Methods and Tools for Building Open Systems of Scientific Research Support

Oleksandr Novytskyi[1], Oleg Spirin[2]

[1] Institute of Software Systems of the National Academy of Sciences of Ukraine, Akademika Glushkova Avenue, 40, Kyiv, 03187, Ukraine
[2] Institute for Digitalisation of Education of NAES of Ukraine, M.Berlyns'koho St., 9, Kyiv, 04060 , Ukraine

Abstract

The article is devoted to the problems of building an integrated environment within the project implementation of the development of open science in Ukraine. To ensure the aggregation of metadata, it was necessary to conduct a study to identify modern methods of metadata harvesting. Since the solution to this problem involves using ready-made solutions, there was a review of the tools that allow the achievement of the project's goals. Therefore, a detailed comparative review was carried out, as well as methods for improving the semantic integration of information for the protocol will be proposed for OAI-PMH.

Keywords

OAI-PMH, Vufind, Data integration, Metadata harvesting, ETL

1. Introduction

When projecting an effective system for sharing scientific results, it is essential to consider the specificity of scientific data. Each field of study has its unique characteristics: in some cases, researchers handle large volumes of experimental data in the form of images, while in others, they work with complex data structures such as chemical formulas or information about astronomical objects. Additionally, each data repository employs its own metadata standard. Thus, the issue of integrating heterogeneous resources and providing unified access to them arises.

In Ukraine, various initiatives have been undertaken to establish an open science system aimed at enhancing the visibility of research outcomes of NASU scientists in the open science information environment using modern technical and informational tools. The effectiveness of these initiatives is to be evaluated using scientometric indicators like citation index, Hirsch index, i-index, g-index, which will promote the development of science in Ukraine, international scientific collaboration, and broaden access to the scientific community, organizations, and enterprises both in Ukraine and internationally to the research results of NASU. As part of this project, a series of programmatic and technical measures are planned to integrate into the scientific and educational space [1].
For the publication and preservation of scientific results, digital libraries (DL) and journals are used, providing access to a vast array of resources in the form of digital objects and a wide variety of tools for searching, viewing, and utilizing digital content. Complex and flexible metadata schemas, such as Dublin Core, MODS, and METS, have been developed and are used to describe digital

14th International Scientific and Practical Conference from Programming UkrPROG'2024, May 14-15, 2024, Kyiv, Ukraine

* Corresponding author.
† These authors contributed equally.

o.novytskyi@iss.nas.gov.ua (O. Novytskyi), oleg.spirin@gmail.com (O. Spirin)
0000-0002-9955-7882 (O. Novytskyi), 0000-0002-9594-6602 (O. Spirin)

objects in collections [2]. The semantic layer allows for more effective extraction of necessary information than the metadata level.

2. Methodology

The study presents a review of current approaches to data integration and identifies the methods most suitable for the integration of scientific information. An analysis of software designed for the creation of integrated environments has been conducted. Publications in Google Scholar on the topic of scientific data integration from 2020 to 2024 were analyzed. Software with new releases in 2024 was identified. Each of these software packages was installed to evaluate its capabilities for addressing the problem of scientific data dissemination and integration. The results were tested in the open science project implemented by the National Academy of Sciences of Ukraine.

3. Approaches to the integration of structured data

In information integration, several challenges can be identified, such as schema integration, data warehouse integration, data integration (also known as enterprise information integration, EII), catalog integration, and the construction and storage of big data. The most complex step is identifying correspondences between semantically related entities in local and global ontologies. Interoperability is the ability of separate systems to exchange information and use it. The term interoperability is widely applicable, particularly concerning the effective coexistence of information resources. This issue can be defined in various aspects, including the semantic aspect. Semantic interoperability is the ability of information systems to exchange and interpret the content automatically. Achieving semantic interoperability involves resolving the heterogeneity of information being exchanged. Semantic heterogeneity is more complex than syntactic and structural heterogeneity as it deals with varying contexts [3], [4], [5], [6], [7]. Syntactic heterogeneity results from requirements for metadata formats. Standardizing data formats is a common approach to solving syntactic heterogeneity issues. For example, XML is used as a standard format for all web-accessible data.

