Using Dspace Platform for Creation of Open Access Local Repositories

George Simeonov[1], Peter Stanchev[1, 2]

[1] Institute of Mathematics and Informatics, BAS, Bulgaria
[2] Kettering University, Flint, USA
gsimeonov@math.bas.bg, pstanche@kettering.edu

Abstract. The paper presents a brief review of the Dspace software platform for long-term data storage with indexing and search system used for open access repository creation. The experience of using and maintaining the platform for building BulDML and BGOpenAIRE repositories are highlighted.

Keywords: Dspace, Open Access Repositories, BulDML, BGOpenAIRE

1 Introduction

A digital repository is a collection of online resources/collections. Each community can have a number of sub-communities and a number of collections. Each collection may contain a number of publications, reports, data, and any other digital material.

A common repository information model is given in Figure 1.

The digital repository base functionalities include ingestion and storage of data, support of various file formats, data integrity checks, persistent identification, navigation and advanced search, user rights management, optional subscriptions.

A given repository could be connected to other repositories for data exchange, according to the FAIR data principles. FAIR data are data which meet standards of findability, accessibility, interoperability, and reusability (The FAIR Guiding Principles, 2016).

2 Dspace Platform

An overview of the content structure in a Dspace platform is shown in Figure 2 (Dspace platform, 2018).

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

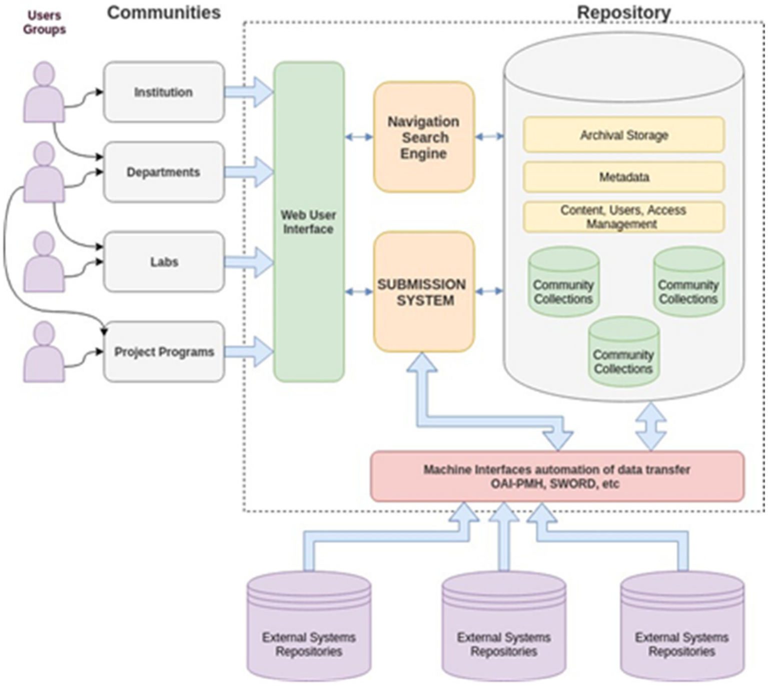

Fig. 1. Repository information model

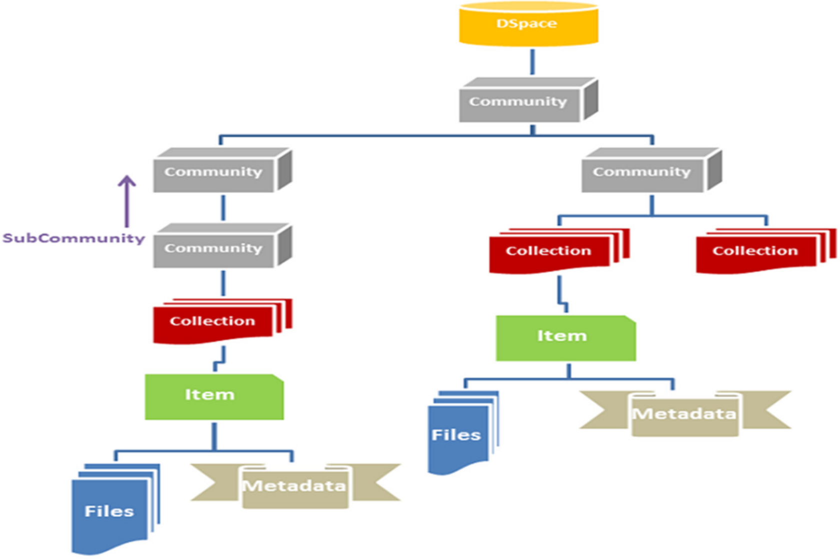

Fig. 2. Overview of a content structure in a Dspace platform

The digital repository based on a Dspace open source software platform provides good opportunities for long-term storage of research results, articles, work papers, reprints, technical reports, conference reports, and data-sets in various digital formats. The platform is made of the following main components:

- RDBMS Database;
- Search engine;
- Users and groups management;
- Different levels of access to content management;
- API interface for system interoperability with external systems.

