Data, Information and Technology services for Research and Management of Science

Gabriel Juhás*‡, Ludovít Molnár†‡, Miriam Ondrisová§ and Ana Juhásová¶
*Faculty of Electrical Engineering and Information Technology
Slovak University of Technology, Bratislava, Slovakia
gabriel.juhas@stuba.sk
†Faculty of Informatics and Information Technologies
Slovak University of Technology, Bratislava, Slovakia
ludovit.molnar@stuba.sk
‡Interes.Institute s.r.o., Bratislava, Slovakia
interes@interes.institute
§Faculty of Arts
Comenius University in Bratislava, Slovakia
miriam.ondrisova@uniba.sk
¶ BIREGAL s.r.o., Bratislava, Slovakia
ana.juhasova@biregal.sk

***Abstract*—Science in inherently connected with data. Data are a crucial entity during whole life cycle of research activities: they are inevitable as an input to research, they are produced as a research output - which serve as an input for further research, they are present in the analyzing and evaluation of the research etc. Nowadays, data science offers new possibilities of storing, sharing, analyzing and processing huge amount of data and to generate, derive new, more accurate and more complex information about all research areas. In the past and still today sharing of scientific data was a real bottleneck. Concept of open data and new IT technology development such as cloud, next generation networks etc. offer possible solutions of this problem. In this contribution we discuss and suggest an architecture for implementation of an ecosystem of storing, sharing, analyzing and processing of research data including data for management of science. The architecture supposes sharing of data not only using a central storage, but it is understood as a bridge or an "open market" of scientific and research data, dealing with issues such as reusability, durability, possible use in the future - storing a snapshot of today world for future generation of researchers in a decentralized way. A crucial issue, which is enabled by new technologies is quality and high accessibility of data, which enables to speed up the research process. The architecture also covers social and economic concepts to ensure the sustainability of the whole research ecosystem, which we consider as a necessary condition of the realization of open data concept. As a use case, we will show an outlook of the ecosystem for research and management of science in Slovakia. The use case includes further development of systems for storing of research and scientific data running at CVTI, Slovak Center of Scientific and Technical Information.**

I. Introduction

Science can principally be understood as a process of information and knowledge mining from data a and previously gathered information and knowledge. Thus, data, information and knowledge create an input of the process of scientific exploration. They are a necessary condition to perform research activities. A lot of today research results as well as recent scientific development was enabled by availability of data from the different periods of past. Well known examples includes Darwins's evolution theory, geological history of Earth, which together enable to reconstruct the history of life on Earth. The data samples of the same kind from different points of the time axis are often crucial for the particular research. A lot of the data from the past used in today research were not recognized as useful in the time of their origin. There is no surprise that data have been recognized as the commodity with a multiple utilization value in the knowledge based society. The same data can be used in different research areas and with different purpose, even those which are not known by the data recording. From the point of view of the knowledge based society data belongs to the most important heritage of the human kind, because of science as an accelerator of the knowledge based society. It follows the commitment of the society to devote the data attention corresponding to their importance, including suitable infrastructure for storing, sharing, analyzing and processing all the data, which are used or can be used in any future research. As a side effect, the data can be used for more accurate evaluation of research and more effective distribution of resources.

The necessary infrastructure, single processes of data retrieval, platforms for storing, sharing and processing data, processes of research evaluation, schemes for distributing resources for research etc. [1] cannot be treated separately, but should be investigated as a *ecosystem*.

Such ecosystem must deal with different aspects of data, including size, quality, availability (social aspects such a as type of access, e.g. open access) accessibility (including technical aspects such as speed etc.), format of data, durability, etc.

In this paper we try to outline a set of requirements and an architecture of such ecosystem, which we call Science Data

Ecosystem, shortly SDE. We also present illustrative examples of scenarios and use cases. As a use case, we will show an outlook of the ecosystem for research and management of science in Slovakia. The use case includes further development of systems for storing of research and scientific data running at CVTI, Slovak Center of Scientific and Technical Information [2].