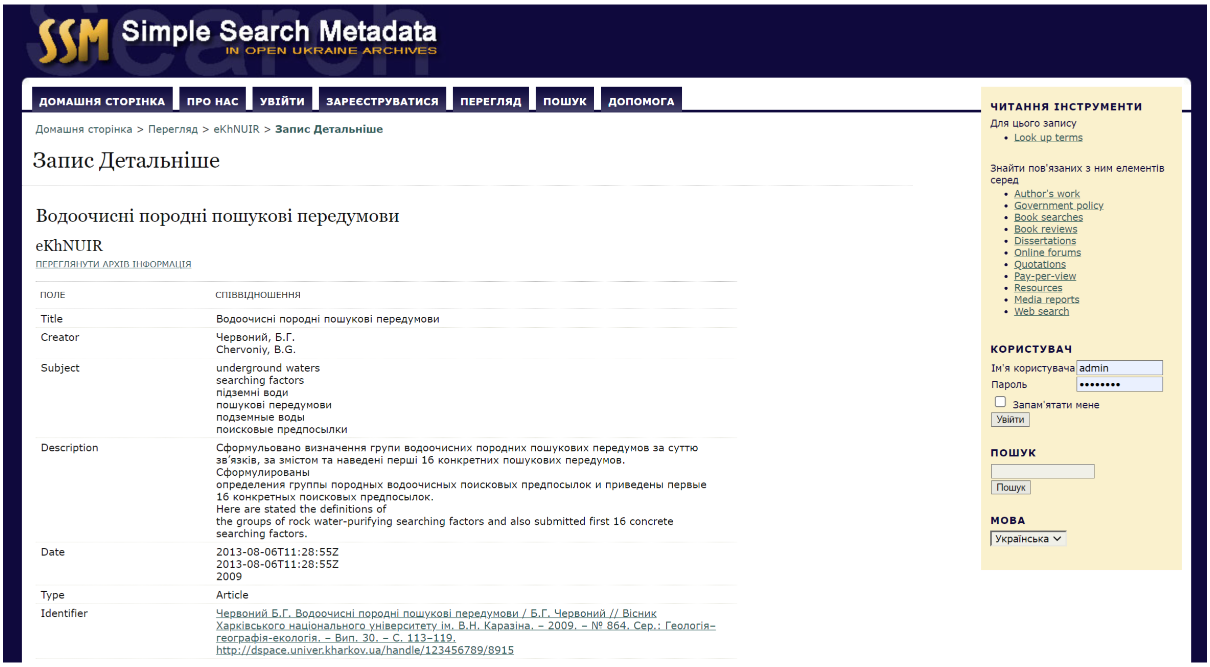
In open science, the volume of such data approaches that defined as Big Data. The primary feature of these data is their exponential growth. Many efforts are directed towards solving Big Data issues, requiring the development of new methods and algorithms for BD processing to address integration challenges. The definition of Big Data primarily relates to the difficulty of quantitatively defining a set of information objects. The most accepted definition is found in a report [8] wher e Big Data management issues are based on the three Vs: Volume, Velocity, and Variety. These represent the growth in data volume, the heterogeneity of data formats and metadata, complicating rapid data management. Later [9], veracity was added as a criterion to the Big Data definition, refining and supplementing criteria that affect data complexity and unstructuredness [10], [11]. However, semantics and structure are provided through external ontologies and fixed through metadata for semantic Big Data. This does not solve the operational issues of such data and creates additional problems related to extracting information from such BD sets. Our semantic data model should meet the requirements of Findable, Accessible, Interoperable, and Reusable data, known as the FAIR principle [12].

The semantic issues for Big Data based on the understanding that data semantics means the meaningful and efficient use of data objects to represent concepts or objects in the real world [13]. This broad concept encompasses various application areas [14]. Semantic knowledge of Big Data (BD) pertains to numerous aspects of rules, expert knowledge, and domain information [15]. A specific property of Big Data in the semantic environment is the complexity of inference, even when the data were not large initially. Thus, a characteristic feature of large semantic data is inference complexity, related to the speed of such operations. Internet web applications are highly sensitive to response delays, and web technologies require high-speed performance for Big Data.

Volumes of metadata that are aggregated in an open science project have the potential to approach big data, and therefore the ability of the system to operate with such data should be taken into account during design [16].

Metadata for repositories are crucial for organizing and providing access to digital resources. It involves creating, managing, and applying descriptive information about digital objects such as books, articles, images, audio files, and other digital content. The issue of metadata representation and exchange between libraries remains relevant, although significant progress has not been achieved over time. A distinctive class of integration systems is those based on the Open Archives Initiative (OAI) technology. In most known systems of this category, their information resources are collections of textual documents, primarily scientific publications, autonomously formed at global network nodes, maintained, and administered by their owners. Metadata aggregation for repositories is performed according to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), providing global access and search services [17], [18]. The essence of the open archives approach is to allow web access to information resources located in interoperable repositories by organizing shared use, publication, and archiving of metadata for such resources. The OAI-PMH protocol provides data providers with a simple way to present their metadata, making it accessible to service providers. HTTP is used as the transport protocol and XML as the data exchange format to address syntactic heterogeneity issues. However, in the OAI-PMH protocol, the Dublin Core format is specified to ensure a basic level of interoperability. Thus, metadata from various heterogeneous sources are combined in a single database to provide a range of services based on such aggregated metadata.

The OAI-PMH protocol concept distinguishes two parts: data provider and service provider. A data provider is a service that supports the creation and maintenance of one or more repositories (document bases, archives, electronic libraries), publishes its resources, and provides access to its metadata for use in other systems. A service provider collects and stores metadata provided by data providers to offer various services to end users. For a long time, a service provider for electronic libraries in Ukraine operated based on the PKP Harvester. The PKP Open Archives Harvester (PKP OAI Harvester) is a good tool for collecting metadata from various archives via the OAI-PMH protocol. This system allows metadata collection from DL in Ukraine, indexing 76 repositories with over 630 thousand records. The page listing electronic libraries is shown in Figure 1.