The Dspace data submission tools include:

- An administrative web interface for management web forms as well as machine to machine API interface for data exchange.
- A web interface for publishing in a collection.

The data entry is passed sequentially through the following seven steps:

1. Describe – Select options to specify what additional input fields will be available in the next steps;
2. Describe – Fill in the main metadata fields;
3. Describe – Set keywords classifying the topic of the material as well as fill in additional metadata fields;
4. Upload – load one or more files. Calculate checksums. Auto-detect file formats;
5. Check – check integrity of attachments and file formats; Edit in case of errors;
6. Check – check compliance with already existing data; Edit in case of errors;
7. Apply – apply a license agreement.

Repositories have support for different user types/roles and user groups with different levels of access and predefined rights in the system. For each published document, a unique HandleNET identifier is used.

When a document or data becomes part of the repository, a permanent URL is assigned. This means that unlike most URLs, this identifier will not need to be changed when the system migrates to new hardware or when system changes are made. The repository provides maintaining the integrity of this identifier, so the users can use it as a permanent link when they quote it in publications or other systems.

Meta data used to describe stored items in digital repositories contain:

- A standard Qualified Dublin Core metadata schema describing the stored data (Dublin core, (n.d.);
- A DSpace Dublin Core Metadata Registry. This registry initializes the default scheme where dc is used to identify the name-space. Because this registry is intended to track the Dublin Core standard, it is recommended that the DSpace administrator does not add/remove metadata fields from this name-space; instead, the "Local Metadata Registry" is used.

Data file formats and data integrity include:

- A mechanism to periodically compute and monitor digital “fingerprints” of file data and to periodically check the integrity of the data with any defects/changes in the data triggering restoration of the original version;
- A system log, storing information for each change of the data, with time and contributing users recorded;

- Unicode / UTF-8 encoded text data;
- The Repository can store files of many types such as DOC, PDF, XLS, PPT, JPEG, MPEG, TIFF, etc.

3 Using Dspace Platform for Building and Maintaining BulDML and BGOpenAIRE Repositories

IMI-BAS has so far provided open access to a total of over 2800 scientific publications, reports from national, international conferences, journals. Documents and data in the contents of the repository include articles, books, collections, and dissertations. Archival collections of OCR digitized documents from the 1953 Bulletin de l'Institut de Mathématiques are included. The default file format for scanned documents is PDF. Part of the scanned documents is processed with OCR. The original type is stored as a Recognized OCR text in a hidden layer of the PDF, which is enough for a full text search. BulDML repository of IMI-BAS is indexed by the European project for mathematical libraries EuDML (Rákosník, Stanchev, Pavlov, 2016). There is an established link for indexing and access to metadata. The IMI-BAS repository has been tested and can be stored and searched for mathematical formulas.

BGOpenAIRE include publications and reports on projects funded by EU programs.

The base 15 elements of the abbreviated version of DC used are:

1. TITLE - The name given to the resource by the CREATOR or PUBLISHER;
2. CREATOR - The person(s) or organization(s) primarily responsible for the intellectual content of the resource; the author(s);
3. SUBJECT - The topic of the resource; also keywords, phrases or classification descriptors that describe the subject or content of the resource;
4. DESCRIPTION - A textual description of the content of the resource, including abstracts in the case of document-like objects; or a content description in the case of visual resources;
5. PUBLISHER - The entity responsible for making the resource available in its present form, such as a publisher, university department or corporate entity.
6. CONTRIBUTORS - Person(s) or organization(s) in addition to those specified in the CREATOR element, who have made significant intellectual contributions to the resource but on a secondary basis;
7. DATE - The date the resource was made available in its present form;
8. TYPE - The resource type, such as home page, novel, poem, working paper, technical report, essay or dictionary. It is expected that TYPE will be chosen from an enumerated list of types;
9. FORMAT - The data representation of the resource, such as text/html, ASCII, Postscript file, executable application or JPG image. FORMAT will be assigned from enumerated lists such as registered Internet Media Types (MIME types). MIME types are defined according to the RFC2046 standard;
10. IDENTIFIER - A string or number used to uniquely identify the resource. Examples from networked resources include URLs and URNs (when implemented);

