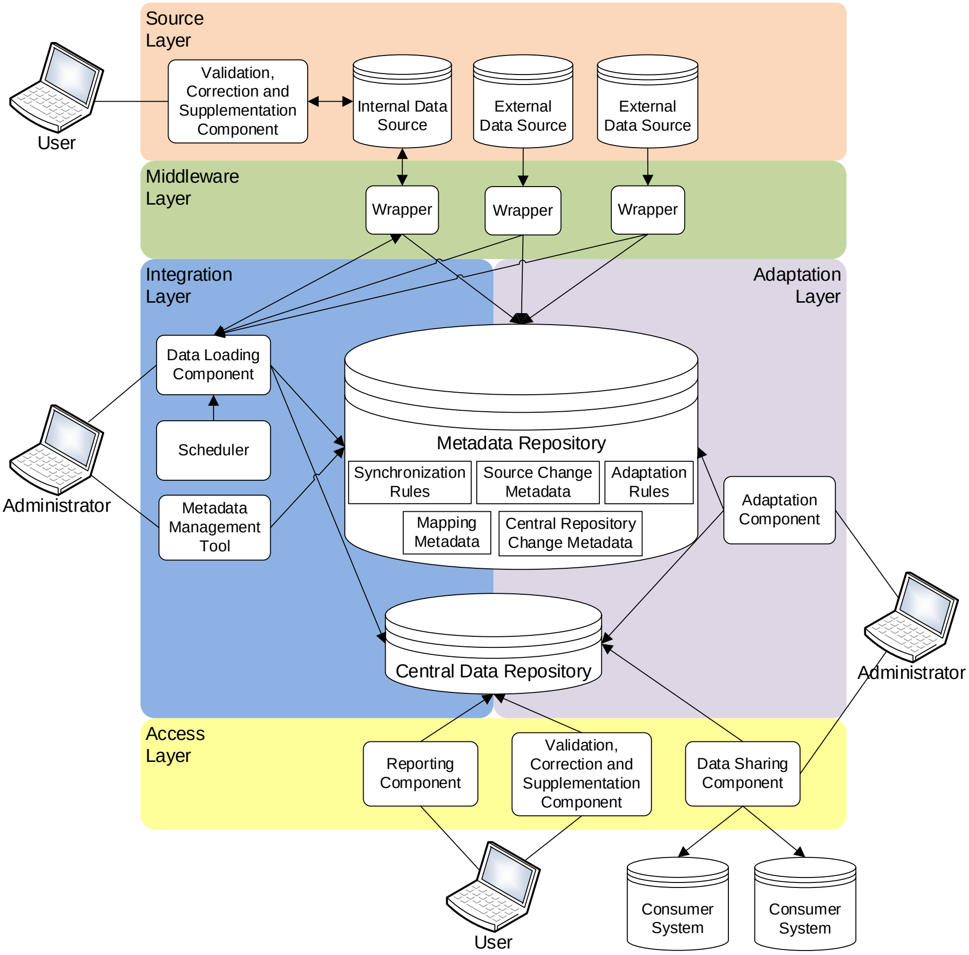

Figure 1. Adaptive integration architecture

4.1.1. Source Layer

In the source layer, there are external and internal data sources. External data sources are used only to obtain data for the central repository. Internal data sources are used not only to obtain data for the central data repository but also for data synchronisation. Data available in internal data sources is validated, corrected and supplemented by users, by means of the corresponding component of the architecture.

4.1.2. Middleware Layer

The middleware layer is composed of wrappers that implement interfaces for data acquisition supported by corresponding data sources (for example, web services, API, etc.). Wrappers of external data sources implement only one-way communication to obtain data that must be loaded into the central data repository. However, wrappers of internal data sources are used to load data from the central repository to the data source, as well as to support data sharing and synchronisation with other systems.

To implement the adaptation of integration processes, both wrappers of internal and external data sources must also be capable of tracking changes in views of source data provided by interfaces. The information about changes is necessary to adapt the integration processes semi-automatically.

4.1.3. Integration Layer

Integration processes that populate the central data repository with new data and update existing data are included in the integration layer. Data are loaded into the central data repository and updated by the data loading procedures executed via the data loading component, which may be launched by the system administrator or scheduled to be executed regularly. The data loading component must also be able to generate data loading procedures based on the metadata. Another function performed by the data loading is to prepare and provide data for wrappers of internal data sources.

The operation of the integration and adaptation layers is based on the metadata in the metadata repository. The metadata management tool is used by the administrator or developer to define metadata in the metadata repository. The metadata repository incorporates five types of interconnected metadata. Mapping metadata defines the logics of the data loading procedures. It stores the correspondences between data obtained from the sources and tables and columns in the central data repository as well as necessary transformations that must be made during the data loading process. Synchronisation rules define constraints for properties of data records gathered from different data sources that must be met to consider these data records to be the same. During the integration process, synchronisation rules are applied according to their priority. Information about changes in data sources obtained from wrappers is accumulated in the source change metadata. Adaptation rules specify adaptation options that must be implemented for different types of changes. Finally, the metadata repository also includes the central repository change metadata, which accumulates potential changes in the central repository schema.

