Bulgarian Open Science Digital Library - First Prototype

Atanas Georgiev, Krassen Stefanov

Faculty of Mathematics and Informatics, SU, Bulgaria
krassen@fmi.uni-sofia.bg

Abstract. In this paper we present the process of development of Bulgarian Open Science Digital Library (BOSDL) as the first and main ingredient of Bulgarian Open Science Cloud (BOSC). We introduce and discuss main principles involved in the Open Science movement. We also give a lot of technical details related to BOSDL development.

Keywords: Open Science, Open Access, Open Data, European Open Science Cloud, Bulgarian Open Science Cloud, Research Information System, Digital Library, Metadata.

1 Introduction

At the beginning we will present some basic concepts laying the ground for the development of the EOSC and BOSC as an integral part of EOSC.

What are Open science, open access, open education? All they advocate free access to all kind of results obtained from research projects and activities at Universities and all relevant scientific institutions. In fact, with the fast spreading of Internet, the traditional scientific publishing, used in centuries for dissemination of scientific results, is constantly replaced by new digital methods for dissemination through Internet. This new form of dissemination of scientific assets was named like Science 2.0, Open Research, and most recently as Open Science. The main goals for the Open science are to speed up the process of dissemination of research outcomes, to minimize the costs, to raise the quality and by shortening the path of every research outcome, to achieve vast speed up of the overall research productivity, and as a result significantly to shorten the path of innovations to society. This new trend is also beneficial for individual researchers, as they receive more visibility, with less costs and efforts.

There are different definitions of Open Science, but in this paper we will use the following one (Bartling & Friesike, 2014), (Access.nl), (Suber, 2012): "the dissemination of scientific knowledge that is as wide as possible, free of charge to all users, and accessible online."

Open Science (Bartling & Friesike, 2014) includes other concepts related to science such as Open Education, Open Software, Open Courseware and Open Research (having two important components: Open Data and Open Access).

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 9, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2019. ISSN: 1314-4006, eISSN: 2535-0366

Open Data (Open Data – An Introduction from the Open Knowledge Foundation), being major component of Open Science, is also freely available for everyone without any restrictions. One of the main features of Open Data is called FAIR (Findable, Accessible, Interoperable and Reusable) (FORCE11, 2016). While Open Data is FAIR data, not all FAIR data is Open Data. However, FAIR data could have clear and accessible data usage license.

Open Access (Access.nl), (Suber, 2012) means free, full and open online access to academic publications. There are two roads to publish scientific work with open access: golden road (in an open access journal) and green road (in a free public scientific repository).

In the next chapter we present the main ideas related to the development of the European Open Science Cloud. After that, we present the vision for the Bulgarian Open Science Cloud (BOSC) as an integral part of the EOSC. In the last chapter we present technical details related to the design and implementation of Bulgarian Open Science Digital Library as the first and main ingredient of the BOSC.

2 European Open Science Cloud (EOSC)

The famous Plan S (About Plan S, 2018) for accelerating the transition to full and immediate Open Access to scientific publications states the following key principle (goal): "After 1 January 2020 scientific publications on the results from research funded by public grants provided by national and European research councils and funding bodies, must be published in compliant Open Access Journals or on compliant Open Access Platforms."

What is a scientific research repository, named as compliant Open Access platform? This is in fact digital library, involving the following components:

- scientific and educational knowledge assets in digital form, like publications, books, movies, data collections, which are automatically indexed, searched and accessed using information technologies;
- all knowledge assets are classified into thematic collections and hierarchies, using different classification schemes and ontologies;
- all knowledge assets are described with relevant metadata format, compliant with major metadata standards used in digital libraries, and conforming to the OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting) standard. All such assets are stored in digital form, assessed using the network and following the Open access rules.

Digital libraries can collect both artefacts developed in digital form, as well as digitized copies of artefacts in other format (books, magazines, photographs, archives, etc.). The main functions of such digital libraries are:

- create and manage digital collections;
- open access to scientific and educational resources;
- sharing and reusing knowledge artefacts for education;
- increase visibility and expand influence of scientific knowledge assets;
- raise quality of scientific publications;

- enable fast access to most relevant and needed scientific results.

Metadata are the key to fulfill all these requirements.