11. SOURCE - The work, either printed or electronic, from which the resource is delivered (if applicable);
12. LANGUAGE - The language(s) of the intellectual content of the resource;
13. RELATION - The relationship to other resources. Formal specification of RELATION is currently under development;
14. COVERAGE - The spatial locations and temporal duration characteristics of the resource. Formal specification of COVERAGE is currently under development;
15. RIGHTS MANAGEMENT - A link (URL or other suitable URI as appropriate) to a copyright notice, a rights-management statement or perhaps a server that would provide such information in a dynamic way.

API Open Interface (OAI-PMH / OAI-ORE) is used. The OAI-PMH standard uses the Hypertext Transport Protocol (HTTP) protocol as a transport layer and defines six query methods (called verbs) that must be supported by an OAI-PMH-compliant data provider (also called a repository).

The used methods are:

- **GetRecord** - Retrieves zero or one full metadata record from a repository;
- **Identify** - Retrieves general information about the repository;
- **ListIdentifiers** - Retrieves zero or more metadata titles (not the complete metadata entry) from a repository;
- **ListMetadataFormats** - Retrieves a list of available formats for metadata saved by the repository;
- **ListRecords** - Retrieves zero or more complete metadata entries from the repository;
- **ListSets** - Retrieves the specified content structure in the repository.

Program Modifications Extensions in the DSpace system for BulDML made by IMI-BAS include:

- Separate lexicographic indexes, created for Cyrillic script, allowing title, author and keywords sorting and output of the results separated from the ones in Latin script.
- Added support for LaTeX math formulas in annotations/abstracts, allowing the user to store the logical structure and visualization of the formulas through the web interface and the ability to search them in textual form.
- Part of BulDML content is being retrieved and processed automatically via system interface to the portal of European Mathematical Library project (EuDML).

Our permanent URLs are registered in the Handle System - a complete system for assigning, managing and authoring permanent identifiers known as handles for digital objects and other resources on the Internet.

4 Future Development of Repositories Software Platforms

A new User Interface based on Angular is coming to replace XMLUI and JSPUI, aiming to unify features from both interfaces into a single UI.

Enhanced REST API (using modern REST best practices) with new REST Contract (describing all API interactions) is developed.

A new configurable object model (tentatively called "Entities"), which allows for the creation of new "typed" Items, and storing relationships between Items is under development;

Support of external identifier systems like ORCID and current research information systems (CRIS), journal publishing systems, etc. is completed;

Alignment with core recommendations from the COAR Next Generation Repositories Report (COAR, n.d.) is necessary.

Next upgrade of Dspace will introduce the following new features:

- New Angular UI - responsive and easy for customization user interface;
- Configurable entities - an optional new item type that allows for complex linked relationships between items;
- Journal Hierarchy - create and link objects for a Journal, Volume, Issue, Article, and Author;
- Research Entities - represent objects that interact in the research life cycle, including Publications, Projects, Faculty, Departments;
- A completely redesigned REST API that is self-documenting and human browsable;
- Redesigned submissions and workflows featuring a one-page submission process with a drag-and-drop interface, and automatic metadata extraction from common formats like PDFs.

Acknowledgments.

This work was partially supported by the Bulgarian Ministry of Education and Science under National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security”, approved by DCM No 577, 17 August 2018.

References

COAR Next Generation Repositories Report, (n.d.). Available at: http://ngr.coar-repositories.org/

Dspace platform (2018). Available at: https://wiki.duraspace.org/display/DSDOC/All+Documentation

Dublin core, (n.d.). Available at: http://dublincore.org

Rákosník, J., Stanchev, P., Pavlov, R. (2016) European Digital Mathematics Library EuDML. Current State and Future Plans. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, Vol. 6, 2016, 29-36.

The FAIR Guiding Principles for scientific data management and stewardship (2016). Scientific Data. p. 160018. Available at: doi:10.1038/sdata.2016.18

Received: June 15 2019
Reviewed: July 05, 2019
Finally Accepted: July 20, 2019