II. Scientific Data Ecosystem

The proposed ecosystem has to work with possibly huge amount of data. An ambition of SDE is not the gather data itself, but it has to enable any researcher to insert the scientific data, information and knowledge, collected during the research, to SDE. The architecture of SDE has to be cloud-based. Here, by cloud-based architecture we mean that a user can insert data themselves to an existing network of SDE nodes or by adding a node running by the user with the stored data to the network of SDE. Thus, SDE includes as a set of possibly distributed nodes storing data, creating together SDE repositories. In order to add a node to a distributed cloud-based SDE, requirements on durable accessibility and security of the node has to be satisfied. The processes defining how to add or remove a node from/to the repository should be defined, including licensing of the data, as well as processes specifying how to deal with data by removing a node (Can they be copied to another node, which is not run by the original node?) etc.

At the same time SDE has to enable to any researcher to access the data inserted to repositories by other researchers. Here, the concept of the open access plays a crucial role. The effective concepts must guarantee sustainability of the system. The suggested concept of SDE for sustainability should satisfy:

- if data, information and knowledge stored in SDE is used for a non-profit scientific research in order to achieve the data, information and knowledge that are proposed to be inserted to the SDE, then the data, information and knowledge stored in SDE are free open accessible;
- if data, information and knowledge stored in SDE is used in order to achieve the data, information and knowledge that are not proposed to be inserted to the SDE, then a fee should be paid to cover SDE operation and further development;
- if data, information and knowledge stored in SDE is used for a profit, then the data, information and knowledge stored in SDE are open accessible, but a part of the revenue should be return back to cover SDE operation and further development;

Thus, behind the main aim, which is to store the data, SDE is about data sharing. In the past and even today sharing of data is a real bottleneck. Open data concept and technology development such as cloud, next generation networks etc. provide necessary tools in order to build an infrastructure for data sharing. Data sharing is not only about repositories, but it should be understood as an ”open market” of scientific and research data, which enables reusability and possible use in the future, storing a snapshot of today world for future generation of researchers.

In order to achieve the first two main goals, to store and to share the data, one has to build data *repositories* and *indexes* of the data stored in the repositories. Although technically repositories and indexes can create a unit, conceptually we propose to make a logical difference between them, with the repositories understood as a tool to store the data and the data indexes understood as a tool to store references to data and metadata about the data. The third main goal of SDE is to offer researchers the tools for data *analytics*. Over this three main building blocks - *repositories*, *indexes* and *analytics* one can build the *services*.

One of the first information systems, which can be understood as a design pattern for SDE is a web search engine, such as Google. A web search engine can actually be understood as a Web Pages Data Ecosytem over Web. It stores copies of crawled web pages in its own repositories, but the servers, where original web pages actually are stored can be understood as external distributed repositories too. A search engine builds an index including references to crawled web pages, web graph including references between crawled web pages. Analytics include functions which compute the rank of each web page, it may contain also other kinds of functions computing for example number of clicks from the search engine page to a web page, ratio between number of appearances and clicks for a web page etc. The core of a search engine is an index over expressions (words, phrases) found in crawled web pages. Finally, the main service over a search engine, namely the search itself is build up over the web page repositories, indexes and analytics, providing for given input search phrase the list of web pages containing the searched phrase, ordered by importance or relevance computed by analytics.

III. Repositories of Scientific Data

There is a lot of literature about data repositories [3], [4], [5], [6], [7], [8], with many different definitions, some of them including technical details, implementation details etc. For the purpose of this paper, we understand *scientific data repositories as possibly distributed information systems which store the scientific data*. The purpose of the data repositories is on one hand to archive the data and on the other hand to enable access to data in order to share them, to analyze them etc. We propose to distinguish two logical parts of repositories, namely *archives* and *operational storages*.

Physically, archives can have different layers or be of a different technical nature. Nevertheless, their common role is to store the data for a long term period and to satisfy the durability. The function of the archives is to restore the data after a physical medium is out of date, to transform the data into an actual data format if the data format is out of date, to keep and actualize the software and hardware tools needed to write and read data etc.

The role of operational storages is to keep the data, which are actually manipulated because of some service request. Here the services are the basic services such as to insert the new

data returning the reference for indexing, to offer data to a user based on a request using an index service, to analyze the data using analytics and further services build over the basic services.