SSM Simple Search Metadata
IN OPEN UKRAINE ARCHIVES
ДОМАШНЯ СТОРІНКА
ПРО НАС
УВІЙТИ
ЗАРЕЄСТРУВАТИСЯ
ПЕРЕГЛЯД
ПОШУК
ДОПОМОГА
Домашня сторінка > Перегляд > eKhNUIR > Запис Детальніше
ПЕРЕГЛЯНУТИ АРХІВ ІНФОРМАЦІЯ
Увійти
Пошук
Українська

Запис Детальніше

Водоочисні породні пошукові передумови

eKhNUIR

ПОЛЕ	СПІВВІДНОШЕННЯ
Title	Водоочисні породні пошукові передумови
Creator	Червоний, Б.Г. Chervoniy, B.G.
Subject	underground waters searching factors підземні води пошукові передумови подземные воды поисковые предпосылки
Description	Сформульовано визначення групи водоочисних породних пошукових передумов за суттю зв’язків, за змістом та наведені перші 16 конкретних пошукових передумов. Сформулированы определения группы породных водоочисных поисковых предпосылок и приведены первые 16 конкретных поисковых предпосылок. Here are stated the definitions of the groups of rock water-purifying searching factors and also submitted first 16 concrete searching factors.
Date	2013-08-06T11:28:55Z 2013-08-06T11:28:55Z 2009
Type	Article
Identifier	Червоний Б.Г. Водоочисні породні пошукові передумови / Б.Г. Червоний // Вісник Харківського національного університету ім. В.Н. Каразіна. – 2009. – № 864. Сер.: Геологія–географія-екологія. – Вип. 30. – С. 113–119. http://dspace.univer.kharkov.ua/handle/123456789/8915

ЧИТАННЯ ІНСТРУМЕНТИ

Для цього запису

- Look up terms

Знайти пов'язаних з ним елементів серед

- Author's work
- Government policy
- Book searches
- Book reviews
- Dissertations
- Online forums
- Quotations
- Pay-per-view
- Resources
- Media reports
- Web search

КОРИСТУВАЧ

Ім'я користувача admin

Пароль ••••••••

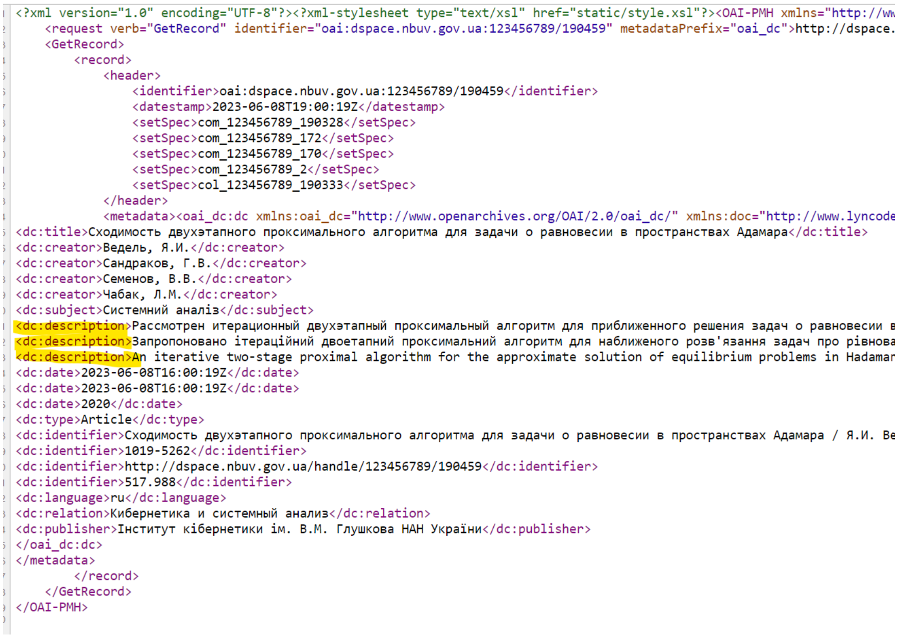
□ Запам'ятати мене

ПОШУК

МОВА

Figure 1: Page listing electronic libraries

PKP Harvester uses PHP 5.6, whose support has ended. This creates a problem for developing and creating new services. In developing software tools to support scientific research, it is necessary to provide not only metadata search but also extended capabilities for their processing and integration with other systems. At the same time, the goal is to create a system oriented towards working with Ukrainian data providers. This prompts the investigation of modern methods and solutions for creating environments that integrate DL. While many projects use the OAI-PMH protocol for data integration, such as BASE, OAIster, and CORE, Ukrainian electronic libraries are not fully represented in these aggregators. This is partly because the metadata language is predominantly Ukrainian. Not all electronic libraries properly provide multilingual metadata. For example, one of Ukraine's largest electronic libraries, the Scientific Electronic Library of Periodicals of the NAS of Ukraine, duplicates data such as resource descriptions without indicating the language, necessitating data processing, as shown in Figure 2. Such problems are common for multilingual repositories, which are generally common for multilingual countries.



```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="static/style.xsl"?><OAI-PMH xmlns="http://ww
    <request verb="GetRecord" identifier="oai:dspace.nbuv.gov.ua:123456789/190459" metadataPrefix="oai_dc">http://dspace.
    <GetRecord>
        <record>
            <header>
                <identifier>oai:dspace.nbuv.gov.ua:123456789/190459</identifier>
                <datestamp>2023-06-08T19:00:19Z</datestamp>
                <setSpec>com_123456789_190328</setSpec>
                <setSpec>com_123456789_172</setSpec>
                <setSpec>com_123456789_170</setSpec>
                <setSpec>com_123456789_2</setSpec>
                <setSpec>col_123456789_190333</setSpec>
            </header>
            <metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode
<dc:title>Сходимость двухэтапного проксимального алгоритма для задачи о равновесии в пространствах Адамара</dc:title>
<dc:creator>Ведель, Я.И.</dc:creator>
<dc:creator>Сандраков, Г.В.</dc:creator>
<dc:creator>Семенов, В.В.</dc:creator>
<dc:creator>Чабак, Л.М.</dc:creator>
<dc:subject>Системний аналіз</dc:subject>
<dc:description>Рассмотрен итерационный двухэтапный проксимальный алгоритм для приближенного решения задач о равновесии в
<dc:description>Запропоновано ітераційний двоетапний проксимальний алгоритм для наближеного розв'язання задач про рівнова
<dc:description>An iterative two-stage proximal algorithm for the approximate solution of equilibrium problems in Hadamar
<dc:date>2023-06-08T16:00:19Z</dc:date>
<dc:date>2023-06-08T16:00:19Z</dc:date>
<dc:date>2020</dc:date>
<dc:type>Article</dc:type>
<dc:identifier>Сходимость двухэтапного проксимального алгоритма для задачи о равновесии в пространствах Адамара / Я.И. Ве
<dc:identifier>1019-5262</dc:identifier>
<dc:identifier>http://dspace.nbuv.gov.ua/handle/123456789/190459</dc:identifier>
<dc:identifier>517.988</dc:identifier>
<dc:language>ru</dc:language>
<dc:relation>Кибернетика и системный анализ</dc:relation>
<dc:publisher>Інститут кібернетики ім. В.М. Глушкова НАН України</dc:publisher>
</oai_dc:dc>
</metadata>
        </record>
    </GetRecord>
</OAI-PMH>
```

