Преглед НЦД 23 (2013), 1–10

Maria M. Nisheva-Pavlova, Pavel I. Pavlov
Faculty of Mathematics and Informatics, Sofia University

OPEN SOURCE SOFTWARE TOOLS FOR CREATING DIGITAL REPOSITORIES

Abstract. The paper presents the main results of a survey aimed at a comparative analysis of some widely used open source software tools for creating digital repositories with respect to their applicability for web publishing of digitized Bulgarian cultural heritage collections and making them accessible via Europeana. The parallel has been made according to a well-defined set of features like availability of convenient authoring/annotation and browsing tools, image scaling and high resolution support, functionalities of the built-in search engine, recognizable file types, user management, OAI-PMH compliance, etc. The Europeana Data Model and the Europeana Semantic Elements standard are briefly described from the viewpoint of publishing in Europeana. Some practical recommendations and conclusions are given as a final result.

Keywors. Electronic Publishing, Digital Repository, Metadata, Search Engine, Europeana.

1. Introduction: Europeana

Europeana (http://www.europeana.eu/portal/) was launched in 2008, with the goal of making Europe's cultural and scientific heritage accessible to the public. It is based in the National Library of the Netherlands, the Koninklijke Bibliotheek. Europeana builds on the experience of The European Library (TEL, http://www.theeuropeanlibrary.org/), which is a service of the Conference of European National Librarians.

In the wide public Europeana is primarily perceived as a web portal exposing increasingly impressive amounts of cultural heritage from various sources to Europe's citizens. On an abstract level Europeana can be seen as a large collection of surrogate objects representing digitally born or digitized cultural heritage objects which themselves remain outside the Europeana data space.

As indicated in Figure 1 [1], Europeana has an API for end user functionality as well as an I/O-API enabling data flow from and to the content providers. The latter creates the option of re-integrating enriched content in the remote applications of the data providers.

Europeana only harvests the metadata describing a digital object and provides a thumbnail with a link to the digital object's location on the provider's or aggregator's system. A user is always directed to the provider's or aggregator's site for viewing the digital object in detail. That enables the content holder to apply their own terms to use to their content. Conditions and formats under which the provider makes the content available vary across providers (paid or free, upon registration, upon order, for viewing via a standard browser or commercial and downloadable plug-in, etc.).

National institutions and portals representing several sectors are the preferred first contact point for Europeana. A new content provider is, therefore, first directed to the national aggregator, if this one is established. However, in some countries these national institutions (portals) are not yet established, and another routing will then be suggested. If an aggregator is proposed, the organization should contact the aggregator directly for information about how

to submit data. The Europeana Office may decide to take the content directly rather than via an aggregator.