4.1.4. Adaptation Layer

The core component of the adaptation layer is the adaptation component that processes changes in data sources and requirements for data. The main idea of the adaptation component is to generate several potential changes in the central data repository for each change in a data source and to allow an administrator to choose the most appropriate change that must be implemented. To achieve the desired functionality, the adaptation component uses data from the metadata repository. To implement certain kinds of changes, additional data may be necessary that cannot be identified automatically, for example, transformations for missing properties of data records. In such case, these data are supplied by the administrator via the adaptation component and are saved in the adaptation rules in the metadata repository.

In addition, the adaptation component also allows the administrator or developer to initiate changes in the central data repository and data loading procedures to handle new or changed requirements for data. The history of chosen central data repository changes that are implemented to propagate evolution of data sources, as well as changes performed directly via the adaptation component, are also maintained in the central repository change metadata.

4.1.5. Access Layer

In the access layer of the architecture, there are two components that are utilised by users. The goal of the validation, correction and supplementation component is to facilitate the best possible data quality in the central data repository. Various data quality issues may arise during the integration process due to insufficient data quality at the sources, the diverse ways in which the same data are stored in different sources, which lead to problems in identifying the same data records automatically. To resolve these issues, in some cases human interaction is necessary to manually validate data and correct errors that have occurred during the integration. Moreover, our proposed architecture also supports the manual supplementation of the central data repository by data unavailable in the sources. Finally, the integrated data available in the central data repository can be used for analysis and reporting via the reporting component.

The proposed architecture also provides the data sharing functionality to other external systems via the data sharing component included in the access layer. The data available to different external systems is limited by the access permissions set by the administrator via the data sharing component. The data sharing component differs from the wrapper of the internal data source, because it only implements one-way communication to provide the data necessary to external systems.

4.2. Supported Changes

The adaptation component of the adaptive integration architecture is capable of handling changes in data sources as well as new or changed requirements for data that must be loaded to the central data repository. The list of supported changes in data sources is highly dependent on the ability of the corresponding wrappers to track changes and register them in the source change metadata. The examples of source changes that may be supported by the adaptation component follow.

The addition of a completely new data source requires the implementation of a new wrapper, definition of new synchronisation rules and correspondences in the mapping metadata. The addition of new properties of data records in an existing data source must be propagated in the architecture by means of supplementation of the data model of the central data repository and the mapping metadata, and possibly definition of new synchronisation rules. To handle the deletion of a property of a data record or deletion of the whole data record, it is necessary to remove the missing data from the mapping metadata and synchronisation rules, or to mark mappings and synchronisation rules as inactive. As a result, the missing data items will not be filled during the data loading process.

The changes in requirements for data accumulated in the central data repository are also handled by the adaptation component. There are two kinds of changes possible: (1) when a requirement for new data is formulated or (2) existing data are no longer required. In the former case, it is necessary to make changes in the model of the central data repository and specify how the new data are obtained. This may require the addition of a new data source together with a new wrapper, or only supplementation of the mapping metadata and possibly synchronisation rules. In the latter case, no changes to the model are required, however, the reports must be changed according to the updated requirements. Currently, the proposed architecture does not support automatic generation and adaptation of reports, therefore, obsolete requirements must be propagated in reports manually.

5. Case Study

The proposed adaptive integration architecture was implemented to support analysis and evaluation of the scientific activity of members of the University of Latvia involved in research, both employees and students. The architecture integrates information about publications from all accessible data sources that include the library information system ALEPH, LU management information system LUIS, SCOPUS and Web of Science databases.

5.1. Data Model of the Central Data Repository

To maintain publication information and support reporting and analysis, the publication data from multiple sources are linked and stored in the central data repository – LUIS publication repository. The data model of the central repository is depicted in figure 2.

The central class to store bibliographical data about a publication, as well as a number of citations in SCOPUS and Web of Science databases, is Publication. The bibliographical data in this class are entered by the author of the publication or the faculty staff, or they are populated during the data loading process from ALEPH or SCOPUS databases. For each publication indexed by SCOPUS or Web of Science, we also include the corresponding number of citations as well as publication identifiers used in both databases to maintain a link with SCOPUS and Web of Science as data sources.

The information about authors of the publication is reflected by the classes Author and SCOPUS Author. The class Author represents ordered authors of the publication, who are affiliated with LU, as well as with other institutions.

Authors recognised as affiliated with LU are also linked to LU Person class, which stores personal and other information used for different functionality of LUIS. For foreign authors, we store only their name as it appears on the paper. If a publication is indexed by SCOPUS, we also collect author information from it, which includes name, surname, H-index and author ID assigned by SCOPUS, which is used in the author matching process.

