# Towards a new generation of Digital Repository of Scientific Institutes

**Tomasz Parkoła and Błażej Betański**
Poznan Supercomputing and Networking Center

**Abstract:** Digital Repository of Scientific Institutes (DRSI) was established in 2010 as a collective initiative of 16 scientific institutes from Poland. The main objective of this initiative is to provide for the research community a country-level, multidisciplinary digital repository composed of archival materials, scientific publications, research documentation and cultural heritage content. DRSI is maintained by the Polish scientific institutes, forming the DRSI Consortium. In 2018 a country-level, EU co-funded project named OZwRCIN, was initiated in order to improve DRSI. The project sets out several objectives that can be reduced to a general statement about sharing, archiving and disseminating public sector information. More than 140 thousands of digital objects will be provided in the scope of the project, using new technologies and innovative approaches. The project specifically aims at providing research datasets. To manage digitisation activities and make the digital content available online the DRSI uses infrastructure and tools developed by Poznan Supercomputing and Networking Center (PSNC). PSNC is an ICT-focused R&D institution, infrastructure provider and operator of the Polish Optical Internet PIONIER. PSNC is a technical partner in the OZwRCIN project. This article discusses in details the key concept of the DRSI, its infrastructure, the software stack as well as organisation of the consortium itself.

**Keywords:** large-scale digital repository, digitisation workflow, long-term preservation.

## 1. Introduction

In 2010 a collective initiative of 16 scientific institutes from Poland has established the Digital Repository of Scientific Institutes (DRSI) and soon received EU funds to conduct a large-scale project named RCIN (eng. DRSI).[1] The assumed indicators of the project have been achieved, and even exceeded significantly. The database has gained great popularity among end-users. By the end of December 2016 over 25 million readers have used DRSI resources, the scientific community in particular. The DRSI board has started to receive applications from successive publishers of scientific journals and institutes wishing to join the DRSI and deposit materials in its database. Unfortunately, the IT infrastructure was provided for 16 institutes and did not allow to deposit and archive collections of several dozen more entities. There was also a need for digitisation and assuring online access to other scientific resources, such as

Received: 6.7.2020 Accepted: 8.9.2020 ISSN 2241-1925

collected research data (inventoried botanical and zoological specimens, elements of the cultural landscape, inanimate nature), historical manuscripts or maps for further scientific research and catalogues. In order to respond to all of the needs of existing members of the DRSI consortium, new interested scientific institutes as well as end-users, it was decided to apply for a new project in the Operational Programme Digital Poland for 2014-2020 [2]. The application was successful and resulted in a country-level, EU co-funded, project named OZwRCIN to improve DRSI in many aspects.

The main objective of the OZwRCIN project is to increase the volume and accessibility of the unique scientific resources coming from 23 institutes. It will be done mainly through digitisation, modernisation and improvement of the DRSI platform, but also  use of existing ICT infrastructure for archiving, processing and providing access to this information.

The OZwRCIN project will make more than 140,000 digital objects available under free licences in the coming years, of which more than 80% is source data / research data. That data includes sets of fauna and flora specimens, archaeological finds, geographical data, documentation of research and experiments (including photos e.g. microscopic photos), records, plans, exaggerations, inventories. Obviously, the platform for sharing these resources is DRSI. Such a wide and diverse set of data required the development of an appropriate data model that would enable its safe storage and effective sharing. As a part of the project work, key changes are being made to the DRSI portal in terms of adapting it to the requirements of new types of research data. These changes concern both the content and metadata storage model. Changes were also necessary in the presentation of shared facilities, e.g. the use of IIIF streaming or hybrid data formats, as well as the development of an approach for the unique identification of individual resources using global identifiers.

## 2. Project purpose and target content of the DRSI

The aim of the project is to increase the volume, availability and use of scientific resources, by making new resources available under free licenses or in the public domain, improving the functionality and development of DRSI, as well as launching new services.