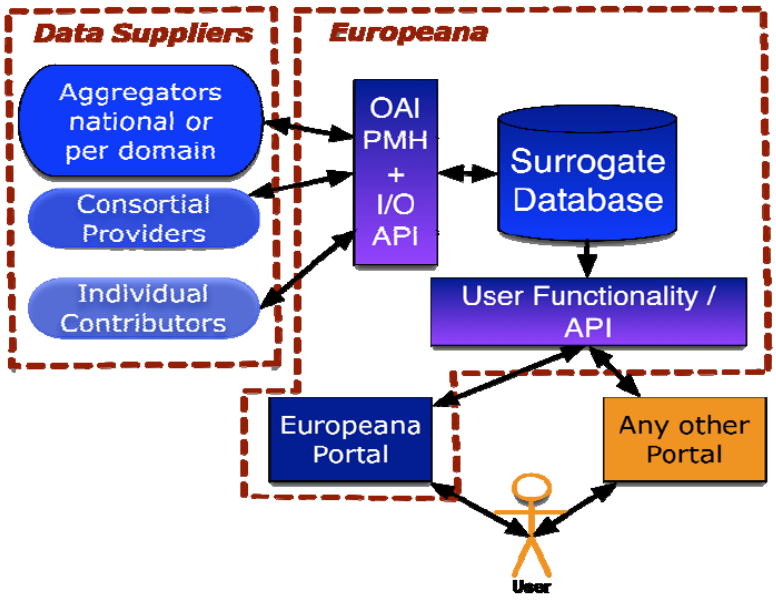

Figure 1 (originally published in [1]). Europeana APIs

Every content provider and aggregator needs to comply with Europeana's technical requirements when submitting data. Europeana provides a common access point to digital cultural heritage objects across different cultural domains. It complements but does not duplicate the source environment of the digital objects. To achieve this, Europeana uses specialized portals. The object is linked to Europeana and is shown in a neutral environment, while at the same time still being available in the domain-specific portal, which may provide greater contextualization.

Europeana aims to provide access to digital objects at the lowest possible level of granularity. This implies giving direct access to the digitised object itself, that is, with a minimum click distance between the description and the object. The minimum granularity can vary, and it is in the discretion of the content provider to decide this. A broadcast provider, for example, might decide to cut down a news programme made of individual sections, and make each one available as a separate digital object. On other occasions, the complete program is of value due to the context it adds to the individual fragments. Europeana asks that content providers keep the users in mind when deciding on the granularity of their data.

Europeana's data model enables search and discovery of digital objects. Europeana maintains a common central index of the objects' metadata. It has, therefore, an object-centric rather than a collection-centric approach.

Europeana stores representations of digital objects and not digital objects themselves. Europeana generates a description and a preview of each digital object with the help of the metadata and thumbnails or previews of the digital objects. This requires that there is a repository on the content provider's/aggregator's side, where the digital objects are stored and can be linked to. It also requires a native website that can be used to view, play and reuse the objects.

A digital object in Europeana is a unique single entity, which can be viewed/played by users (e.g., mpeg movie, mp3 audio, jpeg photo, PDF text, etc.) on their computers. A digital object is the digitized version of a physical cultural artefact. Europeana does not accept descriptions that do not correspond to a digital object.

Europeana harvests, stores and indexes the metadata in a central index. The preferred method for harvesting is the OAI-PMH protocol (http://www.openarchives.org/pmh/) and partners need to set up an OAI-PMH repository comprising their data mapped to the Europeana Semantic Elements (ESE) standard.

A content provider or an aggregator is responsible for making available to Europeana the following data [2]:

- Metadata (descriptive, administrative) describing a digital object. The metadata must be mapped to the ESE v3.2.2. This is the Europeana current data model which consists of the Dublin Core (DC) metadata elements, a subset of the DC terms and a set of twelve elements which were created to meet Europeana's functionality needs [3],
- A preview or thumbnail of the described object,
- Persistent identifiers – active and stable links to the described digital object on the provider's site or the portal's site.

Persistent identifiers are mandatory when submitting data to Europeana, because of the role they play in preventing duplication. Europeana has a wide network of aggregators and content providers, and the possibility of data being ingested more than once must be avoided. The development and attribution of persistent identifiers helps to deduplicate content and provide greater control over the data.

Harvesting in Europeana should follow the implementation guidelines OAI-PMH v2 for repository implementers [4]. Repository implementers should consider exporting DC the first and most important step toward OAI-PMH interoperability. Facilities to export other formats may be added later.

A new approach towards structuring and representing data delivered to Europeana by the various contributing cultural heritage institutions is the Europeana Data Model (EDM). This model aims at greater expressivity and flexibility in comparison to the current ESE, which it is destined to replace [5]. The design principles underlying the EDM are based on the core principles and best practices of the Semantic Web and Linked Data efforts to which Europeana wants to contribute. It acts as a common top-level ontology which retains original data models and information perspectives and enables interoperability at the same time.

Typical object representations in Europeana mostly will be compound entities consisting of several parts, such as for instance metadata attributes, a thumbnail picture and a static html landing page. For this and other reasons the OAI Object Reuse & Exchange specifications (http://www.openarchives.org/ore/1.0/to) were chosen as the structural modelling framework for the EDM ontology.