For example, if a service requesting the insert of new data is called, then the data should be put into operational storage of the repository, the service is called that write the data into archives and return the reference to the data in the archive to the operational storage which is over the archives. The reference is stored in an index over the repository.

If a service requesting a data searched via an index by a user is obtained, then the operational storage is calling its own service reading the data from appropriate archive and loading the data to the operational storage. The data from operational storage are then accessible to the user for further actions (analytics, copying etc.).

The main requirement to archives is to satisfy durability of data and reliability. On the other hand, the main requirement to operational storages is to satisfy fast access. Together, the repositories have to satisfy durability and availability of data. With the huge amount of data expected, not all the data have to necessarily be in an operational storage all the time, but there should be a service, that all the data can be loaded to the operational storage upon request in reasonable time and stored for a requested period of time necessary to perform the actions over the data.

There are many other important issues, which have to be solved by repositories, such as data formats, degree of redundancy to satisfy reliability etc., but go beyond the scope of this paper because of the page limitation.

IV. Indexes and Metadata

Indexes and metadata are crucial for the operation of SDE. We understand indexes as information systems, which store information about data stored in repositories, mainly reference to the location (in which repository and where in the repository are the stored data). Indexes are the key component for searching and finding the data. Once the indexes are lost, the data still exist in repositories but cannot be easily found.

The main problem of indexing scientific data is to choose what should be the information about the data, which is stored in the index as indices[1]. We call this information stored in an index as *metadata*. Briefly, metadata contain information about stored data, which mostly in a structured form describe the content of data. Very simplified, the metadata form the search phrases (search filters), according which one can search for the data.

There has to be made a conceptual distinction between searching *for* data according to the phrases over the *values of metadata* and between searching *in* data according to the phrases over the *values of data* itself.

The main operation function build over an index is to return the references to the data according to a search query over the values of metadata. However, intelligent searching should also support search queries over the values of data too.

Let us illustrate the functionality of indexes on a typical and well known example of an index, namely the index of scientific publications. Imagine one has build repositories of scientific publications - a digital library. An index build over such a repository should at least contain the metadata about the publications stored in the repository, such as the title, authors, publisher, year of publishing, number of pages and key words. The index should not only be a list of metadata about publications, with the possibility to filter the entries in the list according to search queries over metadata, but it should also contain the references (links, pointers) to the places, where the publications are physically stored. After a search filter is returning the list of the publications satisfying the search query, it should be possible to request the publication itself (for example via a link for downloading). Moreover, an intelligent search should also enable to search not only over the publications over metadata, but should enable also searching any phrase in the fulltext of publications. Thus, for example, it should not only return all the publications of some author, but it also should be able to return those publication of an author (metadata value) that contain in fulltext a phrase (data value not contained in metadata).

In the SDE of the future, such repositories and intelligent indexes should be build for any kind of data in any format, including scientific data from experiments with metadata about the methods, how the data are gathered, types of measurment devices etc., with possible intelligent search over metadata values and data values. An example may be an indexed repository of unstructured data such as videos or pictures with searching not only according to metadata, but also according to scenes or persons appearing in the videos, which are able to return e.g. videos or pictures containing a specified object. The methods for such intelligent search can already be understood as a subject of analytics.

In Figure 1 and Figure 2 we illustrate an example of a simplified workflow process of the SDE indexed repository functionality as a Petri net. By internal storage/archive we mean the storage/archive of a user, which produces the data, while by external storage/archive we mean the storage/archive which does not belong to data producer. In order to make this paper self-contained, in the following paragraphs we briefly recall the basic definition of Petri nets, which are one of the most used tools for modeling workflow processes [9], [10], [11]. Using Petri nets one can easily formalize functional requirements on SDE on an abstract and yet formal level.

V. Petri nets

Let $\mathbb{N}$ denote the *nonnegative integers*.