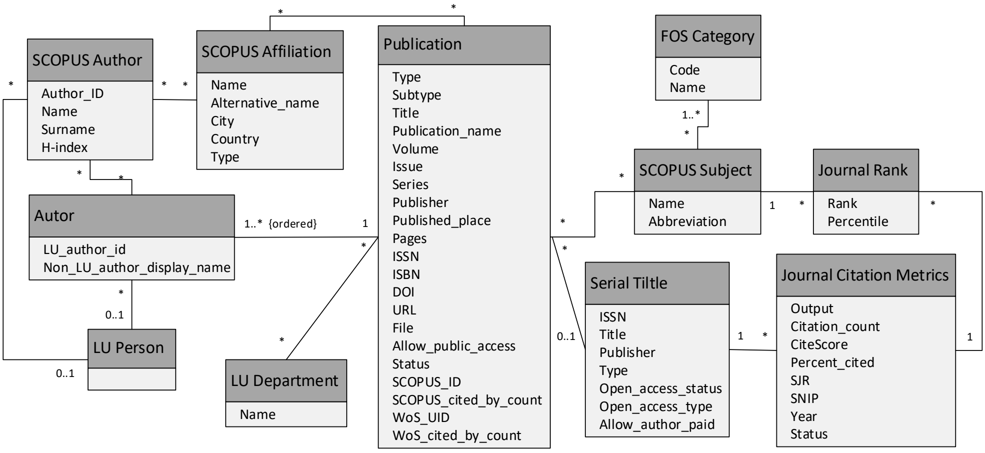

Figure 2. Central repository data model

We also store the information about the affiliation of the publication with the LU department or faculty, which is represented as the class LU Department. This information is obtained automatically from data about the work place of the author and may be corrected by the responsible faculty or library staff. If the publication is indexed by SCOPUS, we also store the information about the institutions the publication authors are affiliated with in the class SCOPUS Affiliation. This information is necessary to analyse connections with co-authors from different institutions or countries.

For the analysis of the quality of publications, we use not only the citation number, but also other citation metrics provided by SCOPUS. The absolute values of such metrics are calculated for journals and serial conference proceedings annually, and their values are represented by the class Journal Citation Metrics. We also collect information about the open access status of the journal or conference proceedings and include it in the class Serial Title. In SCOPUS database, journals are also ranked among other journals that belong to the same subject areas according to their values for CiteScore metric [25], an alternative to WoS Journal Impact Factor. Journal rank information is represented by the class Journal Rank and it is connected with the corresponding subject area (class SCOPUS Subject). For reporting on publications of different OECD categories, we store the correspondence of SCOPUS subject areas and Field of Science and Technology (FOS) categories.

5.2. Scenarios of Research Information System Use

To accumulate the most complete list of publications authored by LU staff and students in the central data repository in LUIS, we gather publication data from different sources, link publications to the correspondent members of LU staff, correct errors and duplicates and provide the consolidated information using a set of reports used by the management of LU, as well as share the information with other consumer systems. In the following section, the scenarios of obtaining and sharing publication data, as well as data flows related to each scenario, are discussed.

5.2.1. Activity of Publication Authors

LU employees, PhD and Master's degree students are able to add information about their co-authored publications to the repository themselves via their LUIS profile, which plays the role of the validation, correction and supplementation component of the adaptive integration architecture. The process where publication data are entered by authors is depicted in figure 3. The LUIS system maintains user profiles for all LU members. Among other information about a user, a profile includes a section devoted to research, which in turn contains a list of an author's publications obtained from the repository. An author can supplement this list by adding newly published articles or articles which were not loaded automatically from various data sources. Before adding them, an author is automatically requested to search for his/her publications in the central data repository and the internal data source – library information system ALEPH, which are not linked with the author's profile, to avoid creation of duplicates. If a desired article is found, it is possible to add it to the profile. In this case, the author does not need to supply any additional information about the article.

If, however, the article is not present in either of the systems, the author has to specify the type and subtype of the publication (for example, journal article, book chapter, book, etc.) and supply bibliographical information. In addition, an author must indicate the status of the publication: published, submitted for publication, developed or under development, attach the publication file (full text or book cover) and indicate whether it can be accessed publicly at the e-resource repository of LU.

After the author has finished entering the publication data, these are prepared by the data loading component and transferred to the library database ALEPH by the corresponding wrapper. To ensure the best possible data quality, the library employees validate publication data, correct any errors where necessary, and approve a publication in the library publication management system. Finally, publication data are synchronised back with the LUIS publication repository and become available for evaluation and reports.

Besides entering new data, LUIS users can unlink erroneously attached publications from their profiles, which were automatically loaded to the repository from other sources or confirm that author matching was performed correctly.

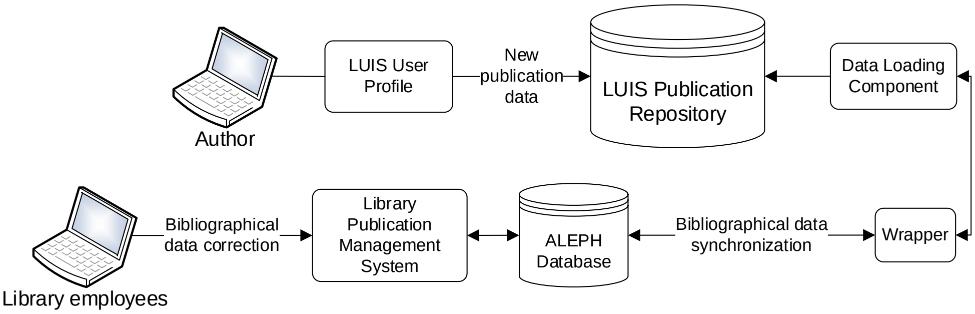
Author
LUIS User Profile
New publication data
LUIS Publication Repository
Data Loading Component
Library employees
Bibliographical data correction
Library Publication Management System
ALEPH Database
Bibliographical data synchronization
Wrapper

Figure 3. Activity of publication authors

5.2.2. Activity of the Faculty Employees

Data about publications authored by the faculty members can also be entered into the LUIS publication repository by specially designated faculty employees (Figure 4). The procedure for adding data is similar to the one that is performed by authors. The differences are that faculty staff can record data about publications authored by other faculty members, correct erroneous links between publications and authors and adjust the list of affiliated LU departments.