This survey is aimed at a comparative analysis of some widely used open source software tools for creating digital repositories with respect to their applicability for web publishing of digitized Bulgarian cultural heritage collections and making them accessible via Europeana.

The comparison is based on the following main features: availability of convenient authoring tools (tools for entering data and metadata), support of high resolution images, image scaling, availability of image manipulation libraries, browsing tools, built-in search engine, metadata format(s), maintained file types, suitable database, public access to content, user management, semantic support, OAI-PMH compliance.

2. DSpace

DSpace (http://www.dspace.org) is an open source repository software tool for creating repositories focused on delivering digital content to end users, and providing a full set of tools for managing and preserving content within the application. DSpace is the most widely used repository software platform, with over 1000 installations worldwide representing a growing and active user community.

DSpace was developed by the MIT Libraries and the Hewlett-Packard Company. It may be characterized as a set of cooperating Java applications and utility programs that maintain an asset store and an associated metadata store. The web applications provide interfaces for administration, deposit, ingest, search, and access. The asset store is maintained on a file system, or similar storage system. The metadata, including access and configuration information, is stored in a relational database.

DSpace is written in Java. It requires the use of either a PostgreSQL or Oracle database. PostgreSQL is open source and free to use. It also requires a Java Servlet Container, which can be Apache Tomcat, Jetty or Caucho Resin.

Technical information (summary):

Authoring/Annotation Tool(s)	Yes (familiar to librarians and archivists)
Supports High Resolution Images	Yes
Image Scaling	No support
Browsing Tool(s)	Yes
Built-in Search Engine	Yes (full-text search for end users; faceted search; can use logical conjunctives across different fields)
Metadata	Qualified Dublin Core (default metadata schema); can also translate metadata from MARC/MODS and other metadata schemas
File Types	Recognizes files of any common format (e.g. TXT, DOC, PDF, JPEG, MPEG, TIFF). Any file type can be uploaded to DSpace
Suitable Database	Uses PostgreSQL
OAI-PMH Compliance	Yes
Public Access to Content	DSpace has a user group called “Anonymous” which can have read access given to it. Thus, public and private access is managed using the same interfaces and rules as repository users
User Management	DSpace manages users and groups, which can have permissions over communities, collections and items
Image Manipulation Libraries	Yes
Semantic Support	Through the DSpace History System, metadata can be exported on the command line into N3 and RDF/XML

DSpace is easy-to-install and running quickly. It does not have much built in support for operations over specific file types such as images and videos. For example, preview thumbnails are not generated for images or videos.

Figure 2. Browsing in the Research Portal of SU

DSpace is popular in Bulgarian academic institutions – e.g., the research portal of Sofia University (see Figure 2) and the digital repository of IMI–BAS run on it.

3. Invenio

Invenio (http://invenio-software.org/) is a free software suite enabling one to run his/her own digital library or document repository on the web. The technology offered by the software covers all aspects of digital library management from document ingestion through classification, indexing, and curation to dissemination. Invenio complies with standards such as OAI-PMH and uses MARC 21 as its underlying bibliographic format. The flexibility and performance of Invenio make it a comprehensive solution for management of document repositories of moderate to large sizes (several millions of records).

Invenio has been originally developed at CERN to run the CERN document server, managing over 1 000 000 bibliographic records in high-energy physics since 2002, covering articles, books, journals, photos, videos, and more. Invenio is being co-developed by an international collaboration and is being used by about thirty scientific institutions worldwide.