All resources digitised and provided online during the project will be available at no charge and based on free licences (e.g. CC BY, CC BY-SA or CC Public Domain Mark). The accompanied metadata will also be available under a free license. This will ensure that both the content of the items and their descriptive metadata will be possible to be reused without any restrictions, not only by individual internet users and companies but also by external (public) databases. The resources made available under the Project can be divided into:

- Scientific resources like monographs, abstracts, doctoral theses, posters, reports, maps and atlases, as well as historical items, including manuscripts, used in scientific research;
- Source data such as assemblies of fauna and flora specimens, archaeological finds, geographical data, etc., collected during field studies;
- Documentation of research and experiments: descriptive documentation, photographs (e.g. microscopic pictures), records, plans, tracings, inventories, etc.

In addition, taking into consideration the format of the data provided, it can be divided into: text resources, graphic resources and hybrid resources, i.e. composed of both textual and graphic elements. More than 80% of the resources that will be made available will consist of digitised unique research data or manuscripts of famous people that are not owned by any other institution. Therefore the project will provide a comprehensive information system, especially in the field of life sciences (biology, biotechnologies, medicine), earth sciences or humanities. With regard to humanities and social sciences, the database will become an important source of cultural heritage datasets. It will potentially enable study of large datasets in order to identify trends, patterns and relationships that cannot be captured through analysis of individual assets.

From the technical point of view DRSI will be upgraded and existing software will be adapted to digitise, archive and share data according to the users' needs. The availability of digital content will be increased in terms of search and readability on various types of devices. Functionality of software managing the digitisation process will be extended by dedicated mechanisms for handling specific types of objects. Automatic data enhancement and automatic verification of the input data quality will be possible. Long term archiving system will be extended with standards, protocols and formats for storing and exchanging information e.g. OAI-PMH, OAI-ORE, METS, PREMIS. A Knowledge base system will be launched to aggregate information of the scientific activities of individual Partners.

The project is also focused on improving back-end routines executed in the context of DRSI. In this regard, various adaptations of the digitisation workflow management and long-term preservation systems will be conducted. It is especially important in terms of dedicated mechanisms for handling specific object types (e.g. research data). The work will also cover automated data enhancement and validation to improve interoperability and quality of the ingested data.

In addition, in scope of the project, it is planned to launch Current Research Information System for each institute that participates in the project. Thanks to that scientific activities of individual Partners will be made available for on-line users.

## 3. Key technical challenges

The technical challenges were directly driven by user needs and concerned issues related to the increased number of digital objects on the DRSI portal, technological changes as well as digital content accessibility needs. The main challenges are as follows:

- Increasing number of digital objects and related source data (e.g. master files, source data) required appropriate scaling up of the existing technical solution and its extension as well as providing appropriate ICT resources.
- New types of objects e.g. source data that include specimens of flora and fauna which are being digitised required dedicated mechanisms that, on the one hand, ensure that such digital content is sufficiently accessible and easy to use for end-users and, on the other hand, enable resource providers to enter this data into the DRSI portal.
- Current DRSI resources which due to their technical characteristics (e.g. file format) were considered to be poorly accessible and not meeting the interoperability requirements, required action to increase their accessibility.
- The modernization of existing institutes' portals and modernization of the main DRSI portal. The work included creation of knowledge bases related to the scientific activities of individual institutes.
- The development of mechanisms related to the automation of the digitalization process in accordance with the requirements of the new institutes that cooperate in the project required appropriate adaptation of the software used to manage the process of digitalization, archiving and sharing resources.
- Ensuring high accessibility of data available in the Project. In particular it concerns modernization of web interface in accordance with accepted practices of user-oriented design and high availability standards e.g. WCAG 2.0 or Linked Open Data.
- Providing interoperability of the data created in the Project in the context of existing external platforms, e.g. Europeana, CrossRef or Google Scholar.

## 4. End users

The target group of the Project are internet users. The results of the online survey that was conducted from 15.05.2016 to 10.10.2016 showed (Figure 1) that the existing DRSI resources were used by the respondents for: scientific work, educational purposes, other professional activity and for hobby purposes.