Figure 2: Duplication of descriptive metadata without language identifier

The article discusses typical approaches to the integration of an electronic library. To achieve integration, it is necessary to determine which metadata will be integrated and which protocols should be used. Similar results were published in articles [19], [20], [21], [22] describing typical components of DL and the main issues related to standard approaches to architecture. In our work, we focused on the development of new methodologies for data integration.

The purpose of this research is to study methods and tools for developing an open science harvester for the NAS of Ukraine as a software tool for automatically collecting metadata of scientific periodicals of the NAS of Ukraine and information resources of NAS institutions. Integration can be based on metadata or a formalized content model. Typically, main protocols for scientific systems are based on metadata exchange, but these metadata can vary. Here are some main types of metadata used in DL:

- Descriptive metadata: provide basic information about a resource, such as its title, author, subject, keywords, abstract, and publication date. These metadata play a significant role in enhanced data exchange architecture.

- Structural metadata: describe the structure of the resource and the relationships between various components and set the relationship in the case when the resource consists of a set of other resources, for example, in the linked data model.
- Administrative metadata: include information about the management and administration of digital resources. It contains details about rights, permissions, access restrictions, file formats, file sizes, technical specifications, and preservation information.
- Archival metadata: are crucial for the preservation and archiving of digital resources. They include information such as the resource's origin, file format, checksums to ensure integrity, migration history, and other technical metadata. These metadata are necessary to ensure that digital objects remain authentic and accessible over time.
- Rights metadata: define intellectual property rights and usage permissions related to digital resources. They contain information about copyrights, licensing terms, usage restrictions, and citation requirements.
- Technical metadata: Provide information about the technical characteristics of digital resources. They include details about file formats, resolution, compression methods, color spaces, and other technical specifications necessary for rendering, reproducing, or processing digital content.
- Usage metadata: Track user interactions with digital items. They include information such as the number of downloads, views, ratings, comments, and user-generated content related to a specific resource.

All these types of metadata can be involved in DL integration. Various protocols and approaches are used to provide access to metadata through a single access point. Metadata work together to provide a comprehensive description of digital resources in a digital library, ensuring effective resource search, discovery, access, and management.

4. Data exchange protocols for digital libraries

Data exchange in a DL involves the transmission of information or resources between various systems, platforms, or repositories within the library ecosystem [23]. This process includes the sharing, importing, exporting, or synchronization of data to ensure that the DL collection remains current, accessible, and consistent across different platforms. In scientific data repositories, the integration process can be ensured at various levels:

- Metadata Exchange: Methods enable the exchange of metadata between DL, allowing them to discover, access, and retrieve resources from diverse sources. Standardization of metadata schemas such as Dublin Core, MARC (Machine-Readable Cataloging), or MODS (Metadata Object Description Schema) promotes interoperability and data exchange. This is facilitated by harvesting metadata and importing it into a data aggregator. An example is the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), which facilitates the collection and exchange of metadata, ensuring efficient data synchronization. Another example is the Journal Article Tag Suite (JATS) format [24], a standardized markup format for scientific articles in web publications. JATS is used for structuring and presenting metadata, text, references, and other elements related to scientific articles. It provides a unified format that facilitates the exchange, integration, and analysis of scientific data across different platforms and systems. JATS is based on the XML standard, which allows for easy processing and adaptation of data for various needs in processing and visualization.
- Distributed Search: Enables users to simultaneously search multiple DL or repositories and retrieve relevant results from each source. Data schema mapping is critical for distributed search when a request to each remote repository must be relayed to a local data schema.

Typical representatives of this protocol are Z39.50 or SRW (Search/Retrieve Web Service) [25].
- Data storage: research data and scientific results require long-term storage, which is ensured by the exchange between aggregation systems and data transfer to centralized data storage systems in various formats. A set of standards ensures the implementation of long-term storage, such as PDF/A.

Let's consider in more detail the OAI-PMH protocol, which is the most widely used protocol for metadata exchange in open access DL. The most common metadata schema supported in OAI-PMH is Dublin Core, which provides a basic set of elements for resource description [26]. The OAI-PMH protocol supports any metadata scheme and is effectively a transport protocol, but for basic compatibility, the Dublin Core (DC) metadata set is included. Typical scientific repositories that use software Dspace, Eprints, or OJS support MODS or METS also MARC is widely known for library catalogs.