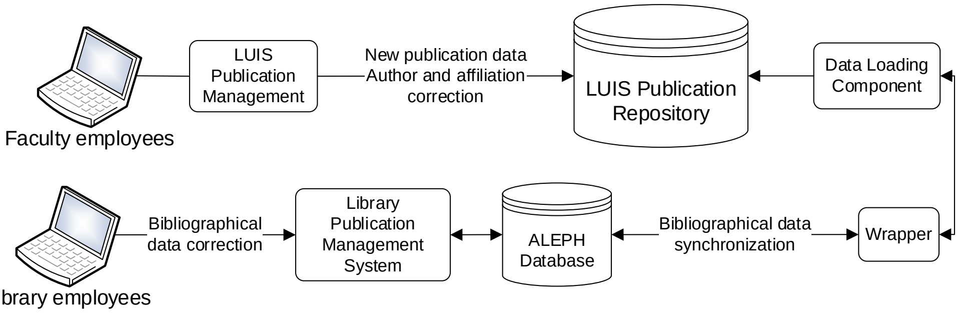
Faculty employees
LUIS Publication Management
New publication data
Author and affiliation correction
LUIS Publication Repository
Data Loading Component
Library employees
Bibliographical data correction
Library Publication Management System
ALEPH Database
Bibliographical data synchronization
Wrapper

Figure 4. Activity of the faculty employees

5.2.3. Data Loading from SCOPUS

There are two external sources of publication data that are used in the data loading process: SCOPUS and Web of Science citation database systems (Figure 5). SCOPUS offers API, which allows searching for publications authored by LU members and extract bibliographical data of such publications, as well as various citation metrics. The extraction and data loading process is run daily. The necessary procedures are launched regularly by the scheduler. Articles that were published during the previous two years are inserted into the repository or updated daily, but all other publications are updated on a weekly basis.

Bibliographical information is extracted from SCOPUS via a wrapper and loaded by the data loading procedures into the repository table, which corresponds to the class Publication of the repository data model. Data about authors (unique author identifier, name, surname, H-index) and publication and author’s affiliation are loaded into tables which correspond to the classes Author, SCOPUS Author and SCOPUS Affiliation of the repository data model. Affiliations are associated with authors as well as with publications directly. In addition to bibliographical and author information, citation metrics are also obtained, that include the current number of citations of individual publications as well as citation metrics obtained for the particular journal or conference proceedings: Source Normalized Impact per Paper (SNIP) [16], the SCImago Journal Rank (SJR) [7], CiteScore [25].

The first step of the SCOPUS data loading process that is executed on any new publication is a *recognition phase*. The main goal of this phase is to identify publications that are already registered in the repository, but that are newly indexed by SCOPUS, to avoid creation of duplicates. The recognition is performed based on the synchronisation rules defined for publications in the metadata repository. The primary criterion used for the recognition is

Document Object Identifier (DOI) which is unique for every publication. The rule specifies that DOI must be precisely the same to consider publication data records coming from different data sources to be the same.

If the matching publication with the same DOI is not found in the publication repository, the search based on the similar title and publication year is performed. To determine the existing publication with the most similar title in the repository, Jaro-Winkler similarity [26] is used because there may be different alternatives of title spelling, as well as data quality issues sometimes being present. Different thresholds for Jaro-Winkler similarity were tested, and experimental evaluation of matching results revealed that the most suitable threshold is 0,93, and currently this coefficient is used in the synchronisation rule to consider titles of publications to be similar.

If the recognition process detects an existing publication in the repository, or if a publication has already been previously updated with SCOPUS data, we update the number of citations for the publication, as well as journal citation metrics and establish a link with a corresponding Scopus record for newly indexed publications by means of filling SCOPUS ID attribute of the Publication class (Figure 2).

If a processed publication is new to the system, a new instance of the Publication class is created with all the bibliographical information obtained from SCOPUS database; publication authors are also represented as instances of Author and SCOPUS Author classes and citation metrics, as well as journal rank data being created or updated if information about a journal has been previously loaded.

In case of a new publication, the author matching process is performed, when for each author affiliated with LU, a corresponding instance of LU Person class is searched for and, if found, is associated with a corresponding instance of the Author class. The primary criterion defined as the synchronisation rule for the author matching process is the SCOPUS author identifier, which allows to uniquely identify authors whose publications have been previously loaded into the publication repository.