Technical information (summary):

Authoring/Annotation Tool(s)	Yes
Supports High Resolution Images	Yes
Image Scaling	No information found
Browsing Tool(s)	Yes
Built-in Search Engine	Yes (customizable simple and advanced

	search interfaces; combined metadata, full text and citation search in one go)
Metadata	Uses MARC 21 as its underlying bibliographic format; handles articles, books, theses, photos, videos, museum objects and more
File Types	Recognizes files of any common format (e.g. TXT, DOC, PDF, JPEG, MPEG, TIFF). Any file type can be uploaded
Suitable Database	Uses MySQL
OAI-PMH Compliance	Yes
Public Access to Content	Ensures open access to the published collections
User Management	Invenio manages users and user groups (provides construction of user-defined document baskets and basket-sharing within user groups)
Image Manipulation Libraries	No information found
Semantic Support	Some (based on the employment of new HTML5 tags and conventions to Invenio output)

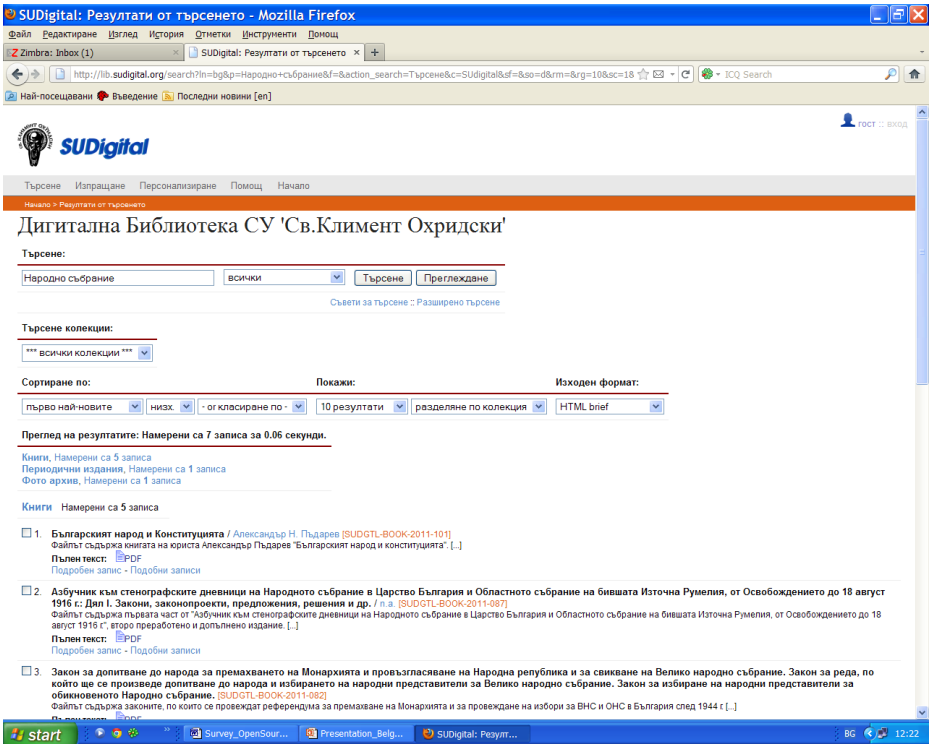

Figure 3. Search results in SUDigital

Invenio is popular in Bulgarian academic institutions – e.g., the digital library in the field of humanitarian studies at Sofia University SUDigital (see Figure 3) runs on it.

4. Nuxeo DAM/DM

Nuxeo DAM (http://community.nuxeo.com/) is a Digital Asset Management system, and Nuxeo DM is a Document Management system. Nuxeo DAM has a rich set of features for working with digital media assets, such as images and videos, while Nuxeo DM is a more mature product for dealing with the management of any files, without specific features for dealing with file types (like images and videos).

It is possible to integrate the two systems together, so that the Nuxeo DAM interface runs on media files that are stored and managed from Nuxeo DM.

The BBC has adopted Nuxeo DAM for their mobile portal, which receives 100 million hits per month. Nuxeo DAM is written in Java and uses JDBC to interface with databases. This allows users to use a PostgreSQL, MySQL, Oracle, Microsoft, or H2 database.