Definition 1 (Petri net)**.** *A* Petri net N *is a quadruple* $N = (P, T, I, O, m_0)$*, where* P *is a set of* places*,* T *is a set of* transitions *such that* $P \cap T = \emptyset$*,* $I : P \times T \to \mathbb{N}$ *is an* input function*,* $O : P \times T \to \mathbb{N}$ *is an* output function*, and* $m_0 : P \to \mathbb{N}$ *is an* initial marking *of* N*.*

[1]In order to make a clear distinction, we use plural *indexes* to denote several index information systems and plural *indices* to denote several entries in an index

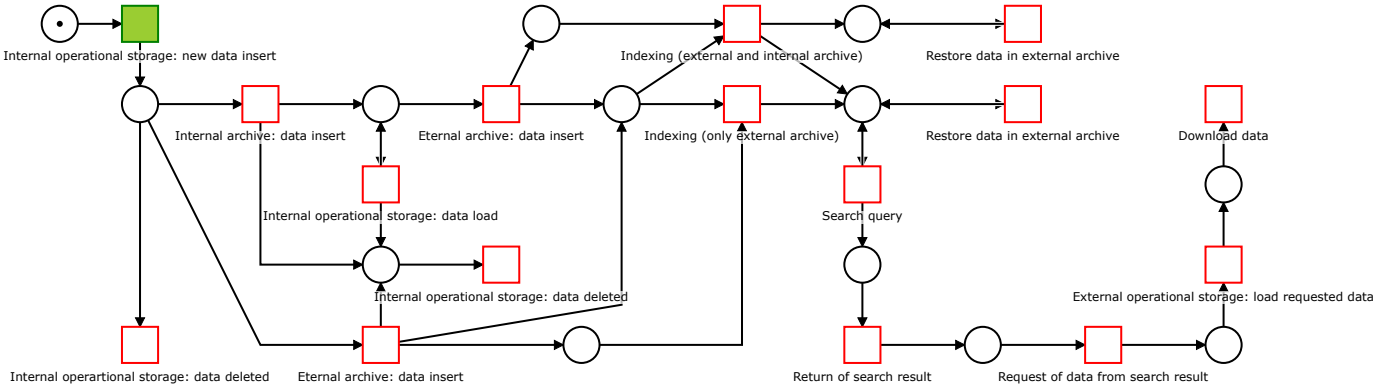

Fig. 1. A Petri net model of an example of a simplified process of the SDE indexed repository functionality.

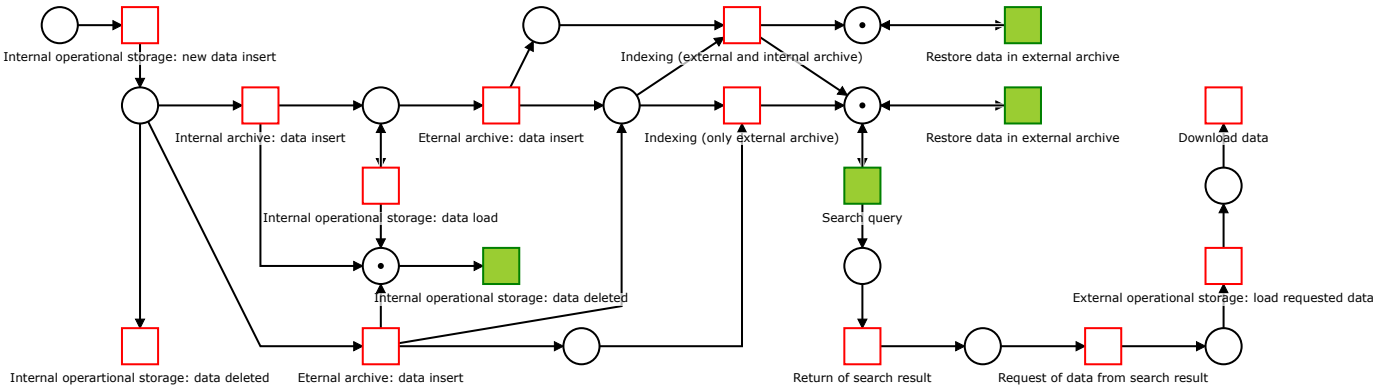

Fig. 2. A Petri net model of an example of a simplified process of the SDE indexed repository functionality after storing data in both internal and external archive and after indexing.