Adjacent to the OAI-PMH protocol is the open archives initiative object reuse and exchange (OAI-ORE) protocol. Both protocols, developed by the OAI, aim to enhance interoperability and exchange of digital resources, albeit through different implementations. The OAI-PMH facilitates interoperability, which is not specific to any application, with a foundation for metadata harvesting. Its data model consists of three primary components:

- the resource, defined by one or multiple metadata records;
- each element of the metadata record is encoded in the form of an XML document, which is controlled by the XML schema. In this way, syntactic uniformity is achieved.;
- the item, a container for metadata records referring to a single resource, which must include at least one Dublin Core metadata record.

In contrast, OAI-ORE focuses on establishing a data model rather than defining an exchange protocol, suggesting potential exchange formats, such as XML/Resource Description Framework (RDF). The OAI-ORE model differentiates three types of resources:

- the aggregation, designed to group other resources referred to as aggregated resources;
- the aggregated resource, representing an information object within a compound object according to ORE standards; and
- the resource map, a serialized depiction of an aggregation that enumerates the aggregated resources and properties regarding the aggregation and its aggregated resources, including relationships with external resources [27], [28].

Since these protocols are very similar and developed under the same initiative, it is appropriate to provide examples of the differences between them, as shown in Table 1.

Table 1
Fundamental differences between protocols OAI-PMH and OAI-ORE

OAI-PMH	OAI-ORE
Intention	
OAI-PMH is designed to facilitate metadata harvesting and retrieval. It establishes a standardized protocol for repositories to provide access to metadata about their digital resources, thereby enabling other systems (harvesters) to retrieve and aggregate this metadata.	OAI-ORE focuses on the aggregation and exchange of complex digital objects. It aims to address the challenges associated with representing and disseminating digital objects composed of multiple interconnected components, such as web pages, multimedia files, annotations, and others.
Structure of data	

OAI-PMH primarily deals with metadata. It defines a protocol for exchanging metadata records, typically in standardized schemas such as DC or MODS, between repositories and harvesters.	OAI-ORE is concerned with the structure and representation of complex digital objects. It provides a framework for describing the relationships and aggregations of individual resources that compose an object, along with associated metadata.
Interoperability	
OAI-PMH promotes interoperability by standardizing the exchange of metadata. It allows repositories to expose metadata in a consistent format, facilitating the aggregation and harvesting of metadata by external systems.	OAI-ORE enhances interoperability by addressing the challenge of exchanging and reusing complex digital objects. It enables the representation of aggregations, annotations, and relationships among resources, rendering it easier to share and reuse complex digital objects across different systems and platforms.
Scope of use	
OAI-PMH is commonly used in DL, repositories, and archives for exposing and disseminating metadata about their collections. It enables metadata aggregation, search, and discovery across distributed repositories.	OAI-ORE is advantageous for managing complex digital objects that span multiple files, formats, or representations. It is useful in scenarios such as scholarly publishing, digital exhibitions, collaborative research, or multimedia collections, where understanding the relationships and structure of digital objects is essential.

Despite the simplicity of the protocol, this feature presents challenges for the semantic integration of data. This issue is exemplified by the OAI–PMH data model. The structure of the OAI-PMH XML documents adheres to a specific schema defined by the protocol.

Validating the XML of a document that describes metadata with the help of XLST is difficult because this mechanism does not provide an opportunity to semantically check the structure of the document. As shown at the top of the article, it is necessary to transfer metadata to a new level using semantic technologies such as RDF. However, the current OAI-PMH realization does not provide it. While facilitating the exchange of metadata, the OAI-PMH does not impose constraints on the semantics of the data being exchanged. This allows for the encapsulation of any type of data within a semantically independent structure. The protocol does not mandate the use of a specific vocabulary for DC elements, thereby allowing resources to be described with varying keywords. Furthermore, the XML schema does not enforce restrictions against duplicating elements or mandate the definition of language attributes, leading to potential issues when data providers submit duplicated data or data lacking language specifications. Semantic data integration aims to harmonize the meanings and interpretations of data elements across different repositories. This process extends beyond the mere exchange of metadata, striving to forge a unified data viewpoint by rectifying semantic discrepancies and establishing connections between disparate datasets. In the context of OAI-ORE, the OAI-PMH protocol is characterized as a transport mechanism that facilitates data delivery.

5. Data processing improvements for the OAI-PMH protocol

In view of the above-described OAI-PMH architecture concept, this paper proposes a semantic extension for data integration in DL [29]. As mentioned, OAI-PMH is a transport layer protocol that can be used only for data exchange, the solution to semantic integration should be implemented by improving the data received and expanding the connections between data. To achieve the schematic integration of metadata from heterogeneous sources, it is suggested to use ontologies of linked data and data mapping. Metadata alignment: usually works at the level of controlled dictionaries, in some cases metadata must be separated or combined, which requires transformation to achieve semantic integrity. Ontology mapping involves the creation of links between two or more source ontologies, and since the application domain is usually common to the metadata model, such mapping is quite simple. 3. Linked Data. Using linked data principles to connect and link data sets between repositories. Linked data allows you to establish explicit relationships between resources using standard protocols such as RDF.

Software processes that facilitate metadata integration are commonly known as extract-transform-load (ETL) processes. ETL (Extract, Transform, Load) are a process and tools that provide extraction of data from several sources, their transformation and normalization with customization, and insertion into a data store. The ETL process model looks like this:

Extraction (E): can be defined as a comprehensive search of all the data contained in the source system, leaving no data unaccounted for or missed during the extraction process.