Technical information (summary):

Authoring/Annotation Tool(s)	Yes (annotations can be added to a whole image or selected areas of the image)
Supports High Resolution Images	Yes
Image Scaling	Can zoom in and out of images. Shows images individually in an inspector as well as in a grid
Browsing Tool(s)	Yes
Built-in Search Engine	Yes (searching the DAM is performed through a faceted filtering interface, where images can be filtered by: folder, category, file type, topic, geographic coverage and authoring date. This is in addition to keyword filtering)
Metadata	No reliable information about the used metadata format(s)
File Types	Nuxeo DAM supports any file type, and can preview images, videos and audio specifically. For images, annotated areas can be selected and commented on. For videos, a storyboard of thumbnails is automatically generated. Videos and audio files can be played inline
Suitable Database	PostgreSQL, MySQL, Oracle, Microsoft or H2
OAI-PMH Compliance	No reliable information found
Public Access to Content	Items can be exported, which provides their public URL. Through Nuxeo DM, workspaces can be explicitly exported as web pages
User Management	Manages users and groups, with access

	control per user and group
Image Manipulation Libraries	Yes
Semantic Support	Nuxeo has a built-in Relational engine that allows predicates to be assigned to items. Vocabularies are managed centrally, where URIs for predicates can be added, with a human readable label. Nuxeo uses Jena internally to model its RDF data. Through the Nuxeo import and export scripts, the RDF data can be exported

Nuxeo DAM offers file downloads for different sizes (original, medium and thumbnails). Through Nuxeo DAM users can comment on items and annotate them. Through Nuxeo DM users can create collaborative workspaces and discussion forums.

5. Alfresco

Alfresco (http://www.alfresco.com/) is an open source enterprise content management system. It is not a dedicated digital asset management (DAM) solution, however, there is support through plugins and configuration options that allow it to support the typical features of a DAM.

Alfresco is written in Java and can connect to either a MySQL or PostgreSQL database. It also uses Spring for its architecture, Hibernate for writing to the database, Lucene for searching, and iPhone apps so that it can be accessed from iPhones.

Technical information (summary):

Authoring/Annotation Tool(s)	Yes (comments can be added to each image)
Supports High Resolution Images	Yes
Image Scaling	It is possible to integrate with ImageMagick for transformations inline, but this is not available in their demo
Browsing Tool(s)	Yes
Built-in Search Engine	Yes (uses a Lucene based search with customisable support for field based searches)
Metadata	No reliable information about the used metadata format(s)
File Types	Manages any file type – documents, images, video and audio
Suitable Database	PostgreSQL or MySQL
OAI-PMH Compliance	No reliable information found
Public Access to Content	Each asset lists its public URL. Also, each asset can be checked out to Google Docs
User Management	Users and groups, with access control per user and group
Image Manipulation Libraries	Yes

Semantic Support	Alfresco manages metadata using "aspects". An aspect is simply a collection of metadata fields, and an item in the document library can have a number of aspects assigned to it, which increases the number of metadata fields the item has. Semantic export is not currently supported

Alfresco is divided into Sites. Each site has users that are members, and groups can also be added to sites. Each site has its own wiki, blog and discussion board, which can be used to make rich notes, make dated updates, and to have discussions about items, respectively.

Alfresco looks just like SharePoint to Microsoft Office, allowing users to upload, check-in, check-out and modify content right from Office. There is a lot of documentation on how to develop plugins for Alfresco. It lacks RDF export, but has a sophisticated metadata system.

6. REPOX

REPOX (http://repox.ist.utl.pt/) is a framework for managing metadata spaces. Some authors characterize it as a Data Aggregation and Interoperability Manager. It comprises several channels to import metadata from data providers, services to transform metadata between schemas according to user's specified rules, and services to expose the results to the exterior. REPOX aims to provide to all the TEL and Europeana partners a simple solution to import, convert and expose their bibliographic data via OAI-PMH, by the following means:

- *Cross platform*
  It is developed in Java, so it can be deployed in any operating system that has an available Java virtual machine.
- *Easy deployment*
  It is available with an easy installer, which includes all the required software.
- *Support for several metadata formats and encodings*
  It supports UNIMARC and MARC21 schemas, and encodings in ISO 2709, MarcXchange or MARCXML. During the course of the TELplus project, support will be added for other possible encodings required by the partners.
- *Metadata crosswalks*
  It offers crosswalks for converting UNIMARC and MARC21 records to simple Dublin Core as also to TEL Application Profile. A simple user interface makes it possible to customize these crosswalks and create new ones for other formats.