A transition $t \in \mathbb{N}$ is *enabled to fire in a marking* $m : P \to \mathbb{N}$ if $\forall p \in P : \ m(p) \geq I(p,t)$. Firing of an enabled t in a marking m causes the chage of the marking m to the marking m' satisfying $m'(p) = m(p) - I(p,t) + O(t,p)$ for each place $p \in P$.

Places are graphically expressed by circles, transitions by rectangles, non-zero elements of input function by integer weighted arcs from corresponding places to transitions and non-zero elements of output function by integer weighted arcs from corresponding transitions to places. Markings are expressed by corresponding number of black tokens inside of places. Those transitions, which are enabled to fire are filled in Figures with green color, while transitions not enabled to fire are drawn by red color and unfilled.

Intuitivelly, firing of a transition consumes the number of tokens given by the input functions from places and produces the number of tokens given by the itput function in places. A transition can fire, if a marking contains enough tokens to be consumed.

VI. Tools for Analytics

Let us recall the example of a search engine over web, which we use as a design pattern for SDE. The main factor of a success of search engines, such as for example Google, was that it returned the most relevant web pages containing the search phrase as the results. In the case of Google, behind this success was the well known page rank algorithm [12], that associated a rank to each crawled web page. Roughly speaking, according to [12] the page rank of a web page, say a was a number given by the sum of a dumping factor (representing a probability that one starts to browse at the web page a) and contributions from all web pages, which contain a link to the web page a. A web page with rank r, which contains a link to n other web pages, contributes to the rank of those web pages by r divided by n which can be interpeted in such a way that a web page propagates its rank, i.e. its *importance*, by dividing its rank by the number of referenced web pages [12]. For the purpose of this paper we understand algorithms such as the page rank from [12] to be functions or tools for analytics. Thus, in analogy to microservice oriented architecture the tools for analytics are losely coupled microservices (such as the page rank algorithm) over which services (such as the rank based search) are build.

More common and more obvious methods and tools for analytics are standard methods of statistics, different methods of data classification, different methods based on artificial intelligence, machine learning, deep machine learning / neural network based methods, methods of scientific computing, which can offer microservices over which services such as intelligent search or any other service leading to knowledge

mining are build.

Naturally, the most obvious use of tools for analytics is to get new research results over scientific data in repositories creating new knowledge (for example proving a correlation between data).

However, tools for analytics can also be used to evaluate research results. Rather than going into details and a complete list of the methods and tools for analytics, let us discuss one more example, which will illustrate how even very simple analytics can bring more accurate results in evaluation of a publication impact. As an example, consider the current system, how the financial resources for scientific publications are granted to public universities in Slovak republic [13]: a university is granted by *the same* amount of financial resources for a paper in a journal indexed in Current Content Connect Database [14] in computer science, metallurgy and ecology. This is still a relict of the past, where not enough data were available to state the *price* of such a paper by relation to average, median or maximal number of papers published per researcher more precisely, distinguishing between fields and even sub-fields.

Another example is evaluation based on impact factor of a journal, which should be used to predict the expected number of citations for a paper. However, after a period of time, a normalized value according to the year of publication and the field should give a more precise impact of the paper [15]. One can for example apply page rank algorithm on a graph with published papers instead of web pages and cited references instead of links between web pages [16]. Other attributes, such as number of views and number of downloads can be used to evaluate the impact of publications as well.

By these examples we want to demonstrate, that having enough data (for example about publications) and good tools for analytics, one can get more precise and more accurate results even in evaluation of science.

VII. Service Layer

The service layer of SDE should enable to assembly services over data stored in repositories, over metadata stored in indexes and over tools of analytics. The proposed architecture of SDE supposes that functionality of repositories, indexes and tools of analytics is implemented according to microservice oriented architecture, providing that these microservices can be composed and causally ordered to workflow processes resulting in services which can include complex reports over data and metadata. It should include the specification and modelling language, which will enable to specify workflow, data and microservices which should be used and roles and user management properties for the composed service. The specification language should also enable to specify logical appearance of user interface.