Transformation (T) means changing the data structure to a new one: data cleaning, which includes removing or correcting errors, inconsistencies, duplicates, and missing values in the data; checking data for integrity, quality and data consistency with predefined rules or business logic; data enrichment, which includes enhancing data by adding additional information, derived attributes, or calculated values based on rules or external data sources; data filtering: selecting or excluding certain data based on predefined criteria.

Loading (L): This process involves loading the transformed data directly into the target system or database using native loading mechanisms or APIs.

It is appropriate to apply these processes when solving the problem of semantic integration within the framework of the OAI-PMH protocol.

6. Overview of software for data integration according to the OAI-PMH protocol

Several software solutions are available for integrating OAI-PMH, allowing organizations to harvest and display metadata from various repositories. We conducted a brief analysis of popular software for creating an OAI-PMH harvester and identified key features for creating an effective scholarly environment. The following requirements were set for forming the list: open-source license, system support and updates, historical system stability, and high-quality system architecture utilizing modern design methods. Below is a comparative table of characteristics of popular software for integrating OAI-PMH DL.

Table 2
List of modern software for integration of scientific repositories

Software	**VuFind**	**DSpace**	**Omeka S**
Description	VuFind is a software that supports the collection and integration of repositories through OAI-PMH, which allows collecting metadata from multiple sources and providing a unified search.	DSpace is a digital storage platform that supports the OAI-PMH protocol.	Omeka S is software for creating digital collections. Includes OAI-PMH support, allowing metadata to be exchanged with other OAI-PMH-compliant systems.
Stack	PHP, MySQL, SORL	JAVA, MySQL	PHP, MySQL
ETL	Yes	No	No
Support of library catalog systems	Yes	No	No
Faceted navigation	Yes	Yes	Yes
Filtering of received records/filtering	Yes	Yes	Yes

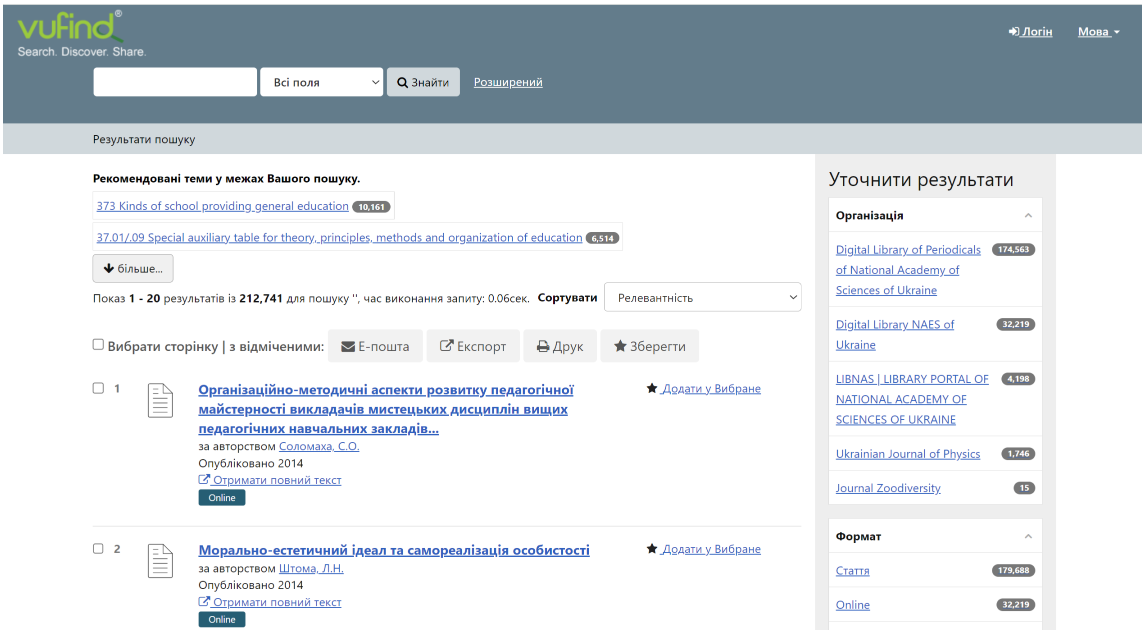
of searches			
Recommendation system in the user interface	Yes	No	Yes
Connection of the full-text extractor	Yes	No	No
Full text search	Yes	Yes	Yes
Fuzzy search	Yes (Sorl)	Yes (Sorl)	Yes (Elasticsearch)
User roles	Yes	Yes	Yes
LDAP authentication	Yes	Yes	Yes
Creating a cover page	Yes	No	No
DOI	Yes	Yes	Yes
EZproxy	Yes	Yes	Yes
Spelling to search	Yes	Yes	No
Export to OAI-PMH	Yes	Yes	Yes
Configuration-based interface	Yes	Restricted	No
Multilingual support for metadata	Restricted	Yes	No
Matomo analytics	Yes	Yes	Yes
API	REST API	REST API	REST API
Metadata Schemas (Import and View)	Dublin Core, METS, Dublin Core Terms, MARC, XML, CSV	Dublin Core, Dublin Core Terms	Dublin Core, METS
Editing metadata	No	Yes	Yes
Web interface for resource management	No	Yes	Yes
Autocompletion when searching	Yes	No	No

When evaluating characteristics, it is important to understand that objectively comparing a number of parameters is very challenging. For example, in VuFind [30], [31] the system architecture is designed in such a way that metadata display management is completely controlled through theme modifications. In other systems, managing the presentation requires simple changes in the system settings through a graphical interface, as implemented in DSpace. Adding new metadata fields in Omeka S requires creating a plugin. Therefore, while the complexity of manipulating the data model in DSpace is the simplest, considering flexibility and speed, VuFind has the advantage. Overall, after analyzing the capabilities of each system, VuFind is a more successful solution for achieving the research objectives.