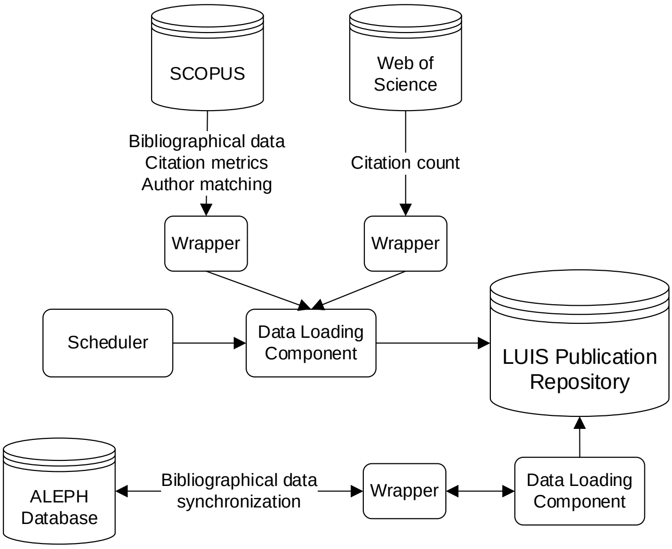

Figure 5. Data loading from external data sources

If author search by identifier is unsuccessful, the matching process uses the secondary criterion, which is a combination of author's name and surname. Author matching by precise name and surname produces insufficient results, because publication authors tend to use different spellings of their names and surnames, that does not always correspond to their full names. Furthermore, a full name can contain special characters of the local language, which may be substituted with English language characters in the publication in a different way. For example, the Latvian letter 'Š' may be substituted with English letter 'S' or with two symbols 'SH'. To solve this data quality issue, the author matching based on names and surnames is performed using Jaro-Winkler similarity between the full name as it appears in the publication and official author's full name, i.e. LU Person instance with the highest Jaro-Winkler similarity coefficient that exceeds a threshold is linked to the publication. We use the same threshold for similarity coefficient 0.93, which was selected based on the experimental matching results and is defined in the corresponding synchronisation rule in the metadata repository.

After a match is found, we also establish an association between the instance of the SCOPUS Author class and the corresponding instance of the LU Person class to use SCOPUS author identifier as the primary criterion for matching future publications.

When a new publication is loaded into the repository, the second process phase – *publication data synchronisation* with the library information system ALEPH is executed. During this phase, publication data are exported to ALEPH, bibliographical information is supplemented and any errors are manually corrected by the library employees, to maintain the best possible data quality, and, finally, the updated data are imported back to the repository.

5.2.4. Data Loading from Web of Science

Another data source used in the integration process is the Web of Science web service (Figure 5). The version of web services available at the University of Latvia does not include journal citation metrics and provides only limited bibliographical information, number of citations, full names and sometimes researcher identifiers of publication authors if they have registered such identifier in the Web of Science database. Since the affiliation of authors is not available, we have discovered that the author matching process for Web of Science data produces too many incorrectly identified authors, therefore, a decision has been made to add new Web of Science publications to the repository manually.

In addition, the integration process also matches publication data obtained from Web of Science with publications already available at the repository. Just as for publications loaded from SCOPUS, the primary synchronisation rule used for matching is DOI and the secondary synchronisation rule is based on title similarity and publication year. The integration process regularly updates the Web of Science number of citations for recognised publications.

5.2.5. Activity of the Library Employees

There are a considerable number of publications that are not indexed by SCOPUS and are authored by LU members, especially in humanities. The information about such publications is necessary to perform accurate evaluation of the scientific activity of the institution.

Therefore, library employees manually add bibliographical information about publications to the library system ALEPH (Figure 6). This is done when a librarian comes across a new publication authored by any LU member in a 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 distributed monthly by Web of Science. When new publication data appear in ALEPH, a synchronisation process is conducted, that integrates bibliographical information into the publication repository to ensure that it always represents an overall view on publication data.

The synchronisation process includes author matching phase. During this phase, for each author of a publication, a corresponding LU person is searched for, using the same synchronisation rule based on the full name similarity as in other matching phases. If the corresponding LU person is found, the publication is attached to his/her LUIS profile. We use the same minimal Jaro-Winkler similarity coefficient 0,93 to consider names to be similar.

In addition to entering new publications to ALEPH, library employees are also responsible for correcting and supplementing bibliographical data for publications added by authors and by faculty staff, and for publications data imported from Scopus database. This activity is performed via LUIS publication management component which plays the role of the validation, correction and supplementation component of the adaptive integration architecture.

Finally, the library employees are also responsible for preparation of reports about the publications authored by LU members. These reports are used for different purposes. For example, they are supplied to the management of the institution for analysis purposes as well as to the Ministry of Education and Science, or special reports are prepared and submitted to other external organisations as one of the criteria used in the evaluation of the activity of the University of Latvia.

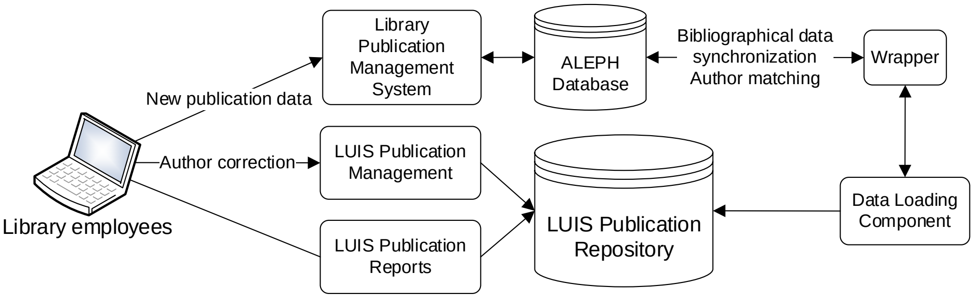
Library Publication Management System
ALEPH Database
Bibliographical data synchronization Author matching
Wrapper
New publication data
Author correction
LUIS Publication Management
LUIS Publication Repository
Data Loading Component
Library employees
LUIS Publication Reports

Figure 6. Activity of the library employees

5.2.6. Data Sharing with Consumer Systems

The access layer of the research information system also incorporates the data sharing component (Figure 7). The publication data export tool provides data available at the central data repository to two consumer systems. The data about all publications, as well as full texts of publications that are allowed to be publicised, are exported from the research information system to the e-resource repository and become available to the public.