The service layer with the help of tools for analytics, indexes and repositories should enable an intelligent search. Such an intelligent search should take into account not only what is searched but also who is searching and for what purpose. Thus, for different needs and the same searched phrase the results might be different. It depends whether the same keywords are searched by a user writing his diploma thesis or by a user writing a new original research paper. For the first user, the most relevant results could be the survey papers, while for the second user these could be the most recent results from the subject containing the keywords. It means that quality of data, for example quality of scientific publications depends on a use case. We can distinguish different dimensions of quality, such as originality, state of the art, educational dimension, pioneering (founding) dimension. Clearly, different methods of data science and different systems can be used to evaluate the quality of publications, such as antiplagiarism system to detect duplicate content, citation indexes to detect the impact of the paper etc.

VIII. Use Case

As a use case, we will show an outlook of the ecosystem for research and management of science in Slovakia. The use case includes further development of systems for storing of research and scientific data running at CVTI, Slovak Center of Scientific and Technical Information [2]. The comprehensive description of the state-of-the-art at CVTI can be found in the paper [17] in this proceedings.

The further development of the existing systems should start with the analysis w.r.t. the SDE architecture resulting in the categorization of single systems. For each system a detailed plan of development has to be made in order to extend the system according to the processes and functionality stated by SDE architecture. The integration layer should be added. The tools for analytics, which are not yet the part of the CVTI infrastructure should be build according to microservice oriented architecture. Use cases have to be defined in detailed level. Based on the use cases, the concrete services for the service layer should be assembled.

Now, let us illustrate these steps on particular systems operated by CVTI.

The key information systems at CVTI are/will be:

- Repositories (SCIDAP - Scientific Data Analysis Platform)
- Open Access Publication Platform (planned)
- Modul for Management of Research Data SVD (planned)
- Current Research Information System SK CRIS
- Analytics for Evaluation of Science (planned)

From the SDE point of view, SCIDAP and its further development as planned by CVTI fulfills the repository definition, including possible integration of institutional repositories not operated by CVTI.

Open Access Publication Platform should mainly fulfill a definition of an index linked with published journal and books. It will enable to store links between the publications and scientific data. It also has a functionality of a tool for analytics and a service layer as it should enable to insert data (to publish publications), and to search.

Modul for Management of Research Data SVD should be a modul for planning of research projects and for inserting the research project data into SCIDAP repositories. It should also

have the function of an index over the scientific data from projects in SCIDAP (including the index of datasets).

Current Research Information System SK CRIS should serve for registration and searching of researchers, institutions and projects itself in order to create research teams and to get an overview of research projects. It mainly fulfil a definition of an index.

Analytics Modul for Evaluation of Science should offer tools for analytics and a service layer for evaluation of results of research and scientific institutions. It should serve accreditation committee, evaluators, grant agencies, evaluated organizations, governing bodies, but also private public partnership and even industry and public (e.g. students) to evaluate the research results for different purposes, including distribution of financial resources to scientific institutions and universities (from the state, grant agencies, by endowment etc.). It will be integrated with SCIDAP repositories and will enable scientometric analysis and will offer different rankings of scientific institutions and universities. These rankings and evaluations should be available online. Thus, the complicated accreditation of universities from the past should be replaced by an automatic and continuous evaluation and ranking by this module.

There are other information systems, operated by CVTI, which should be extended and integrated into SDE. Most important of them are:

- The Central Registry of Publications Activity CREPC
- The Central Registry of Artistic Activity CREUC
- The Central Registry of Theses and Dissertations CRZP
- System for plagiarism detection ANTIPLAG.

The Central Registry of Publications Activity CREPC and the Central Registry of Artistic Activity CREUC are nowadays lists of publications, but do not contain the links to all publications from these lists. Thus, they have not full functionality of indexes. They should be extended to full indexes and connected with SCIDAP and repositories of publishers originally publishing the publications.

The Central Registry of Theses and Dissertations CRZP (CRZP) is a repository with an index offering a tool for searching. It will be integrated with SCIDAP.

System for plagiarism detection ANTIPLAG will be integrated with SCIDAP repositories. It has a functionality of a tool for analytics.