REPOX is a standalone OAI-PMH server developed within the TELplus project. It aims to provide all TEL partners with a simplified and flexible solution for exposing their data via OAI-PMH. The technology for creating a data provider and data source as well as the one for mapping metadata to TEL format are described in [6] and [7].

7. Conclusions

According to [8], Alfresco provides the richest collaboration environment of any of the evaluated systems, with easy to use wiki, blog and discussion areas for each sub-project on the site. Nuxeo DAM/DM provides the best tools for access to published rich digital

content (and to its particular components) and allows the richest support for semantic metadata to be associated with digital media assets.

From the point of view of the resources needed for web publishing of existing digitized Bulgarian cultural heritage collections and making them accessible via Europeana, Invenio (possibly in combination with REPOX) seems to be the most adequate solution. However, having in mind the experience gathered in leading national academic institutions and the lack of news during the last year on Invenio website, we recommend DSpace as the preferable repository tool under present conditions.

As an additional advantage of this choice, the possibility of integrating DSpace with DuraCloud [9] should be considered. DuraCloud is an open source software platform which aims to maintain a fully integrated environment where services and data can be managed across multiple cloud providers and in this way a better support for data preservation, data transformation, and data access can be provided.

Acknowledgements. This work has been supported by the Bulgaria-Korea IT Cooperation Center and by the Bulgarian National Science Fund within Project ДДВУ 02/22/2010.

References

[1] Concordia, C. et al. *Not (just) a Repository, nor (just) a Digital Library, nor (just) a Portal: A Portrait of Europeana as an API*. http://www.ifla.org/files/hq/papers/ifla75/193-concordia-en.pdf, last accessed on November 27, 2011.

[2] The Europeana Team. *Europeana Aggregators' Handbook*. http://version1.europeana.eu/c/document_library/get_file?uuid=94bcddbf-3625-4e6d-8135-c7375d6bbc62&groupId=10602, last accessed on November 27, 2011.

[3] Friberg, A. *Europeana*. National Conference on Library Digitalisation (Budapest, March 17, 2010). http://dtt.ogyk.hu/letoltesek/Friberg_Annette.pdf, last accessed on November 27, 2011.

[4] Lagoze, C., Sompel, H., Nelson, M., Warner, S. (Eds.). *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: Guidelines for Repository Implementers*. http://www.openarchives.org/OAI/2.0/guidelines-repository.htm, last accessed on November 27, 2011.

[5] Doerr, M. et al. *The Europeana Data Model (EDM)*. http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf, last accessed on November 27, 2011.

[6] *REPOX Documentation*. http://repox.ist.utl.pt/doc/index.html, last accessed on November 27, 2011.

[7] *D5.3.1 – Europeana OAI-PMH Infrastructure – Documentation and final prototype, Appendix – REPOX User Manual*. http://www.europeanaconnect.eu/documents/01a_Europeana_OAI_PMH_APPENDIX_User%20Manual.pdf, last accessed on November 27, 2011.

[8] Packer, H. *Comparison of Digital Asset Management Systems (DAMs) and Content Management Systems (CMSs)*. ResearchSpace Project, January 2011.

[9] Kimpton, M., Payette, S. *Using Cloud Infrastructure as Part of a Digital Preservation Strategy with DuraCloud*. http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/UsingCloudInfrastructureasPart/206548, last accessed on November 27, 2011.

marian@fmi.uni-sofia.bg
pavlovp@fmi.uni-sofia.bg