Transfer of Metadata into the National Information System of Scientific Research Activities with Automatic Authorship Association

UDC 655.411(0.034.2):004.65]:378.4(497.11)
DOI 10.18485/infotheca.2023.23.2.2

Vladimir Otašević
vladimir.otasevic@rcub.bg.ac.rs
ORCID: 0000-0002-9408-3454

University of Belgrade
School of Electrical Engineering
Belgrade, Serbia

ABSTRACT: Existing and alternative CRIS (Current Research Information System) systems, which deal with solving the problem of tracking scientific research productivity in Serbia, have proven to be incomplete. The new national information system for scientific research activities, named eNauka (eScience), through the application of modern methodologies, integration of external services, and process automation, should provide higher quality metadata and efficiency in the work of all its users. To save time in downloading and assigning records to research profiles and research organizations, eNauka utilizes techniques of automatic association and authorship recognition using persistent identifiers and similarity algorithms.

KEYWORDS: automatic authorship association, persistent identifiers, metadata validation, metadata deduplication, OAI-PMH protocol

PAPER SUBMITTED: 21 October 2023
PAPER ACCEPTED: 14 December 2023

1 Introduction

The eNauka system is a publicly accessible portal for tracking scientific research results, researchers, and institutions in the Republic of Serbia and it is designed for the unified presentation of scientific output, research areas, and achievements of the scientific research community[1]. The system is based on open-source software - DSpace-CRIS[2], which enables the maintenance of researcher profiles and research organizations, the collection of various research results, and tracking of citations, and more.

The main focus of eNauka is on saving time for all end users, regardless of their role in the system. From the perspective of researchers, time savings translate to the time spent entering all references and results they have published so far, while from the system administrator's perspective, it pertains to the time required to verify and validate all results reported by the researcher. For decision-makers, time savings translate to a more comprehensive overview of scientific productivity. At the data level, time savings will be achieved by not easily discarding available data sources that have already consumed researchers' time or administrators' time for verification and data entry. The eNauka system is open to downloading all metadata about scientific results that can meet the international standard for library information exchange, which implies the implementation of the OAI-PMH protocol (Lagoze et al. 2015). Time savings will be realized if metadata sources can provide authorship transfer. Authorship transfer implies that author names resolved by a persistent identifier at the data source can be transferred as such into the eNauka system.

The eNauka system allows for the transfer of metadata via the OAI-PMH protocol. Any information system, database, or internal infrastructure within any institution that can support the OAI-PMH protocol can potentially be integrated into eNauka. For eNauka, OAI-PMH represents a standard in terms of technical characteristics that need to be implemented. In addition to the technical correctness of the server implementing the OAI-PMH protocol, metadata must meet minimum requirements in terms of quality and availability for the results to be transferred to the eNauka system. Therefore, as part of the eNauka development, a separate service called OAIValidator has been created, which allows for transparent, public, and free semantic and syntactic validation of the correctness of metadata downloaded through the OAI-PMH protocol. Syntactic validation is most commonly applied to check identifiers and numerical values such as publication year. Semantic validation is used when recognizing the category of scientific result (e.g., article, conference paper, monograph, patent, technical solution, doctoral dissertation, etc.). The recognized categories of scientific results are aligned with the applicable regulations of the competent ministry (PC 2020). The OAIVal-

[1] eNauka System

[2] GitHub - 4Science/DSpace at Dspace-6_x_x-Cris

idator service supports authorship recognition and can be used to verify the accuracy of transferring researchers' persistent identifiers.[3]

One of the innovations that eNauka introduces compared to previous systems or alternative solutions is openness. Openness provides the opportunity for metadata about scientific research work to be more visible to all end users, regardless of whether they perform any of the roles recognized by eNauka. Openness enables easier verification of the correctness and accuracy of results because researchers are more cautious when submitting scientific results, while system administrators feel responsible for verifying every publication they must check. Logged-in users of the system have no privilege over anonymous users except in terms of functionality. Transparency towards users allows them to be free to write criticisms and praises of the system. Additionally, such an approach has proven to be very successful in collecting ORCID (Open Researcher and Contributor ID)[4] numbers of researchers. For the development team of eNauka, transparency towards users enables them to understand the needs of the scientific research community and to implement additional functionalities that could not have been foreseen in the system design process, but which can significantly contribute to improving the platform's quality and facilitate the work of all participants in the system.

Downloading metadata into the eNauka system is not dependent on the platform on which this metadata was created. Some institutions have been building their local infrastructure based on their knowledge for years, without relying on international standards, while others have opted for the implementation of ready-made (out-of-the-box) solutions. In this way, every publication, regardless of the institutional policy of depositing records, is a candidate for downloading. This means that an institution can continue the independent development of its infrastructure and adapt it to its needs without radically changing the way data is deposited. Institutions that have recognized the importance of information systems and have developed infrastructure for storing scientific research results can continue to work on their system as before, while eNauka will continue to collect results from institutional information systems by regular downloading. Certain institutions have shown a high level of interest in improving their infrastructure and supporting greater transfer of metadata, especially when it comes to transferring

[3] The latest version of the OAIValidator service is available at https://proref.rcub.bg.ac.rs/OAIValidator/

[4] ORCID

authorship through ORCID or other persistent identifiers (Haak et al. 2012; Otašević and Kosanović 2022).