Another functionally important modules of information systems operated by CVTI are: Integration layer, Presentation platform (portal) and Integrated System of Services. They together create the main part of the service layer of SDE. Their functionality should include single sign-on access of users, intelligent searching using a common user interface, central user management, autentification and access control of users of SDE.

As it was already mentioned, the whole architecture of SDE operated by CVTI will be based on microservice oriented architecture.

CVTI operated SDE contains also Databases of third parties EIZ, which contain third party indexes such as Scopus, Electronic Resources Management System ERMs, which contain a database of prepaid information resources of third parties, and a search engine PRIMO. The search engine should be extended in order to offer not only fulltext search, but an intelligent search over as many different types of scientific data as can be stored in repositories.

Conclusion

This paper discusses a concept of an architecture for an ecosystem for collecting, storing, sharing and analyzing scientific data in order to get new information and knowledge in research as well as more accurate information for evaluation and management of science. The concept is illustrated on simple examples and the application of the ecosystem architecture is outlined for information systems operated by Slovak Center of Scientific and Technical Information CVTI or information systems, which are planned to be build by Slovak Center of Scientific and Technical Information CVTI.

References

[1] R. C. Amorin, "A comparative study of platforms for research data management: Interoperability, metadata capabilities and integration potential." *WorldCIST*, vol. 1, pp. 101–111, 2015.

[2] "Slovak center of scientific and technical information: Support of science," *http://www.cvtisr.sk/en/support-of-science.html?page_id=788*.

[3] M. Armstrong, "Institutional repository management models that support faculty research dissemination." *OCLC Systems & Services*, vol. 30, no. 1, pp. 43–51, 2017.

[4] J. Bankier, "Institutional repository software comparison." vol. 33, 2014.

[5] C. S. Burns, A. Lana, and J. M. Budd, "Institutional repositories: Exploration of costs and value." *D-Lib Magazine*, vol. 19, no. 1/2, 2013.

[6] E. Fay, "Repository software comparison: Building digital library infrastructure at lse." *Ariadne*, vol. 64, 2010.

[7] J. Giesecke, "Institutional repositories: Keys to success." *Journal of Library Administration*, vol. 51, pp. 529–542, 2011.

[8] A. Swan, "The business of digital repositories." in *A DRIVER's Guide to European Repositories*. Amsterdam: Amsterdam University Press, 2007.

[9] J. Desel and G. Juhás, "What is a petri net?." in *Unifying Petri Nets*, ser. Lecture Notes in Computer Science, H. Ehrig; G. Juhás; J. Padberg; G. Rozenberg, Ed., vol. 2128. Springer, 2001, pp. 1–25.

[10] H. Ehrig; G. Juhás; J. Padberg; G. Rozenberg, Ed., *Unifying Petri Nets, Advances in Petri Nets*, ser. Lecture Notes in Computer Science, vol. 2128. Springer, 2001.

[11] R. Lorenz, J. Desel, and G. Juhás, "Models from scenarios," in *Transactions on Petri Nets and Other Models of Concurrency VII*. Springer Berlin Heidelberg, 2013, pp. 314–371.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[13] "Metodika rozpisu dotacii zo statneho rozpoctu verejnym vysokym skolam na rok 2017," *https://www.minedu.sk/data/att/10876.pdf*.

[14] "Current contents connect," *http://wokinfo.com/products_tools/ multidisciplinary/CCC/*.

[15] A. Agarwal, D. Durairajanayagam, S. Tatagari, S. C. Esteves, A. Harlev, R. Henkel, S. Roychoudhury, S. Homa, N. G. Puchalt, R. Ramasamy *et al.*, "Bibliometrics: tracking research impact by selecting the appropriate metrics," *Asian journal of andrology*, vol. 18, no. 2, p. 296, 2016.

[16] U. Senanayake, M. Piraveenan, and A. Zomaya, "The pagerank-index: Going beyond citation counts in quantifying scientific impact of researchers," *PloS one*, vol. 10, no. 8, p. e0134794, 2015.

[17] J. Turna, L. Bilský, J. Dzivák, and J. Kasáková, "Utilization of specialized ict infrastructure in the processes of science management in slovakia," in *ICETA*, 2017.