At European level, the key initiative related to the progress and constant advance in the field of open science digital libraries is OpenAIRE (OpenAIRE). It provides standards, tools and services for Open Science implementation and to ensure their uptake on a global level, at least in Europe. OpenAIRE is one of the central initiatives behind the new ambitious plan of the EC to implement the so called European Open Science Cloud (EOSC Declaration, 2017), (Commission Staff Working Document - Implementation Roadmap for the European Open Science Cloud, SWD(2018) 83 final, 2018), (About Plan S, 2018), (November 2018, the European Commission launched the European Open Science Cloud (EOSC) in Vienna, 2018). The main goal is to move from Open Access to Open Science. In other words, to provide open access not just to publications, but to all types of scientific results, including Open/FAIR research data, open source software, open education including free educational resources, open services, open protocols, open methodologies. By linking together all possible research knowledge assets, to achieve the ultimate goal to open science to all European citizens.

3 Bulgarian Open Science Cloud (BOSC)

Following the efforts of all other European scientists, in Bulgaria we also started to think and work on opening the Bulgarian science to the world (Bulgarian Open Science Initiative). The ultimate goal is to develop the so called Bulgarian Open Science Cloud (BOSC), based on the same principles, standards and technologies and fully compliant with EOSC. We will present the main idea for the design and development of the BOSC. Then we will describe the first prototype of the Bulgarian Open Science Digital Library (BOSDL), as the main cornerstone of the new BOSC.

The current digital research libraries in Bulgaria are dispersed, not well integrated and using different data models and standards. Most of these libraries are not compliant with existing models and standards adopted by OpenAIRE and approved by EOSC. In fact, only three such digital libraries are listed as compliant.

The other important problem is related with adopted standards and practices from Bulgarian National Centre for Information and Documentation NACID (Bulgarian National Centre for Information and Documentation (NACID)), supporting the main general registries with all relevant scientific results available from all Bulgarian scientific organizations. They are definitely not compliant with existing models and standards adopted by OpenAIRE (OpenAIRE) and approved by EOSC.

So, we were forced to develop new model for storing all research knowledge assets in BOSDL, compliant with OpenAIRE and EOSC, and to design and implement various transition schemes, in order to transfer all available information from national registries supported by NACID, into new BOSDL.

The main work was focused on the development of open research digital repositories, preserving research outcomes and assets, working in multilingual mode and storing the full assets (either text, data or programs), and following well established metadata standards from OpenAIRE initiative (OpenAIRE) and supported by EOSC.

The main portal and the necessary repository were designed and the relevant models for research asset description and storage were developed and implemented. The model chosen is in full compliance with models offered by OpenAIRE and CRIS (BOSC portal). This data model relies on a set of basic entities as defined by the Common European Research Information Format (CERIF) model (Common European Research Information Format (CERIF)) maintained by the non-profit organization euroCRIS (BOSC portal) – see Picture 1.

On the base of this model, the first software prototype was developed, utilizing the open source system DSPACE-CRIS (DSPACE-CRIS).

The existing registries of NACID were analyzed and used to populate the prototype. The first version of automatic tool for extracting metadata from existing sources was developed. The information was further checked in the official public registries using DOI and ISBN, and also using the Crossref API (Crossref REST API).

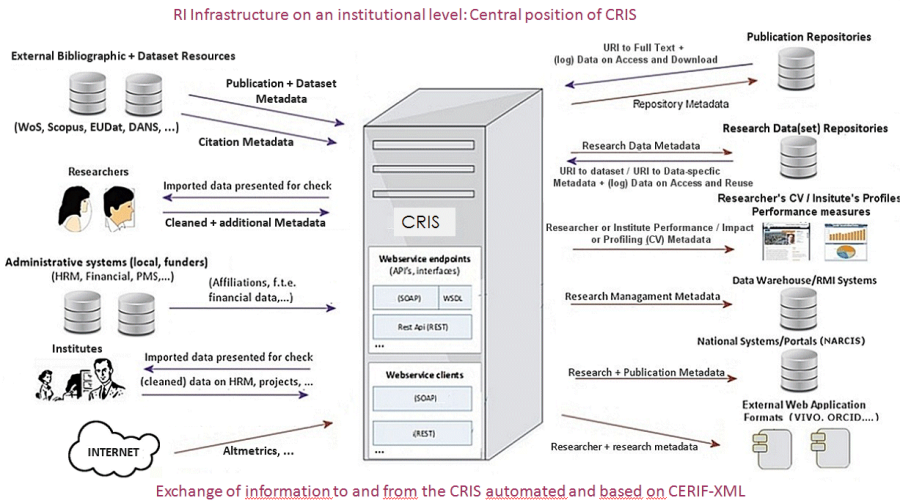

Fig. 1. CRIS and CERIF model in action

In addition, the available full text sources for all research assets were analyzed from their metadata descriptions, and relevant automatic tools for their extraction and storage in the BOSC were implemented.