The system design is based on the CRIS data model, which has been present and applied in various systems for many years (Jeffery and Asserson 2009). The eNauka portal consists of three entities: Research Organizations, Researchers, and Results. Additionally, the portal has been expanded with a statistics module, with the idea of providing continuous monitoring of scientific research productivity and other parameters relevant to science. The data model is created so that each entity can be maintained independently, but relationships can also be established between multiple instances of the same entity or between different types of entities. The presence of persistent identifiers for researchers plays a significant role in the accuracy of relationships. The data model is dynamically expandable, allowing for easier system upgrades with new entities if the need arises.

One of the main ideas behind creating the eNauka system is to gather all information related to the overall scientific productivity of researchers and institutions in the Republic of Serbia in one place. This implies that eNauka periodically collects results from other systems as well as from external services such as the ORCID platform, keeping track of where the results originated from. The eNauka system deduplicates and consolidates scientific research results in one place. In addition, the consolidated data is automatically checked for syntactic and semantic correctness. Furthermore, it allows for the enrichment of results with additional metadata, providing end users with a better understanding of scientific productivity.

The system is designed to support two processes that are very significant for every researcher and SRO (scientific research organization)(srb. *NIO - naučnoistraživačka organizacija*). Those processes are institutional scientific accreditation and selection for scientific research positions. The system aims to achieve a high degree of transparency in these processes because one of the main principles of this system is openness.

To ensure sustainability, eNauka relies on international standards and best practices in a technical sense, while also being aligned with legal regulations that define scientific research activities more closely. By listening to the needs of the scientific community, eNauka considers new solutions to make it easier for users to navigate the system. Additionally, eNauka strives to educate and inform end users through its functionalities and the implementation of new services, providing them with the opportunity to acquire new knowledge and skills (for example, the importance of an ORCID profile and international visibility (Arunachalam and Madhan 2016)).

2 Existing Solutions

So far, there have been several attempts and ideas on how to implement and technically support a system that would collect scientific research results and could support various processes. Each previous solution has not been able to fully meet the needs of the scientific research community. There has been room for improvement in the process, especially if the focus is on saving time. Each of the mentioned solutions has served as an example from which good and bad practices could be drawn. Identified practices have influenced eNauka in such a way that it does not neglect known and previously encountered problems, but actively works on resolving them.

Looking at sustainability, implementation, and development of other software, there are several different solutions. One technical implementation of a system could be ready-made solutions like Pure[5]. On the other hand, some solutions are implemented entirely independently or rely significantly on independent development. Such a system is the RIS - Serbian Researcher Registry[6]. Another similar example is the Novi Sad CRIS-UNS[7] used at the Faculty of Sciences in Novi Sad (Ivanović et al. 2017). Yet another possible solution is the eCRIS system based on data from the COBISS database, a commercial solution maintained and developed by IZUM (Institute of Information Science in Maribor), actively used in several scientific research libraries as well as university libraries, such as the University Library "Svetozar Marković" (Tomic and Ljubišić 2020).

The eNauka system is a kind of combination of various implementations, but its foundation is based on open-source software. However, eNauka significantly deviates from the original initial open-source code, and the entire system upgrade represents independent development. For eNauka, it is very important that previous results, previous development, and data collected through other systems are not discarded, and that redundant data entry is avoided.

The RIS system, in its structure, closely resembles the CRIS model to a significant extent. The system represents independent development, meaning it relies less on the implementation and integration of existing software solutions. During its active use and application period, a significant number of various references on scientific research results have been collected. When considering the level of information availability and open data, the

[5] Pure | The World's Leading RIMS or CRIS | Elsevier
[6] RIS – Registar Istraživača Srbije
[7] CRIS-UNS

RIS system is largely implemented as a closed system. This closed nature also extends to the data, as researchers are identified based on RIS IDs, internal identifiers within the system. The lower transparency and visibility of this system reduce the accountability of various users towards the accuracy of the collected data. RIS represents one of the main sources of data initially incorporated into eNauka, for the following reasons:

- It contains references to scientific research results that were exclusively available within the RIS system,
- In the RIS system, researchers typically enter references themselves. This means that results could only be associated with individuals through manual entry,
- In addition to research references, the RIS system contains other administrative data that are important for establishing and identifying other entities.

The Pure platform is another possible technical solution that supports the CRIS data model and relationships between entities. It is a commercial solution developed and maintained by Elsevier. Pure is a modern system based on contemporary information technologies, which utilizes data available in the Scopus database[8] as its foundation. One of the main challenges for such a system would be expanding the data model because scientific research productivity, primarily its evaluation, is primarily based on data available in the Web of Science[9], and to a lesser extent on Scopus data (Kosanović 2004). Another challenge for this system would be incorporating domestic productivity. Research results from social sciences and humanities are less visible and available in international databases, but significantly more present in domestic index databases.

The solution that is commercial, but understands much better how scientific research productivity emerges when published in domestic journals or by domestic publishers, is the COBISS system. The system recognizes several different user roles depending on the level of responsibility and capabilities associated with each role. Data in the system are publicly available, which significantly adds another level of responsibility towards the data. For eNauka, COBISS is a significant source of metadata, primarily because it represents the main source of metadata on published and released results by domestic publishers. In combination with the eCRIS system, it is possible to establish the transfer of metadata along with authorship information.