Another consumer system included in the architecture is the national science information system that is being developed by the Ministry of Education and Science of Latvia. The system will gather various pieces of information about the scientific activity of institutions in Latvia. This information will be used, for example, to calculate the financing of science, inform the public about the use of public financing and scientific results. The data about different aspects of the scientific activity of the University of Latvia, including publication data, will be automatically imported to the national science information system via the data sharing component of the adaptive integration architecture.

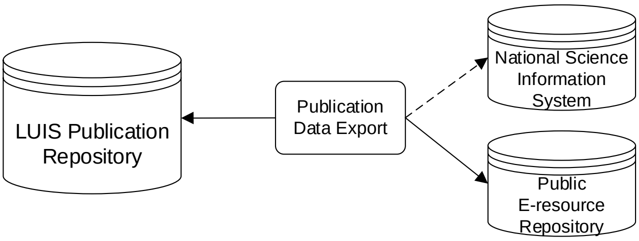
LUIS Publication Repository
Publication Data Export
National Science Information System
Public E-resource Repository

Figure 7. Data sharing with consumer systems

5.3. Adaptation of the Research Information System

Let us consider examples of changes that affected the research information system during its operation and were implemented via the corresponding components of the presented adaptive integration architecture.

5.3.1. Addition of New Data Sources

The first change that was processed in the architecture was the addition of new data sources to SCOPUS and Web of Science citation databases. To handle this change, firstly, we had to implement wrappers manually in the middleware layer which collect data from the new data sources. To store new data, we augmented the data model of the central data repository by tables that correspond to the classes Scopus Author, Scopus Affiliation, Scopus Subject, Serial Title, Journal Citation Metrics and Journal Rank. We also supplemented the existing class Publication with additional attributes: Scopus_ID, Scopus_cited_by_count, WoS_UID and WoS_cited_by_count.

To make it possible to generate data loading procedures, we supplemented the metadata in the repository. We defined mappings between fields obtained from both new data sources and columns of tables of the data model. The majority of columns are filled directly by the fields obtained from the data sources. Only the type and subtype of the publication is mapped from the type used in Scopus to the classifier used at the University of Latvia.

Subsequently, we specified the primary and secondary synchronisation rules. To identify the same publication, firstly, DOI is used and, secondly, the title and publication year are compared with the similarity coefficient of 0.93. The similar synchronisation rules were defined to consider author records the same.

The information about the change and its solution was registered in the source change metadata and central repository change metadata. Using the information in the metadata, the adaptation component was able to adapt the data loading procedures to collect data from the new data sources and incorporate it in the central data repository.

5.3.2. Changes in Schema of a Data Source

Another example of changes that occurred during the operation of the research information system was deletion of formerly available data properties. Previously, it was possible to obtain Impact per Publication (IPP) metric [9] from Scopus data source, which became unavailable in Scopus. We discovered the missing data property automatically when it became impossible to collect IPP by the wrapper.

The wrapper registered the change in the source change metadata and the adaptation component generated potential adaptation options for the integration process, which were saved in the source change metadata. There were two options: (1) to substitute the missing metric with data from another data source and (2) to stop collecting the missing metric. If the former case had been selected, the administrator would have to provide the information about how to obtain the missing data from the alternative data source. In our case we did not have IPP metric available at other data sources, therefore, the latter adaptation option was selected, so the missing metric has been retained for previously loaded publications and has not been loaded for the new ones.

5.4. Publication Data Analysis

The context of the data collection and integration can be described with the total number of publications of LU researchers for the last 30 years, which is equal to 47477. Of these, 7917 publications are indexed by Scopus and 8786 publications are indexed by WoS. The case study data analysis is performed using data that correspond to the LU Faculty of Computing and the timeframe that was chosen for research results evaluation was 2012 – 2017. We have already performed an initial evaluation of research performance by means of quantitative metrics [18]. To provide context for the further data analysis, some figures, e.g. publication count and Scopus publications, should be mentioned. The total number of publications decreased at the faculty from 101 in 2012 to 61 in 2017, but the number of Scopus publications grew. In 2012, there were 42 publications indexed by Scopus, but in 2016 there were 51. The data for 2017 are not complete at the time of preparation of this paper, so the current number of publications indexed by Scopus for 2017 (37 publications) will still be supplemented.

The goal of the data analysis was to define metrics based on the data attributes provided by the data model of the integrated publication information system, and with the goal of evaluating the quality of the publications, to find out the positive trends, as well as the problems with quality.

The quality aspects of a publication can be indirectly described with the source quality characteristics, e.g. journal quartiles that are computed from the citation count of all journal publications, because we can presume that the journal with the highest quartile Q1 will accept the best publications. Another group of quality indicators directly describe the quality of the publications and are computed from citation counts of publications. Further in this section, different analysis scenarios for research output evaluation with quality metrics that can be implemented with the new publication module and data integration infrastructure are described.

For the 1st analysis scenario the following research question was formulated: “How many faculty publications in Scopus are published in sources with and without computed quartiles?” Later more detailed analysis was performed to find out how the publication count is divided among quartiles.