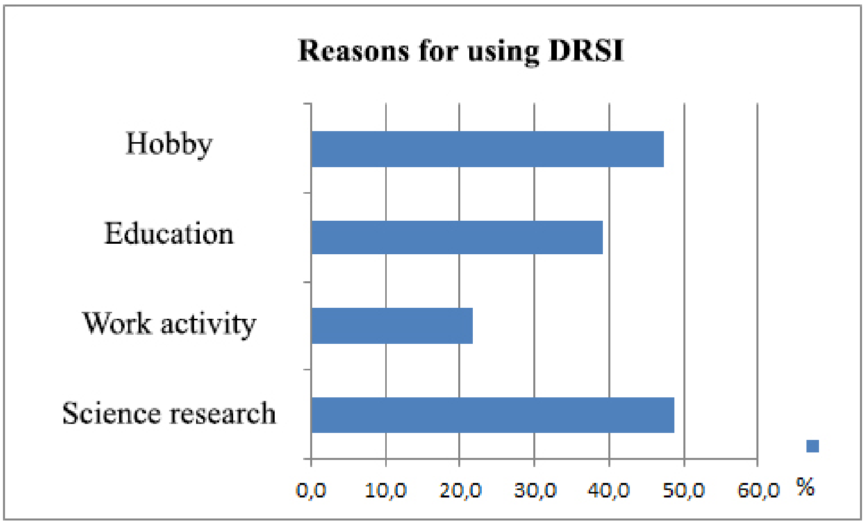

***Figure 1. DRSI survey results - question about the reasons for using DRSI.***

The survey also showed (Figure 2) that the portal is used by academics, PhD students, students, librarians, professionals from art and culture sectors, representatives of enterprises as well as a large group of people describing themselves as hobbyists.

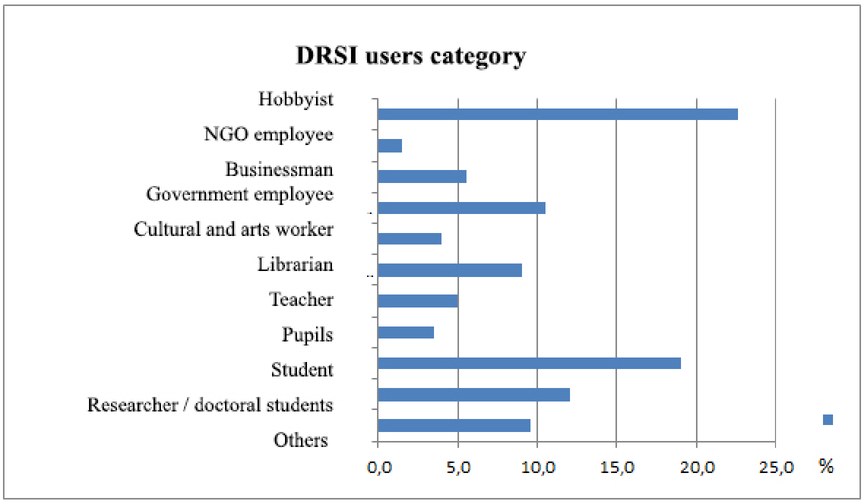

***Figure 2 DRSI user survey results - question about the user category.***

We can also specify two main types of users visiting the portal:

1. Users of digital resources that are available in the DRSI portal;
2. Resource providers, including both those developing services and those preparing digital content for ingestion into DRSI.

Among the users of digital resources provided through the DRSI portal, there are three user segments:

1. The scientific community, including:
    a. employees and doctoral students of Partner institutes;
    b. scientists from institutions cooperating with Partners;
    c. employees of other Polish and foreign scientific institutions;
2. Other internet users, including:
    a. students, pupils, teachers;
    b. professionals from art and culture,
    c. administration employees
    d. hobbyists
    e. librarians and information workers
    f. representatives of enterprises
    g. local government officials
    h. NGO employees
3. Services aggregating data from DRSI, e.g. Google Scholar [4], Europeana [5], FBC.[6]

From the survey it is visible that the scientific community uses DRSI not only to obtain information related to their work but also to make their publications available online with a high search rank. Internet users e.g. hobbyists look for information related to their interests and specific materials related to their hobbies. Students look for information in a specific subject area, related to the topic of their studies and aligned with their current assignments. These kind of analysis and other conclusions from the survey directed technical work planned in the project. In general, improvements apply to the search and view features, including build-in support for various types of devices (e.g. mobile) as well as better accessibility aligned to WCAG [7]. For future improvements, the repository will also be monitored in terms of various statistics and user experience, e.g. usage of resources, user behaviour or user satisfaction.