VuFind is software for creating a library resource portal, primarily aimed at improving user interaction by transforming the traditional online public access catalog (OPAC) [32]. This platform is an open-source library search engine developed by Villanova University's library, first released to the public in a stable version in 2010. The software architecture of this product is very well implemented, achieved through a developer-oriented toolkit, the Laminas framework, and a large number of system settings. This allows for changing the structure of metadata to be displayed to the user without needing to modify the system's code. The formatting rules of an object are controlled from the theme's code, establishing rules for which metadata to use and which system methods will handle data retrieval. This is sent to the backend, and after processing, the result is returned to the interface for user display. Thus, in the system's architecture, data and the formatting rules for this data are separated, which is very convenient for customization.

Let's examine in more detail how the ETL method is implemented in VuFind. In this software product, the data transformation process is actually divided into two stages. Upon receiving data, the initial transformation of metadata occurs to change the record identifier from the archive. This is due to the need for unique identifiers and, on the other hand, the identifier structure should not contain slashes. Since each resource has a URL corresponding to its identifier in the primary electronic library, which is the source of metadata. Based on these advantages, this software product was chosen as the foundation for creating an integrated environment to support scientific research.

The result of deploying and indexing the scientific electronic library of periodicals of the NAS of Ukraine is presented in Figure 3.



Рекомендовані теми у межах Вашого пошуку.

373 Kinds of school providing general education 10,161

37.01/.09 Special auxiliary table for theory, principles, methods and organization of education 6,514

Показ **1 - 20** результатів із **212,741** для пошуку '', час виконання запиту: 0.06сек. **Сортувати** Релевантність

□ Вибрати сторінку | з відміченими:

□ 1 **Організаційно-методичні аспекти розвитку педагогічної майстерності викладачів мистецьких дисциплін вищих педагогічних навчальних закладів...**
за авторством Соломаха, С.О.
Опубліковано 2014

□ 2 **Морально-естетичний ідеал та самореалізація особистості**
за авторством Штома, Л.Н.
Опубліковано 2014

Організація	
Digital Library of Periodicals of National Academy of Sciences of Ukraine	174,563
Digital Library NAES of Ukraine	32,219
LIBNAS | LIBRARY PORTAL OF NATIONAL ACADEMY OF SCIENCES OF UKRAINE	4,198
Ukrainian Journal of Physics	1,746
Journal Zoodiversity	15

Формат	
Стаття	179,688
Online	32,219

Figure 3: Resource list interface with a faceted filter

Semantic data integration is not provided in VuFind by default, but it can be achieved through user functions that can map semantic data. One of the advantages of VuFind is the ability to use such calls. The process of integration and organization of access to information in VuFind consists of the following stages:1) Metadata collection using the OAI-PMH protocol; 2) Data transformation according to the ETL model. At the extraction stage, VuFind allows for partial transformation operations. This process enables VuFind to create a unified and comprehensive index of resources from many sources, providing users with centralized search capabilities; 3) User search on aggregated data using a convenient interface with deep configurability; 4) Resource access. Each resource is directly accessible through provided links, including necessary identifiers (e.g., URLs or DOIs) to the full content hosted by the original data providers; 5) Metadata display. The collected metadata is presented in a standardized ETL process and a user-friendly format. This can enrich the metadata with additional information or aspects to improve search and assist in resource discovery.

7. Conclusions

To enhance the visibility of research results from NASU (National Academy of Sciences of Ukraine) scientists in the open science information environment, modern software tools are required. This will ultimately enable the evaluation of effectiveness through scientometric indicators, promoting the development of science in Ukraine and international scientific cooperation.

Building an integrated environment for the aggregation of scientific resources requires addressing a number of challenges. The article discusses approaches to the integration of electronic archives and describes the practical experience of integrating Ukrainian electronic archives using the OAI-PMH protocol.

The construction of an integrated environment for the aggregation of scientific resources requires solving several problems. The article examines approaches to the integration of electronic archives and describes the practical experience of integrating Ukrainian electronic archives using the OAI-PMH protocol. The main protocols for the integration of electronic libraries are considered. As the analysis has shown, since 2015, no significant exchange protocol alternative to OAI-PMH has emerged. Approaches to the structural integration of electronic libraries have been analyzed, and a comparative analysis of the functional capabilities of each software has been carried out. It has been shown that VuFind is the most effective tool for the integration of DL.

Future research will focus on the integration of full-text search and the improvement of metadata quality [33] and display through semantic technologies such as ontologies and linked data.

References

[1] V. O. Kopanieva, L. I. Kostenko, O. V. Novytskyi and V. A. Reznichenko, "The task of digital transformation of the scientific information environment," *Problems in programming,* vol. 1, pp. 3-10, 2023.

[2] W. M. Beyene, "Metadata and universal access in digital library environments," *Library Hi Tech,* vol. 35, no. 2, pp. 210-221, 2017.