[8] More about Scopus https://www.elsevier.com/products/scopus

[9] Web of Science

A good example of how it's possible to implement a CRIS system as a reliable solution in the work of a scientific research organization (SRO) is CRIS-UNS. The CRIS-UNS system was developed as an independent solution. However, this solution is primarily designed and applied for operation at the level of a single institution. The primary role of such solutions is to support the work of individual SROs and enable the introduction of reference input procedures simply. This solution is recognized as an example of best practice that has been applied in several other SROs as well. In technical terms, the system meets the requirements to be integrated into eNauka.

3 Implementation of the data model and infrastructure of the eNauka system

The eNauka system is based on the open-source software solution Dspace-CRIS. Dspace-CRIS software was developed by 4Science but heavily relies on the community and individuals who continuously contribute to its improvement (Mornati and Bollini 2013). Additionally, Dspace-CRIS is based on another open-source software solution, namely the Dspace platform (2017). As of the time of writing this paper, there are 206 accredited SROs. Out of this number, 111 have locally developed or utilized shared infrastructure that meets the minimum technical requirements for integration with eNauka. Out of the 111 registered sources harvested via the OAI-PMH protocol, a significant number are based on Dspace or Dspace-CRIS platforms (64%), which are easily extensible to support authorship transfer (Kosanović et al. 2019; Smederevac et al. 2020).

As a well-established data model due to its significant prevalence, eNauka is based on the Dspace-CRIS 6.3 data model as well as a major portion of the source code. The reason for choosing version 6.3 is that at the time when the eNauka project was initiated, there was no international stable production version of Dspace 7.X or Dspace-CRIS 7.X. Most production instances are based on versions 6.X or 5.X, which are very similar to each other. These versions differ drastically, both technologically and in terms of data model, from version 7.X.

Dspace-CRIS 6.3 represents an upgrade from Dspace 6.X, which was developed using two different technologies. One is based on the XMLUI technology (Sarang 2006), which is flexible, extensible, and easily applicable, but faces a significant sustainability issue as the technology has long been considered outdated and surpassed. On the other hand, Dspace offers an implementation developed in JSPUI technology (JCP 2000). A solution based on Java programming language is highly popular among applications that have been present and active for a long time. Dspace-CRIS 6.3. was developed as an upgrade from Dspace 6.X, specifically in JSPUI technology. Therefore, eNauka's main component (core) is based on JSPUI technologies. The eNauka system aims to develop its enhancements primarily related to expanding existing functionalities or introducing new components using more modern technologies. Additional tools, scripts, and add-ons, depending on the needs, utilize development frameworks such as Angular (Google 2024) and Spring (Deinum et al. 2012), REST API structure (Fielding 2000), and other technologies commonly found in modern software solutions. Special attention is given to writing scripts and designing tools that are periodically executed, including scripts that generate variants of names (e.g., permutations, transliterations, etc.) and tools for automatically associating works with researchers.

The eNauka system recognizes several different user roles. Anonymous users represent the largest group. The system aims to be open and transparent in every aspect, allowing end-users to be the main controllers of the system's work quality and displayed data. Regarding the group of researchers, eNauka strives to simplify work processes within the system. In addition to the ability to edit additional personal data and manually input their references solely through external services, researchers are gradually informed and educated through the system. The biggest novelty for researchers is the ORCID profile, whose representation was very low (<15%) before the introduction of eNauka, reaching 83,24% four months after production release. This percentage will be higher when eNauka starts to be promoted within the research community. The third very important role in the eNauka system is the NIO editor. NIO editors are recognized as individuals engaged by SRO and represent the main guardians of the quality of metadata about research results. Most functionalities related to verification, approval, and editing of records are tailored to the work of NIO editors, saving them time needed for record review and authorship assignment.

In technical terms, eNauka, through its enhancement, should support process automation in every segment. The main component of the CRIS system is entities. The main entities in eNauka are research results, researchers, and SRO. The goal is to automate every process using information technology that would lead to saving time for researchers and NIO editors. The main challenge lies in automatically recognizing PIDs (persistent identifiers). PIDs are applied to all entities, like publications, researchers, and organizations. Their application provides unique identification of each entity and their continuous integration with external services. Since the main data in eNauka is scientific research results, i.e., their affiliation to individual researchers or SRO, the main challenge is achieving automation in establishing relationships between entities (Figure 1).

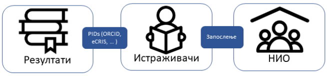


Figure 1. Representation of three entities in eNauka and the data defining direct and indirect relationships

Each of the entities has the characteristic of a strong entity, which allows them to exist independently. This implies that it is possible to edit, add, or remove each of them regardless of other entities. However, besides entities, CRIS systems also consist of relationships that exist between them. Therefore, regardless of their independence, every change to the data should be made carefully and cautiously to avoid disrupting other entities. In addition to research data, eNauka also includes other data that more closely describe or uniquely identify each researcher or research organization. Such data represent administrative data and are recognized as a separate entity within the system. Therefore, the data in eNauka should be divided into two groups:

1. Research Data - Metadata about references. Publicly visible and accessible to everyone after approval by NIO editors or downloading from reliable sources.
2. Administrative Data - Managed by NIO referents. A separate closed module with authorized access rights. Contains a larger amount of data about researchers and SROs (e.g., ORCID and employment). Important for starting procedures for academic promotion or accreditation.