## 5. Requirements of the content providers

The technical work in the project is aimed at adapting and developing further the software solutions used by DRSI, as well as increasing the availability of scientific materials held by the Partners cooperating in the project. Technically, DRSI is composed of main portal and one sub-portal for each cooperating institute. This way there is one entry point for the end users that want to search & browse DRSI content. On the other hand, each institute can also promote their content via their own institutional repository. Therefore data ingested by the project partners are available both on the main portal as well as on respective, individual sub-portal of a specific institute. It is important to note that data is uploaded only once and then it is accessible in the main portal and respective sub-portal.

The process of digitization and sharing of objects in DRSI is supported by the DInGO toolset developed by PSNC in cooperation with representatives of the DRSI consortium. DInGO allows users to monitor and control the workflow related to digitization, long-term data storage and the process of making digital objects available on-line. DInGO allows users to handle the digitalization process in accordance with international standards of good practice, e.g. data archiving is carried out according to the OAIS model, or a self-archiving feature is organised according to the green Open Access path, or relevant mechanisms for automated verification of data quality (e.g. verification of technical parameters of ingested data). Compliance with international open standards and good practices is essential to ensure that users can access the data provided through the DRSI portal in the long term. The key technical challenges identified in the project are directly related to the users' needs and include:

- Increased number of digital objects in the DRSI portal and the ongoing technological changes, especially accessibility needs.
- Increased number of objects uploaded to the system and related source data (e.g. master files, source data) that require appropriate scaling of the technical solution and implementation of relevant features.
- New types of objects (e.g. source data - including specimens of flora and fauna) which are to be digitised. They require dedicated mechanisms that, on one hand, ensure that the users of the resources have straightforward access to digital content, and on the other hand, enable content providers to ingest data into the DRSI portal.
- Improvement of the digitisation process in accordance with the requirements of the new institutes involved in the project, including adaptations of the digitisation workflow management software as well as long term preservation environment.

Improved accessibility of the data provided via DRSI will be ensured by application of best practices of user-centered design and well-known accessibility guidelines, i.e. WCAG 2.0 or "5 Star Open Data" principles. There will be also dedicated mechanisms used to provide DRSI data to external aggregation platforms like Europeana or Google Scholar. It is done using global standard and accepted practices such as OAI-PMH [3] protocol or Sitemap format.

Based on the analysis conducted among content providers it is clear that the digitalization staff requires the possibility to work in batch mode i.e. prepare a set of data during the day and then ingest the information to the server at once at a later stage or execute the ingestion process in the background. Therefore improvements in the usability of the user interface are important, e.g. by performing certain operations simultaneously for many digitised objects. In addition, it is planned to introduce automated verification of technical parameters of the ingested master files (source data).

For managers of the digitisation workflows it is important to have a high-level information on technical parameters of the data stored in the system. This is crucial in the context of the inventory of owned digital resources and for ensuring their availability to target users in the long term. In terms of security it is important to replicate data in various storage systems.

Scientific managers of the partnering institutions expressed the need for tools that could support reporting of the scientific achievements to central units or for internal purposes. It would also help to make information about the scientific achievements of the institution available to the public. Therefore, in the scope of the project, it is planned to launch a CRIS system for each partner. The system will include features enabling the preparation of relevant reports, statements and summaries.

## 6. OZwRCIN technical solution architecture

To meet the essential needs of individual user groups, with a particular focus on aspects related to the high availability of digital resources, the technical work covers two key areas. First concerns the use of existing hardware and software infrastructure and services available at PSNC. The second area of technical work is related to developments and adaptations of the existing toolset for digitalisation, long time preservation and online publishing.