The results are shown in figures 8 a) and b). The results show an unsatisfactory trend for the faculty, that in 2015 the number of publications in sources without quartiles increased. The detailed analysis showed that in all years, except the last two - 2016 and 2017, the biggest number of publications belongs to quartile Q3. Regarding the excellence, the number of publications with Q1 has increased in the last years.

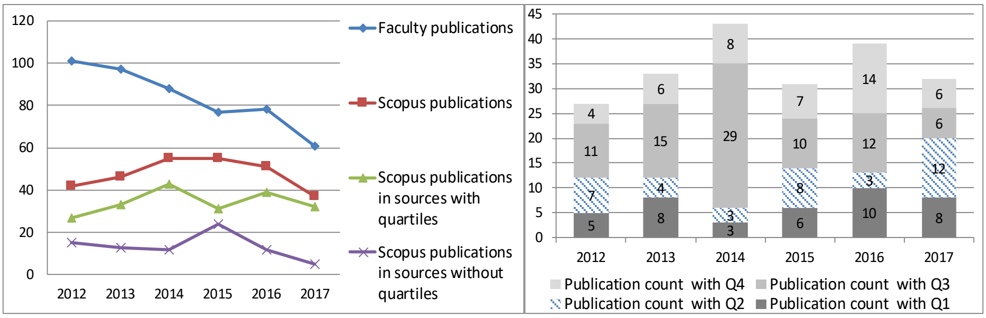

Figure 8. Publication count in Scopus sources a) with and without quartiles b) detailed count by quartiles

For the 2nd analysis scenario, the following research question was formulated: “How many faculty publications in Scopus are not cited comparing with all publications and the publications in sources with computed quartiles?” Figure 9 shows the trend that the proportion of uncited publications remains unchanged in sources with computed quartiles, but is growing among all Scopus publications.

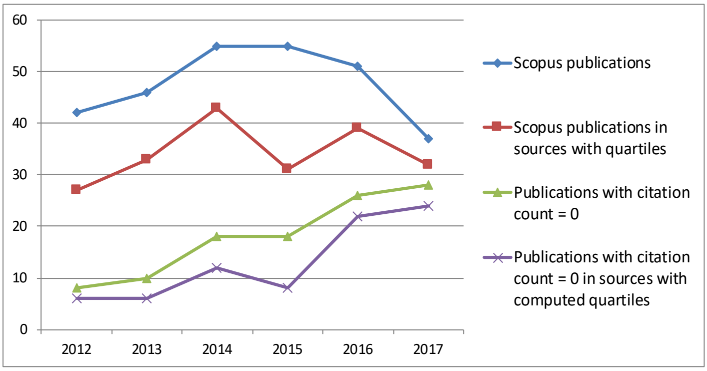

Figure 9. Publications that are not cited

For the 3rd analysis scenario, the research question was formulated: “How many citation counts in Scopus sources are there with quartiles and without computed quartiles?” Figure 10 a) shows the trend that the citation count in Scopus sources without computed quartiles is decreasing. More detailed analysis (see figure 10 b) shows a significant citation count of Q3 publications, however this can be explained with a greater amount of Q3 publications among all others.

These results can help decisions to be made at each individual researcher’s level to try publishing their works in sources with one quartile higher, but for the faculty, the shift from sources with Q3 to Q2 may be the most promising and realistic.

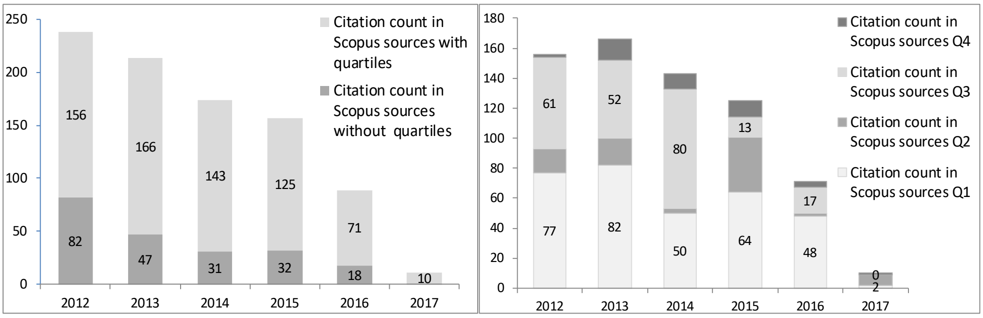

Figure 10. Citation count in Scopus sources a) with and without quartiles b) detailed citation count by quartiles

6. Conclusions and Future Work

The main contribution of this paper is the adaptive integration architecture capable of processing changes in data sources used during the integration process, as well as evolving requirements for data. We described the components of the architecture, the necessary metadata and gave examples of changes that are supported by the architecture together with their implementations within the architecture. We demonstrated the practical application of the proposed architecture that implements different flows of data and integrates them into one consistent system for research output evaluation. The research information system is based on the idea of ensuring data quality, control, and also the integration process transparency, involving publication authors in different roles – as information providers or approvers.

The central part of the architecture is the central data repository that is implemented as a relational database. However, traditional relational database solutions may become unable to handle large volumes of data incoming from data sources. Therefore, we plan to extend the integration architecture to employ big data technologies for the central data repository. Another direction for future work involves automatic or semi-automatic adaptation of reports on data gathered during the integration process and available at the central data repository.