4 Sources of Metadata

Researchers and SROs as entities in eNauka are defined by administrative data. If a specific researcher or institution cannot be found in eNauka, or if there is an error in administrative data such as names, surnames, employment, titles, ORCID, etc., then the NIO referent within their institution should enter, modify, or delete individual researchers or specific data through the administrative module. Data between the public eNauka portal and the administrative data editing module are synchronized daily. This synchronization model was chosen because changes in administrative data are not massive or frequent now when establishing relationships between entities.

For the public part of eNauka, in addition to administrative data, metadata about scientific research results also play a significant role. Additionally, each record that is regularly downloaded from existing sources, carries metadata (`dc.identifier.uri`), which represents a permanent link to the source from which the record was downloaded. Such an approach ensures the verifiability of available information and authorship verification. Two important factors that influenced the selection should be considered: with what priority the metadata sources will be integrated and in what manner:

1. Reliability of metadata sources - if the data comes from sources where someone responsible and qualified has already reviewed the metadata, then there is no need to verify such records at those sources. However, if the data was entered solely by the researcher, then the NIO editor must verify them before assigning them to the author.
2. Importance of metadata sources - if the metadata source has been previously used as an information system for tracking scientific productivity or contains a large number of significant publications, then such sources were prioritized during integration and this is because the availability of such metadata would significantly save time for researchers and NIO editors.

Metadata sources can be categorized based on the type of results they collect. The first group comprises information systems and databases that collect results with a common attribute of publication. Such sources can gather metadata records published by domestic publishers or those published solely in Web of Science, etc. On the other hand, there are infrastructures developed within individual SROs that collect all metadata on the scientific research results of their employees. The problem is that certain SROs cannot develop their infrastructure, so they opt for shared repositories. According to the acquisition model, eNauka enriches its records by regularly harvesting reliable sources that do not need to be verified or by researchers initiating the download of publications from external sources, which NIO editors then verify.

Various software solutions enable the implementation of the OAI-PMH protocol. In technical terms, to support the OAI-PMH protocol, there must be a software solution (referred to as an OAI-PMH server) capable of sending data in the expected format, and there must be another software solution on the receiving end capable of fetching that data. Each SRO has the freedom to choose a software solution that facilitates sending metadata. Correct implementation of the OAI-PMH protocol has been established by 111 SROs, and this number of implementations is expected to continue growing. An increase in servers implementing the OAI-PMH protocol is anticipated as the system is promoted within the scientific research community. Recognized implementations of servers implementing the OAI-PMH protocol are listed in Table 1.

Table 1. Types of software solutions used as sources from which metadata is retrieved

Software Solution	**Description**
Institutional Repository	Repository the importance of repositories, whose application extends far beyond integration into eNauka.
Shared Repository	The common infrastructure of universities or libraries
International Repository	International infrastructure
Independent Development	Some institutions have mastered the OAI-PMH standard and have managed to independently develop an OAI-PMH server
External Services	Researchers can only retrieve metadata from external services. SRO does not support any infrastructure.

SROs that have recognized the importance of their infrastructure have mostly opted to establish institutional repositories. Institutional repositories are used and exist independently of eNauka. As another option, some SROs have chosen to collaborate, usually at the university level, and utilize shared infrastructure. Universities such as the University of Kragujevac and the University of Criminalistics and Police Studies have university repositories. Additionally, libraries like the University Library "Svetozar Marković" provide their infrastructure and support to SROs to join the shared repository[10]. As a third option, SROs aspiring to establish repositories may opt for international infrastructure. So far, Zenodo[11] has been recognized as an international repository. From a metadata perspective, Zenodo collects all types of results (a catch-all repository). Some SROs had local databases and information systems even before eNauka, which they managed to advance to support an OAI-PMH server. The Faculty of Sciences in Novi Sad, through CRIS-UNS, integrated into eNauka and helped other SROs to integrate into eNauka following a similar model. The Faculty of Electrical Engineering in Belgrade also upgraded its information system and met the technical requirements for integration into eNauka, assisting other SROs using the same information system to integrate into the eNauka system. Institutions have begun to recognize the importance of ORCID and are expanding their databases with new data, knowing that transferring them to eNauka will save time on record editing.

The metadata available in the COBISS database is also downloaded into eNauka using the OAI-PMH protocol. For records downloaded from the COBISS database, it is important to have a certain typology. When eNauka downloads records from the COBISS database, it automatically recognizes the category of the scientific result. The OAI-PMH implemented to disseminate records from the COBISS database has been expanded so that downloading is done according to the COMARC specification (IZUM 2023a, 2023b). The visibility and transparency of records that have arrived in eNauka from the COBISS database have made it easier to identify and correct errors at the metadata level at the source.

5 The system infrastructure

The public portal of the eNauka system was developed in an environment commonly found in non-commercial information systems. Ubuntu operating

[10] PHAIDRA, Digital Repository Of University of Belgrade.

[11] Zenodo

system was used as the operating system. Apache Tomcat® 9.x was utilized as the web server (The Apache Software Foundation, 2017). The foundation of the eNauka public portal is the Dspace-CRIS 6.3 platform based on the Java programming language and implemented as a Maven project. Apache Tomcat® 9.x server is suitable and compatible with Dspace-CRIS version 6.3. PostgreSQL was used as the relational database management system (PostgreSQL Global Development Group, 2023), and Apache Solr was used for indexing and search (The Apache Software Foundation, 2014).