DRSI is built using well-established DInGO toolset, that has been developed by PSNC since 2002, and which currently has more than 140 deployments in Poland and abroad. The logical division of the DInGO toolset components within OZwRCIN is shown in Figure 3.

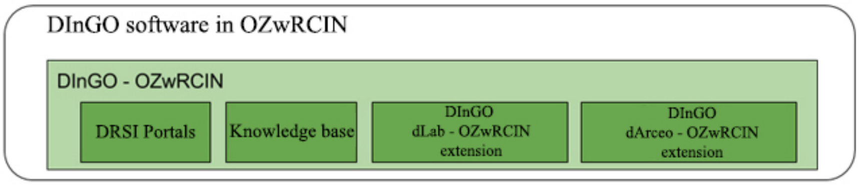
DInGO software in OZwRCIN
DInGO - OZwRCIN
DRSI Portals
Knowledge base
DInGO dLab - OZwRCIN extension
DInGO dArceo - OZwRCIN extension

*Figure 3 The logical division of DInGO software package components.*

In DRSI portals a new user interface has been designed and developed, with the use of Responsive Web Design techniques, to allow access to data from various devices. The compliance of individual portals with WCAG 2.0 requirements and Linked Open Data practices (based on "5 Star Open Data"[8]) was also ensured. Functions increasing the availability of resources have been developed for existing digital objects, e.g. for those created in legacy DjVu format. The mechanisms include: displaying objects in DjVu format with the use of HTML5 technology and development of mechanisms enabling presentation of an object using various (alternative) file formats. Dedicated mechanisms for displaying objects with specific characteristics have been developed or are under development. This particularly applies to objects such as: maps, herbariums or 3D objects. The metadata scheme of digital objects available in DRSI has been adapted to the needs of new Partners cooperating in the Project, e.g. the possibility to provide metadata using Darwin Core scheme. Related search mechanism has also been adapted to the DRSI metadata schema.

There are also improvements in the context of particular metadata fields, for instance, tagging with geolocation will be possible and hence the display of digital objects on the map will be possible. It will allow users to see direct location of the object as well as related objects available nearby. The ability to quote an article directly from the DRSI website will be added, e.g. displaying ready-to-copy quotes in PN-ISO 690:2012, APA, MLA, CSV and Chicago standards, as well as the ability to download one or several quotes at once in ENW, RIS and BIB format from the preview of object information and the list of search results. This mechanism will be supplemented with support for popular bibliography management systems like Zotero.

The work related to DInGO dLab - OZwRCIN extensions aims to increase ergonomics in the context of handling the digitisation process. Thanks to the development of the batch ingest feature, the process of digitisation has been improved and allows transferring large volumes of data at a convenient moment, e.g. at the end of a day or overnight. In addition, the extensions will ensure proper quality of resources processed in the digitisation workflow by validating ingested master files, e.g. using the JHOVE tool. The quality profile may contain information about technical parameters of the files, e.g. for TIFF no compression or LZW compression, indicated colour depth or scanning resolution. The system will also assure proper execution of individual activities related to the digitisation workflow, e.g. through communication with external systems to validate data that should be available as a result of such activity. Availability and interoperability of data will be also increased, by named entities recognition using existing tools and linking such resources to existing databases, e.g. Viaf.

DInGO dArceo component is responsible for increasing the security of data in terms of its long-term storage. This will be done by introducing functions related to the analysis of stored master files. It particularly concerns the analysis of formats, time of creation of files, correctness of files, the program used to create the files, etc. The data validation function in the data storage system will be extended by the possibility of validation against the file storage policy applied by the Project (e.g. technical parameters of files, file structure, etc.). The function of source data storage system for creating data replicas in independent data storage systems, e.g. storage on a disk array and storage in a cloud service, will also be added.

The Knowledge base will aggregate the scientific activities of the participating institutes. Relevant data model has been already developed and the knowledge base will be integrated with DRSI repository portals. The integration will allow e.g. to enrich the description of the objects with information from the knowledge base (additional information about the authors, the institute, etc.) or to present related articles (e.g. same author, same issue of the journal, same conference). Data will be imported into the knowledge database from the existing systems of individual institutes.