A dedicated workflow for transferring entities from the Bulgarian Current Re-search Information System supported by NACID into the Bulgarian Open Science Digital Library (BOSDL) was developed. This workflow was designed as a multi-stage process, similar to the process of extraction, transformation and loading (ETL) data in data wharehouses. The multi-stage process designed consists of the following steps:

- **Metadata harvesting.** BOSRL objects are harvested using the REST API provided by NACID National Research Information System in JSON format and stored in a single MongoDB database organized in collections.

- **Extraction and finding of persistent digital identifiers DOI and ORCID.** First we are extracting DOI identifiers from publications using regular expressions (Crossref REST API). Then we find DOIs of publications referenced by a given reference (unstructured) using Search-Based Matching with Validation (SBMV) algorithm (Search-Based Matching with Validation (SBMV) algorithm). On the base of available information, we find authors' ORCID identifier by personal names and DOI identifier of an own (author) publication (ORCID Public API V 2.1).
- **Extracting and enriching bibliographic metadata and content.** Each DOI is associated with bibliographic metadata about the object, including one or more URIs where the object can be found. Using the DOI resolution service (http://dx.doi.org/) bibliographic metadata for each object is retrieved and stored. Using the bibliographic metadata all URIs are extracted and crawled for pdf content, if any, and downloaded.
- **Validation.** In order to validate the metadata record, we need to match the bibliographic metadata from DOI resolution service and source data from BOSRL. This step ensure that the extracted and found DOI identifiers are relevant for the corresponding documents. The match is performed by measuring the Levenshtein distance (Levenshtein, 1966) between the source and the target with pre-set threshold for measuring the similarity between two strings.
- **Transformation.** In contrast to the ingest of flat metadata formats like Dublin Core, the import of CRIS objects requires a denormalization of the entities in order to add to a given entity (e.g. publication) the properties of related entities (e.g. information from the person entity). A spreadsheets are generated that reflects the structure of the DSpace-CRIS data model. Finally the content files and metadata spreadsheets are transformed into a Simple Archive Format Package (Simple Archive Format Package) for batch import to the BOSDL repository.

4 Conclusions

The first prototype of Bulgarian Open Science Cloud was implemented and can be accessed by: https://cris.fmi.uni-sofia.bg/. It includes one main ingredient – Bulgarian Open Science Digital Library, developed using the DSPACE-CRIS open source digital repository software.

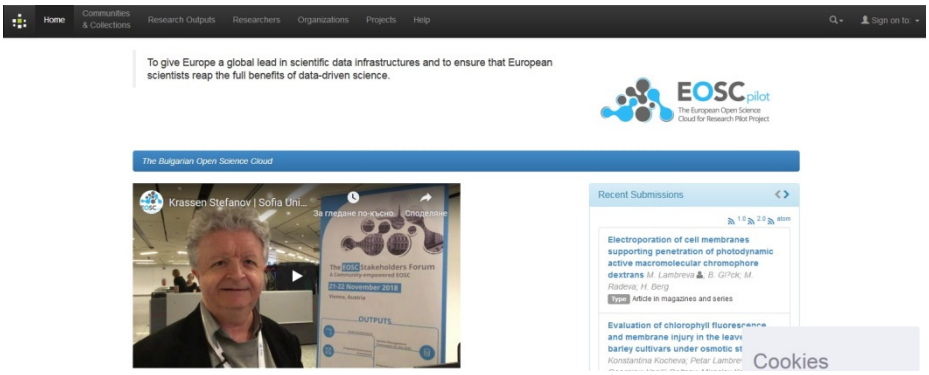

Fig. 2. BOSC portal in action

At the moment, we implement the main library, which will harvest and populate the information from all institutional libraries. Currently, only five such repositories were integrated, while all of the rest are under development and should be integrated in the next year. In such a way we have now the first proof-of-concept prototype of the Bulgarian Open Science Research Library, which will be further developed and improved in the next two years in order to fully satisfy all the requirements for the development of the Bulgarian Open Science Cloud (BOSC). The BOSC portal will be fully compatible with the registries supported by NACID and will involve all research outputs coming from projects funded by MES and BNSF.