In addition to the technical characteristics of the environment in which eNauka was developed, the eNauka system consists of a public portal and several services. Due to the complexity of the implemented processes, as well as the multiple sources of metadata and administrative data, the eNauka system is divided into several separate units (modules). Each module is developed as a standalone application that can exist and operate independently of other modules. However, for the eNauka system to function as a service, some modules must be in frequent communication with each other. Figure 2 illustrates a simplified infrastructure model representing the eNauka system. The orange arrow shows the path of metadata movement during downloading. The blue arrow indicates synchronizations with internal and external services/modules

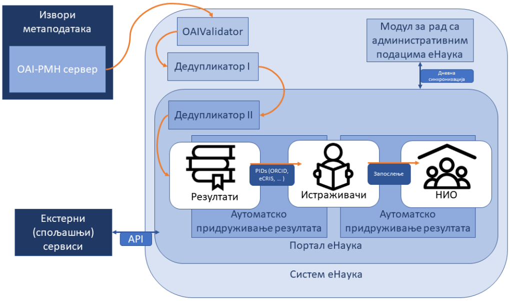

Извори метаподатака
OAI-PMH сервер
OAIValidator
Дедупликатор I
Модул за рад са административним подацима еНаука
Дневна синхронизација
Дедупликатор II
Резултати
PIDs (ORCID, eCRIS, ...)
Истраживачи
Запослење
НИО
Аутоматско придруживање резултата
Аутоматско придруживање резултата
Екстерни (спољашњи) сервиси
API
Портал еНаука
Систем еНаука

Figure 2. Presentation of the eNauka infrastructure.

In Figure 2, the eNauka portal is observed. It represents the main segment of the eNauka system, as this segment is publicly accessible and serves as the main component that integrates all other modules. The eNauka portal consolidates administrative data about researchers and research organizations, as well as metadata about scientific research results. Additionally, the eNauka portal serves as the main hub for other external services integrated within it. Table 2 provides an overview of external services that are significant for authorship affiliation in eNauka.

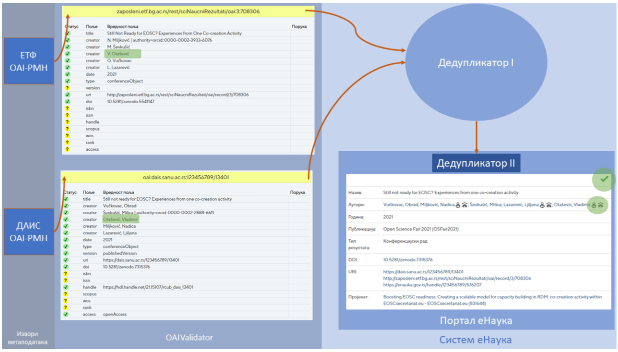
Table 2. List of external services that are integrated into eNauka, and are important for the transfer and association of authorships.

External Service	Application
ORCID	Authentication; Authorization; Retrieval of works; Automatic authorship affiliation; *Submission of works(push)
CrossRef	Enrichment of records; Quality verification
KoBSoN	Automatic categorization of results; Retrieval of records
KNR	Retrieval of records; Automatic authorship affiliation; Link to researcher's profile within the KNR platform
eCRIS	Automatic authorship affiliation; Link to researcher/SRO profile within the eCRIS system
RIS	Automatic authorship affiliation

In Table 2, the most significant external services integrated with eNauka are presented. These services contribute to eNauka by regularly updating researchers and result entities with new data. From the perspective of automatic authorship affiliation, these services provide necessary data for establishing mechanisms in eNauka that automatically affiliate results with researchers, verify the accuracy of transferred authorships, and check for duplicates. There are other external services integrated with eNauka, but their application is specific and not of interest to end users. Records found on the eNauka portal are regularly enriched with data obtained from these external services. In the "Application" column, certain values are marked with an asterisk "*", indicating that the specified application is expected but not available at the time of writing this paper.

6 Metadata verification and processing phases

The metadata retrieved through regular harvesting before reaching eNauka must undergo several phases of verification and processing. The first step is the validation of the OAI-PMH server. To check the OAI-PMH server and metadata, the OAIValidator has been developed. Checking the OAI-PMH server involves verifying if the server is accessible, if it has implemented functions that return lists of sets by which records are grouped, if there is an Identify page with basic information about the metadata source, and if the expected format for transferring metadata is supported. In addition to the basic functionality of checking the validity of the OAI-PMH server, OAIValidator helps repository administrators verify the semantic and syntactic correctness of metadata in one place. Also, OAIValidator provides basic statistical indicators about records as well as detailed messages about invalid metadata. In the context of eNauka, all sources of metadata that are regularly harvested automatically undergo a validation process that will machine-recognize records/metadata that can enter the eNauka system. Figure 3 shows an example of validation of one metadata source and how eNauka views one of the retrieved records with authorships. The NIO editor has additionally reviewed the shown example to resolve the authorship request.