The logical architecture of the components of DRSI portals and knowledge bases is shown in Figure 4. The assumption for repository portals is analogous to the current DRSI architecture. The main DRSI portal presents all the scientific resources made available by individual Project Partners. Additionally, each Project Partner has its own portal - an institutional repository (logically separated from the main database) - where resources only of a given Partner are presented. This means that within the DRSI platform, and within the Project, there will be 17 portals - one main and 16 portals for each of the Partners. In order to enable the presentation of information about the scientific achievements of each of the Partners participating in the Project, the knowledge bases were launched. Each Partner will be able to enter data into their own knowledge base, which will be linked to their DRSI repository portal. In general, information about scientific achievements of the employees of a given Partner will be stored in the knowledge base, and some of this information will be presented in the repository portal DRSI of the Partner together with access to the full content of the scientific article.

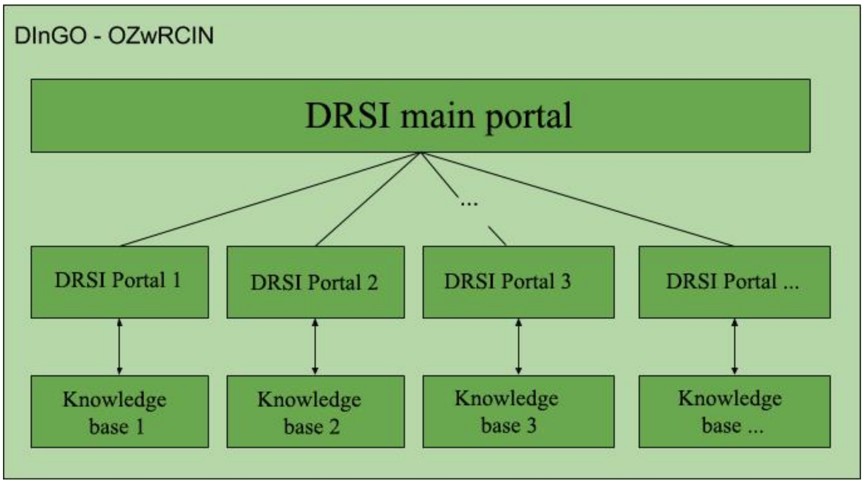

*Figure 4. The logical architecture of the components of DRSI portals and knowledge bases.*

## 7. Conclusions

Although the project includes institutions representing various fields of science, DInGO toolset has made it possible to present their collections and digital objects in a unified form with the preservation of their individual charisma in a clear and convenient way to browse. The current stage of the OZwRCIN project is on a good way to being as successful as it was in the previous project. The main goals like increasing volume and accessibility of unique scientific resources are continuously undertaken with support of the DInGO toolset. New developments are underway, based on discussions with representatives of the project Partners. Rich experience in developing tools for online presentation of cultural resources and constant desire to improve existing solutions should result in further ideas setting new directions of development for tools supporting digital repositories.

**References**

**Web Site.**

[1]rcin.org.pl, Retrieved 4 May 2020, from
https://rcin.org.pl/dlibra/text?id=Projekty&language=en

[2] Portal Funduszy Europejskich (2014, December 5) Operational Programme Digital Poland for 2014-2020 Retrieved 4 May 2020, from
https://www.funduszeeuropejskie.gov.pl/media/1655/POPC_eng_1632015.pdf

[3] Open Archives Initiative Protocol for Metadata Harvesting Retrieved 4 May 2020, from https://www.openarchives.org/pmh/

[4] Google Scholar Retrieved 4 May 2020, from
https://scholar.google.com/intl/en/scholar/about.html

[5] europeana.eu, About us, Retrieved 4 May 2020, from
https://www.europeana.eu/pl/about-us

[6] FBC, Retrieved 4 May 2020, from http://fbc.pionier.net.pl/pro/en/

[7] w3.org, Web Content Accessibility Guidelines (WCAG) Overview, Retrieved 4 May 2020, from https://www.w3.org/WAI/standards-guidelines/wcag/

[8] 5stardata.info, 5 star Open Data, Retrieved 4 May 2020, from
https://5stardata.info/en/