[3] Daniela Florescu, Ioana Manolescu, Donald Kossmann, "Answering XML queries over heterogeneous data sources," in *27th International Conference on Very Large Data Bases (VLDB 2001).*

[4] Leonidas Galanis, Yuan Wang, Shawn R. Jeffery, David J. DeWitt, "Locating data sources in large distributed systems," in *29th International Conference on Very Large Data Bases (VLDB 2003)*, Morgan Kaufmann, 2003.

[5] M. Lenzerini, "Data integration: a theoretical perspective," in *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2002)*, New York, 2002.

[6] A. Y. Levy, "Combining artificial intelligence and database for data integration," in *In Artificial Intelligence Today: Recent Trends and Developments*, Berlin/Heidelberg, 1999.

[7] Alon Y. Levy, Anand Rajaraman, Joann J. Ordille., "Querying heterogeneous information sources using source descriptions," in *22nd International Conference on Very Large Databases*, Bombay, India, 1996.

[8] D. Laney, "3D data management: Controlling data volume, velocity and variety," META group, 2001.

[9] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales and P. Tufano, "Analytics: The Real-World Use of Big Data," IBM, 2012.

[10] I. C. Intel IT Center, "Centre. Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data," Santa Clara, 2012.

[11] S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning.," *ACM SIGMETRICS Performance Evaluation Review,* vol. 41, no. 4, pp. 70-73, 2014.

[12] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, P. Bourne and J. Bouwman, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data,* pp. 1-9, 2016.

[13] P. Ceravolo, A. Azzini, M. Angelini, T. Catarci, P. Cudré-Mauroux, E. Damiani, A. Mazak, M. Van Keulen, M. Jarrar, G. Santucci and K. Sattler, "Big data semantics," *Journal on Data Semantics,* vol. 7, no. 2, pp. 65-85, 2018.

[14] R. Amsler, "Application of Citation-based Automatic Classification," Austin, 1972.

[15] W. A. Woods, "What's in a link: Foundations for semantic networks.," *Representation and understanding,* pp. 35-82, 1975.

[16] S. Roy, B. Sutradhar and P. Das, "Large-scale Metadata Harvesting—Tools, Techniques and Challenges: A Case Study of National Digital Library (NDL)," *World Digital Libraries: An International Journal.,* vol. 10, 2017.

[17] H. Van de Sompel, M. Nelson, C. Lagoze и S. Warner, «Resource harvesting within the OAI-PMH framework,» *D-lib magazine,* № 10, 2004.

[18] "The Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 of 2002-06-14," [Online]. Available: http://www.openarchives.org /OAI/2.0/openarchivesprotocol.htm.

[19] R. Gartner, Metadata for digital libraries: state of the art and future directions, JISC, 2008.

[20] A. Getaneh, B. Stevens and P. Ross, "Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: A social constructivist approach," *New Library World,* vol. 113, pp. 38-54, 2012.

[21] К. Лобузіна, "Сучасні підходи до інтеграції електронних інформаційних ресурсів бібліотек," *Вісник Книжкової палати,* vol. 12, pp. 24-28, 2012.

[22] О. М. Спірін, С. М. Іванова, О. В. Новицький, З. Савченко, В. А. Резніченко, А. В. Яцишин, Н. М. Андрійчук and В. Ткаченко, Електронні бібліотечні інформаційні системи наукових і навчальних закладів., Педагогічна преса, 2012.

[23] M. Agosti, N. Ferro and G. Silvello, "Digital library interoperability at high level of abstraction," *Future Generation Computer Systems,* vol. 55, pp. 129-146, 2016.

[24] National Center for Biotechnology Information, U.S. National Library of Medicine, "Journal Article Tag Suite," 2024. [Online]. Available: https://jats.nlm.nih.gov/. [Accessed 10 2024].

[25] A. S. Lingam, "Federated search and discovery solutions," *IP Indian J. Libr. Sci. Inf. Technol.,* Vols. January-June 5, no. 1, pp. 39-42, 2020.

[26] C. Lagoze and H. Van de Sompel, "The Open Archives Initiative Protocol for Metadata Harvesting," 2015. [Online]. Available: http://www.openarchives.org/OAI/openarchivesprotocol.html.

[27] C. Lagoze and H. Van de Sompel, "ORE User Guide - HTTP Implementation," [Online]. Available: https://www.openarchives.org/ore/1.0/http. [Accessed 2023].

[28] C. Lagoze and H. Van de Sompel, "ORE User Guide - Resource Map Implementation in RDF/XML," [Online]. Available: https://www.openarchives.org/ore/1.0/rdfxml. [Accessed 2023].

[29] В. А. Резніченко, О. В. Новицкий and Г. Ю. Проскудіна, "Інтеграція наукових електронних бібліотек на основі протоколу OAI-PMH," *Проблеми програмування,* no. 2, pp. 97-112, 2007.

[30] Villanova University's Falvey Library., "VuFind® - Search. Discover. Share.," [Online]. Available: https://vufind.org/. [Accessed 2023].

[31] D. Katz, R. LeVan and Y. Ziso, "Using authority data in VuFind," *Code4Lib Journal,* vol. 14, 2011.

[32] H. Yu and M. Young, "The impact of web search engines on subject searching in OPAC," *Information technology and libraries,* vol. 4, no. 23, pp. 168-180, 2004.

[33] O. Novytskyi, G. Y. Proskudina, V. Reznichenko and O. Ovdiy, "Evaluation of the quality of digital libraries in the web environment," *Software engineering,* vol. 20, no. 4, 2014.