zaposleni.etf.bg.ac.rs/rest/sciNaucniRezultati/oai:3:708306

Статус	Поље	Вредност поља	Порука
✓	title	Still Not Ready for EOSC? Experiences from One Co-creation Activity	
✓	creator	N. Miljković | authority=orcid::0000-0002-3933-6076	
✓	creator	M. Ševkušić	
✓	creator	V. Otašević	
✓	creator	O. Vučkovac	
✓	creator	L. Lazarević	
✓	date	2021	
✓	type	conferenceObject	
?	version		
✓	uri	http://zaposleni.etf.bg.ac.rs/rest/sciNaucniRezultati/oai/record/3/708306	
✓	doi	10.5281/zenodo.5541147	
?	isbn		
?	issn		
?	handle		
?	scopus		
?	wos		
?	rank		
?	access		

oai:dais.sanu.ac.rs:123456789/13401

Статус	Поље	Вредност поља	Порука
✓	title	Still not ready for EOSC? Experiences from one co-creation activity	
✓	creator	Vučkovac, Obrad	
✓	creator	Ševkušić, Milica | authority=orcid::0000-0002-2888-6611	
✓	creator	Otašević, Vladimir	
✓	creator	Miljković, Nadica	
✓	creator	Lazarević, Ljiljana	
✓	date	2021	
✓	type	conferenceObject	
✓	version	publishedVersion	
✓	uri	https://dais.sanu.ac.rs/123456789/13401	
✓	doi	10.5281/zenodo.7315376	
?	isbn		
?	issn		
✓	handle	https://hdl.handle.net/21.15107/rcub_dais_13401	
?	scopus		
?	wos		
?	rank		
✓	access	openAccess	

Назив:	Still not ready for EOSC? Experiences from one co-creation activity
Аутори:	Vučkovac, Obrad; Miljković, Nadica; Ševkušić, Milica; Lazarević, Ljiljana; Otašević, Vladimir
Година:	2021
Публикација:	Open Science Fair 2021 (OSFair2021)
Тип резултата:	Конференцијски рад
DOI:	10.5281/zenodo.7315376
URI:	https://dais.sanu.ac.rs/123456789/13401 http://zaposleni.etf.bg.ac.rs/rest/sciNaucniRezultati/oai/record/3/708306 https://enauka.gov.rs/handle/123456789/576207
Пројекат:	Boosting EOSC readiness: Creating a scalable model for capacity building in RDM: co-creation activity within EOSCsecretariat.eu - EOSCsecretariat.eu (831644)

Figure 3. An example of validation, deduplication, and consolidation of records that were automatically transferred from two metadata sources.

The second step involves the deduplication of metadata. Since the deduplication process is complex and time-consuming, it takes place in a separate module. Deduplication is based on two recognition patterns:

1. Identifiers (DOI, WoS UT, Scopus ID) - the mentioned identifiers must be unique; otherwise, it's considered a duplicate. There are exceptions, but in those cases, a different pattern is applied;
2. Title, year, publication type - Not all publications have PIDs. Publications with similar normalized titles, publication years with an error of ± one year, and the same type of research result are considered duplicates.

The identified duplicate is not discarded; instead, an attempt is made to merge the records automatically. The most significant metadata in the record is authorship. The deduplicator will attempt to recognize authors who do not have assigned authorship (e.g., ORCID, eCRIS-ID, etc.) if the duplicate record contains additional identifiers. In addition to authorship, the deduplicator will merge all PIDs and URI values for recognized duplicates to permanently retain the trace of where the records were downloaded from.

The eNauka portal is continuously active, with its users constantly editing or withdrawing records from eNauka. Since the deduplication process is time-consuming, the data in the deduplicator may be not synchronized with the data in eNauka. Therefore, when transferring deduplicated records to eNauka, another automatic deduplication will be performed in real-time if it is recognized that the record already exists or is pending verification by the NIO editor. Additionally, for records that already exist in eNauka, metadata merging and enrichment will be conducted. It is essential if a duplicate record is received from other metadata sources containing new information about authorships.

7 Data Model

The data model in eNauka is based entirely on the data model existing in Dspace-CRIS 6.3., so there was no need to introduce additional tables or columns. This approach allows eNauka to remain standardized and not deviate technically from the original code when it comes to the data model. However, certain modifications have been made to accommodate values from the database. By default, Dspace-CRIS 6.3. recognizes ORCID as authorship values. On the other hand, integrated systems store authorships through other PIDs such as eCris, RISiD, Karton Vojvodine ID, and others.

Therefore, eNauka is designed to allow the storage and association of other identifiers.

The eNauka system retrieves metadata from external sources through regular harvesting via OAI-PMH servers, so it was necessary to define which standards and specifications for metadata would be applied. It is possible to download metadata about records that follow one of the following specifications:

1. Dublin Core (abbreviated as dc) - an internationally accepted standard specification for metadata exchange (DCMI 2012; Kunze and Baker 2015; ISO 2017; Хокнер and Будрони 2011);
2. DSpace Intermediate Format (dim) - a non-standardized metadata specification primarily designed for data exchange within the DSpace platform itself.