Acknowledgements

This work has been partly supported by the European Regional Development Fund (ERDF) project No. 1.1.1.2./VIAA/1/16/057 “Handling Adaptation of Big Data Warehouse” and by University of Latvia project No. AAP2016/B032 “Innovative information technologies".

References

[1] Aagaard K., Bloch C., Schneider J.W., Impacts of performance-based research funding systems: the case of the Norwegian Publication Indicator. *Research Evaluation* 24, 2, 2015, 106–117.

[2] Cabinet Regulation No. 1316 Regulations regarding calculation and assignment of grant-based funding for research institutions, https://likumi.lv/doc.php?id=262508

[3] Chen Z., Wu D., Lu J., Chen Y., Metadata-based information resource integration for research management, *Procedia Computer Science*, 17, 2013, 54-61.

[4] Di Tria F., Lefons E., Tangorra F., Academic data warehouse design using a hybrid methodology, *Computer Science and Information Systems*, 12, 1, 2015, 135-160.

[5] DSpace-CRIS Home. https://wiki.duraspace.org/display/DSPACECRIS/DSpace-CRIS+Home

[6] Galimberti P., Mornati S., The Italian model of distributed research information management systems: a case study, *Procedia Computer Science,* 106, 2017, 183-195.

[7] Golfarelli M., Lechtenbörger J., Rizzi S., Vossen G., Schema versioning in data warehouses: Enabling cross-version querying via schema augmentation, *Data & Knowledge Engineering,* 59, 2, 2006, 435-459.

[8] Gonzalez-Pereira B., Guerrero-Bote V.P., Moya-Anegon F., A new approach to the metric of journals’ scientific prestige: The SJR indicator, *Journal of Informetrics,* 4, 3, 2010, 379-391.

[9] Hardcastle J., New journal citation metric – Impact per Publication, 2014, http://editorresources.taylorandfrancisgroup.com/new-journal-citation-metric-impact-per-publication/

[10] Hicks D., Wouters P., Waltman L., De Rijcke S., Rafols I., The Leiden Manifesto for research metrics, *Nature,* 520, 7548, 2015, 429-431.

[11] Inmon W.H., *Building the Data Warehouse,* 3rd edition, Wiley Computer Publishing, 2002.

[12] The International Organisation for Research Information, http://eurocris.org/cerif/main-features-cerif

[13] Jörg B., CERIF: The common European research information format model, *Data Science Journal*, 9, 2010, CRIS24-CRIS31.

[14] Kosten J., A classification of the use of research indicators, *Scientometrics,* 108, 1, 2016, 457-464.

[15] Kulczycki E., Korzeń M., Korytkowski P., Toward an excellence-based research funding system: Evidence from Poland, *Journal of Informetrics,* 11, 1, 2017, 282-298.

[16] Moed H.F., Measuring contextual citation impact of scientific journals, *Journal of Informetrics*, 4, 3, 2010, 265-277.

[17] Nadal S., Romero O., Abelló A., Vassiliadis P., Vansummeren S., An integration-oriented ontology to govern evolution in big data ecosystems, in: *Proceedings of the Workshops of the EDBT/ICDT 2017 Joint Conference (EDBT/ICDT 2017)*, Venice, Italy, 2017.

[18] Niedrite L., Solodovnikova D., University IS Architecture for the Research Evaluation Support, in: *Proceedings of 11th International Scientific and Practical Conference "Environment. Technology. Resources"*, Rezekne Academy of Technologies, Rezekne, 2017, 112-117.

[19] Niedrite L., Solodovnikova D., Niedritis A., Publication Data Integration as a Tool for Excellence-Based Research Analysis at the University of Latvia, in: M. Kirikova, K. Nørvåg, G. Papadopoulos, J. Gamper, R. Wrembel, J. Darmont, S. Rizzi (eds.), *New Trends in Databases and Information Systems. ADBIS 2017. Communications in Computer and Information Science*, 767, Springer, Berlin, 2017, 125-136.

[20] Nikolić S., Penca V., Ivanović D., Surla D., Konjović Z., Storing of Bibliometric Indicators in CERIF Data Model, in: *International Conference on Internet Society Technology*, 2015.

[21] Parmenter D., *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs*, Second Edition, Jon Wiley & Sons, Inc., 2010.

[22] Quix C., Matthias J., Information integration in research information systems, *Procedia Computer Science*, 33, 2014, 18-24.

[23] Sivertsen G., Data integration in Scandinavia, *Scientometrics*, 106, 2, 2016, 849-855.

[24] Solodovnikova, D., Data Warehouse Evolution Framework, in: *Proceedings of Spring Young Researcher's Colloquium on Database and Information Systems*, Moscow, Russia, 2007, 4.

[25] Teixeira da Silva J.A., Memon A.R., CiteScore: A cite for sore eyes, or a valuable, transparent metric?, *Scientometrics*, 111, 1, 2017, 553-556.

[26] Winkler W., The state of record linkage and current research problems, Technical report, Statistics of Income Division, US Census Bureau, 1999.

[27] Winter R., Strauch B., A method for demand-driven information requirements analysis in data warehousing projects, in: *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Washington, DC, 2003, 231.1.

This paper is a revised and extended version of the work originally presented at ADBIS Workshop: Data-Driven Approaches for Analyzing and Managing Scholarly Data: Systems, Methods and Applications (AMSD), Cyprus, September 2017
Received 11.01.2018, Accepted 21.03.2018