In order to improve BOSDL data accuracy, consistency and integrity at the level of links between authors and their contributions e.g. publications, research projects etc., a unique persistent and international identifiers for researchers (for example: ORCID) should be adopted. The adoption of ORCID (ORCID Public API V 2.1) at a national level will give the opportunity to enhance the quality of the metadata and content files. As an example, the ability for researchers to login using their ORCID credentials and do a profile claiming, will improve the correctness and completeness of their publications, affiliations and bibliography details. Disambiguation of researchers' names will allow proper and robust interoperability, allowing data from the institutional libraries to be harvested at the moment of their appearance and/or change, which will enable the rapid refreshing and accuracy of research data collected in BOSDL.

Acknowledgements

This work was supported by the project "Information and Communication Technologies for a Single Digital Market in Science, Education and Security" of the Scientific Research Center, NIS-3317, funded from the Ministry of Education and Science.

References

About Plan S. (2018). Retrieved from About Plan S: https://www.coalition-s.org/

Access.nl, O. (n.d.). *What is open access?* Retrieved from http://www.openaccess.nl/en/what-is-open-access

Bartling, S., & Friesike, S. (2014). *Opening Science – The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing.* Heidelberg, Germany: Springer.

BOSC portal. (n.d.). Retrieved from https://cris.fmi.uni-sofia.bg/

Bulgarian Ministry of Education and Science (MES). (n.d.). Retrieved from http://mon.bg/en/100000

Bulgarian National Centre for Information and Documentation (NACID). (n.d.). Retrieved from http://nacid.bg/bg/

Bulgarian Open Science Initiative. (n.d.). Retrieved from https://npict.bg/node/49

Commission Staff Working Document - Implementation Roadmap for the European Open Science Cloud, SWD(2018) 83 final. (2018). Retrieved from http://ec.europa.eu/transparency/regdoc/rep/10102/2018/EN/SWD-2018-83-F1-EN-MAIN-PART-1.PDF

Common European Research Information Format (CERIF). (n.d.). Retrieved from https://www.eurocris.org/cerif/main-features-cerif

CRIS - data model used for the definition of research information systems. (n.d.). Retrieved from https://www.eurocris.org/

Crossref REST API. (n.d.). Retrieved from https://www.crossref.org/services/metadata-delivery/rest-api/

DSPACE-CRIS. (n.d.). Retrieved from https://dspace-cris.4science.it/ ; https://wiki.duraspace.org/display/DSPACECRIS/DSpace-CRIS+Home

EOSC Declaration. (2017). Retrieved from EOSC Declaration: https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none

FORCE11. (2016). *The FAIR Data Principles*. Retrieved from https://www.force11.org/group/fairgroup/fairprinciples

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady. 10 (8)*, 707-710.

November 2018, the European Commission launched the European Open Science Cloud (EOSC) in Vienna. (2018). Retrieved from https://eosc-launch.eu/home/

Open Data – An Introduction from the Open Knowledge Foundation. (n.d.). Retrieved from https://okfn.org/opendata/

OpenAIRE. (n.d.). Retrieved from https://www.openaire.eu/

ORCID Public API V 2.1. (n.d.). Retrieved from https://pub.orcid.org/v2.0/#!/Public_API_v2.1/searchByQueryV21

Search-Based Matching with Validation (SBMV) algorithm. (n.d.). Retrieved from https://github.com/CrossRef/search-based-reference-matcher

Simple Archive Format Package. (n.d.). Retrieved from https://github.com/DSpace-Labs/SAFBuilder

Suber, P. (2012). *Open access [PDF version]*. Retrieved from Open access [PDF version]: https://mitpress.mit.edu/sites/default/files/9780262517638_Open_Access_PDF_Version.pdf

Received: June 04, 2019
Reviewed: July 05, 2019
Finally Accepted: July 25, 2019