The Dublin Core specification was easier to apply and integrate because it is a widely adopted metadata schema that can be propagated through the OAI-PMH server (Jackson et al. 2008). However, such an approach also carries certain limitations. According to the dc specification, it is not possible to transmit authorship (e.g., ORCID). To overcome this issue, eNauka followed the guidelines of international organizations and best practices by implementing a solution that extends the authorship label to include an additional attribute (`"id"`) where the OAI-PMH server can place the authorship value (2018). Other alternative solutions can be applied for transmitting authorship, which were considered when extending the OAI-PMH metadata transmission that implements the `dc` specification (BASE 2023). Another significant drawback of this specification is the limitation in recognizing labels. Each label consists of two parts, `"schema"` and `"element,"` which serve for recognition. This problem arises when transmitting PID values, which is a major issue for eNauka as most operations are based on these values. For example, a record may have a DOI, WoS UT, Scopus ID, PMID, COBISS ID, and an internal numerical identifier. All these PID values will arrive through the `"dc.identifier"` field, which means that recognized identifiers exist but it is not known for a specific value which identifier it represents. Of course, eNauka has mechanisms for semantic recognition and syntactic verification of PID values and will attempt to categorize values according to precisely determined identifiers, but there is a possibility that a NIO editor will have to subsequently correct such a record.

On the other hand, eNauka offers the possibility of downloading data according to the `dim` specification. Since this specification is non-standardized, it is possible to make all kinds of modifications to it. There is no problem with identifiers and their recognition because each label can be defined with a maximum of three segments, i.e., besides "`dc`" and "`element`," there is the "`qualifier`" attribute. Additionally, the dim specification supports the transmission of authorship values on all labels without additional modifications. The only difference is that the attribute is called "`authority`." By applying XSLT (Extensible Stylesheet Language Transformations) transformation (W3C 1999), it is possible to translate `dim` into the `dc` specification unambiguously. XSLT is a language for transforming XML documents. It can be concluded that `dc` represents a subset of the dim specification. It is possible to transform dc into dim specification using XSLT, but such transformation is not straightforward or reliable, especially if the record is rich in identifiers.

The eNauka portal has been enhanced with additional mechanisms that facilitate the work of NIO editors and researchers when checking the quality of metadata. In real-time, the eNauka portal will flag potential duplicates based on titles that could not be automatically deduplicated because a sufficient level of confidence that they are genuine duplicates has not been achieved. These are situations where human review of the record is necessary. In addition to recognizing duplicate titles, eNauka will perform a check during record editing if additional identifiers are added to verify if a record with such identifiers already exists. A problem has been observed with records that have invalid or non-existent DOI numbers. The eNauka system automatically checks each DOI number and marks those that are not valid.

8 Conclusion

The eNauka system represents a significant advancement compared to existing alternative solutions. It is tailored to the work of NIO editors and researchers. The automatic assignment of authorship significantly saves time for all users. If metadata sources meet technical requirements, meaning they have an OAI-PMH server and their metadata meets syntactic and semantic correctness criteria, then eNauka will support the integration of such sources. Centralizing data collection, while preserving links to original sources, allows for the review of individual records in eNauka and the verification of the same records in the original sources from which they were retrieved. Data sources are decentralized, allowing each SRO to continue using their information systems as they have done before. On the other hand, the eNauka system will automatically perform validation, deduplication, verification, and enrichment of records through regular data retrieval, providing end users with better insights into the records.

Furthermore, anonymous users are recognized as having a distinct role in this system, and their contribution through continuous quality checks of available information enables this system to be better. Transparency in information introduces a new level of accountability and reduces the possibility of (un)intentional errors slipping through.

Regarding the infrastructure, it is designed to achieve maximum modularity. Such an approach ensures that each module can be developed and maintained independently of others. Additionally, modules can operate independently of each other, but for the entire system to function, the modules must be accessible and for there to be mutual synchronization. With its innovation and advancement compared to alternative solutions, the eNauka system has become a source of scientific information. End users have the opportunity to acquire new skills that are necessary for anyone who wants to be part of the modern international scientific research community.

The eNauka system with entities of researchers, SROs, and scientific research results represents a CRIS system. The next step for eNauka would be to add projects as a new entity, in addition to the existing ones. Projects would contribute to opening up a new perspective on monitoring scientific research productivity.

References

Arunachalam, SSubbiah, and Muthu Madhan. 2016. "Adopting ORCID as a Unique Identifier Will Benefit All Involved in Scholarly Communication." *The National Medical Journal of India* 29 (4): 227–234.

BASE. 2023. *Bielefeld Academic Search Engine. n.d. Golden Rules for Repository Managers.* Преузето 24.10.2023, https://www.base-search.net/about/en/faq_oai.php#dc-creator.

DCMI. 2012. *Dublin Core™ Metadata Element Set, Version 1.1: Reference Description.* Преузето 25. 10. 2023, https://www.dublincore.org/specifications/dublin-core/dces/.

Deinum, Marten, Koen Serneels, Colin Yates, Seth Ladd, and Christophe Vanfleteren. 2012. “Pro Spring MVC: With Web Flow.” Chap. Spring Framework Fundamentals, edited by Marten Deinum, Koen Serneels, Colin Yates, Seth Ladd, and Christophe Vanfleteren, 25–50. Berkeley: Apress. https://doi.org/https://doi.org/10.1007/978-1-4302-4156-0_2.

DSpace 6.x Documentation. 2017. Preuzeto 28. 10. 2023, https://dspace.umkt.ac.id//handle/463.2017/19.

Fielding, Roy Thomas. 2000. “Architectural Styles and the Design of Network-Based Software Architectures.” Преузето 25. 10. 2023, https://ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf. PhD diss., University of California, Irvine.

Google. 2024. *Angular - Introduction to Angular Concepts.* Преузето 6. 1. 2024, https://angular.io/guide/architecture.

Haak, Laurel L., Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. 2012. “ORCID: A System to Uniquely Identify Researchers.” *Learned Publishing* 25 (4): 256–264. https://doi.org/https://doi.org/10.1087/20120404.

ISO. 2017. *ISO 15836-1:2017: Information and documentation The Dublin Core metadata element set.* Преузето 25. 10. 2023, https://www.iso.org/standard/71339.html..

Ivanović, Dragan, Dušan Surla, Miroslav Trajanović, Dragan Misić, and Zora Konjović. 2017. “Towards the Information System for Research Programmes of the Ministry of Education, Science and Technological Development of the Republic of Serbia.” *Procedia Computer Science,* no. 106, 122–129. https://doi.org/https://doi.org/10.1016/j.procs.2017.03.044.

IZUM. 2023a. *COMARC/A. Format za normativne podatke: priručnik za korisnike.* Preuzeto 25. 10. 2023, https://home.izum.si/izum/e-prirucnici/2_COMARC_A/Ceo_2_COMARC_A.pdf.

IZUM. 2023b. *COMARC/B. Format za bibliografske podatke: priručnik za korisnike.* Preuzeto 25. 10. 2023, https://home.izum.si/izum/e-prirucnici/1_COMARC_B/Ceo_1_COMARC_B.pdf.

Jackson, Amy S., Myung-Ja Han, Kurt Groetsch, Megan Mustafoff, and Timothy W. Cole. 2008. “Dublin Core Metadata Harvested Through OAI-PMH.” *Journal of Library Metadata* 8 (1): 5–21. https://doi.org/https://doi.org/10.1300/J517v08n01_02.

JCP. 2000. *JavaTM Servlet Specification Version 2.3.* Preuzeto 27. 10. 2023, https://jcp.org/aboutJava/communityprocess/first/jsr053/servlet23_PFD.pdf.

Jeffery, Keith, and Anne Asserson. 2009. "Institutional Repositories and Current Research Information Systems." *New Review of Information Networking* 14 (2): 71–83. https://doi.org/https://doi.org/10.1080/13614570903359357.

Kosanović, Biljana. 2004. "The Availability of Scientific Information in Serbia: Trends and Perspectives." *Hemijska industrija* 58 (4): 158–160. https://doi.org/https://doi.org/10.2298/HEMIND0404158K.

Kosanović, Biljana, Milica Ševkušić, Vasilije Rajović, and Nenad Popović. 2019. *Setting the Scene for a Sustainable National Repository Network in Serbia.* Преузето 24. 10. 2023, https://zenodo.org/records/3509971.

Kunze, John A., and Thomas Baker. 2015. *The Dublin Core Metadata Element Set. RFC 5013.* Preuzeto 27. 10. 2023, https://datatracker.ietf.org/doc/rfc5013/.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner, eds. 2015. *Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0.* Preuzeto 27. 10. 2023, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

Mornati, Susanna, and Andrea Bollini. 2013. "DSpace-CRIS: An Open Source Solution." EuroCRIS Membership Meeting, https://dspacecris.eurocris.org/bitstream/11366/73/1/CINECA_DSpace_CRIS_An_Open_Source_Solution_v2.pdf.

Otašević, Vladimir, and Biljana Kosanović. 2022. "Resolving authorship using the ORCIR identifier." In *Zbornik radova 28. IKT konferencije "YU INFO 2022",* 48:167–172. Informaciono društvo Srbije.

Sarang, Poornachandra. 2006. "Pro Apache XML." Chap. The Apache Cocoon Framework, 279–325. Berkeley: Apress. https://doi.org/https://doi.org/10.1007/978-1-4302-0166-3_7.

Smederevac, Snežana, Dejan Pajić, Sanja Radovanović, Silvia Gilezan, Petar Čolović, and Branko Milosavljević. 2020. *Otvorena nauka: praksa i perspektive.* Novi Sad: Univerzitet u Novom Sadu.

Tomic, Emina Cano, and Ljubica Ljubišić. 2020. “General Overview of the Growth and Development of the Services Provided by Virtual Library of Serbia Network and Their Impact on Collection of Statistical Library Data in Serbia.” *Qualitative and Quantitative Methods in Libraries* 9 (1): 77–98.

Using Persistent Identifiers with Literals in Dublin Core-Based Metadata in XML. 2018. Preuzeto 28. 10. 2023, https://github.com/dcmi/pids_in_dc/blob/master/proposal/The_Association_of_Persistent_Identifiers_with_Literals_in_XML-formatted_Metadata_using_Dublin.md.

W3C. 1999. *XSL Transformations (XSLT).* Преузето 29. 10. 2023, https://www.w3.org/TR/xslt-10/.

РС. 2020. *Правилник о стицању истраживачких и научних звања ("Службени гласник РС", бр. 159 од 30. децембра 2020, 14 од 20. фебруара 2023.)* Преузето 25. 10. 2023, https://www.pravno-informacioni-sistem.rs/SlGlasnikPortal/eli/rep/sgrs/ministarstva/pravilnik/2020/159/18/reg.

Хокнер, Маркус, and Паоло Будрони. 2011. “Пројекат Репозиторијума Униврзитета у Бечу.” Преузето 25. 10. 2023, http://infoteka.bg.ac.rs/pdf/Srp/2011-1/INFOTHECA_XII_1_August_23-33.pdf, *Инфотека* 12 (1): 23–33.