Преглед НЦД 36 (2020), 38–47

Andrija Sagić
Library "Milutin Bojić", Belgrade

ISLANDORA
MODULAR DIGITAL REPOSITORY OF LIBRARY "MILUTIN BOJIĆ"

Abstract. Islandora, as a collection of Drupal modules, Fedora repository software and Solr software for indexing and search, gives a wide palette of possibility to create a digital repository. Modular digital repository of library "Milutin Bojić" is made on Islandora. This paper explains a process of building the repository and shows its background infrastructure, also it presents an introduction to Islandora with some detailed technical and methodological descriptions for further exploration.

1. Introduction

The process of making a functional digital repository has some planning and realization tasks that must be followed.This paper describe a process of forming digital repository of "Milutin Bojić" Library. The first task was finding an appropriate solution to get needed results. Islandora is a collection of modules built to manage and configure a digital repository according to the needs of an institution. Islandora shows as a solution which best suits to demands additionally to the demands of construction of repository "Milutin Bojić" Library. The background architecture of Islandora is presented in parts 3-5, while part 6 describes the main elements used in construction of "Milutin Bojić" Library digital repository.

2. Repository properties and findings

Digitized collection of poet Milutin Bojić consists of various types of material, it includes published books, manuscripts, studies, articles and photos. This variety of types became one of the demands for repository, to be able to deliver material in a different viewer. Also, it is important, for repository, to support a metadata standards and protocols (MODS[1], MARC[2] and DC[3], OAI-PMH[4]), tools for administration, adaptable design and must be published under the FOSS[5] license. After getting a list of requirements, the next task was search for software that support previous list of requirements which can give such support for a collection.

An excellent source to start searching a software for a digitization and librarian tools is website **foss4lib.org** where tools are categorized according to type, specific tags (platforms, software language, etc.). Site also provides good advice for software according to user needs. There is a good "Guide to Institutional Repository Software" [6] published by UNESCO.

After testing several repositories on virtual machines and comparing different active online repository solutions, the results was that Islandora repository could handle required demands. Modular repository, as Islandora, is upgradable to new functions and demands and gives a space to administrators and developers to further improve a repository. Islandora gives a possibility to adapt or make a Solution Pack for your needs and there is a good documentation how to do that [7]. There is also an Islandora Lab

repo a list [8] of additional modules and packs that are not officially included in Islandora but can be modified and included in own digital repository.

3. Structure of Islandora

Islandora is an open source software framework for managing and publishing a digitized material, it is built on Fedora[9], Drupal[10] and Solr[11]. **Fedora** is a robust, modular, open source repository system for the management and dissemination of digital content, **Drupal** is one of the best content manager systems (CMS) for a website creating, **Solr** is a powerful indexing tool for searching. Islandora includes **Islandora Solution Packs** which allow users to manage their repository according to needs. Figure 1 shows a diagram of Islandora architecture with three layers Fedora – Repository Layer, Islandora – Integration Layer and Drupal – User Interface Layer.

Fedora – Repository Layer stands in the bottom and serves to store objects and metadata, it preserves integrity of collection and includes:

- connected to MySql database
- connection to Solr through GSearch
- XACML policies encode access control (part of Fedora secure policies)[13]
- Content Models, which are an integrated structure for persisting and delivering the essential characteristics of digital objects in Fedora[14] and Mulgara serves as a triple store for a resource index[15].

Islandora – Integration Layer is connected with Fedora through Tuque API in a form of Drupal library[16]. Tuque API is a PHP library that allows managing objects and data streams in Fedora. Bridge module establish a connection between Drupal and Fedora, with Drupal-filter define an access to Fedora. Depending on repository's specification this layer implements appropriate software. For example, if OCR tool required then tesseract can be used[17]. In the case when presentation of material is needed then djatoka[18] or Cantaloupe[19] can be used. This layer integrates software tools installed on server levels.

Drupal – Integration Level is a workflow constructed on Drupal CMS and serves for administration of digital repository and public access to digitized material. In this level Drupal expansion modules and themes are in use. This level is used also add and define users and their roles and finally control the whole repository.

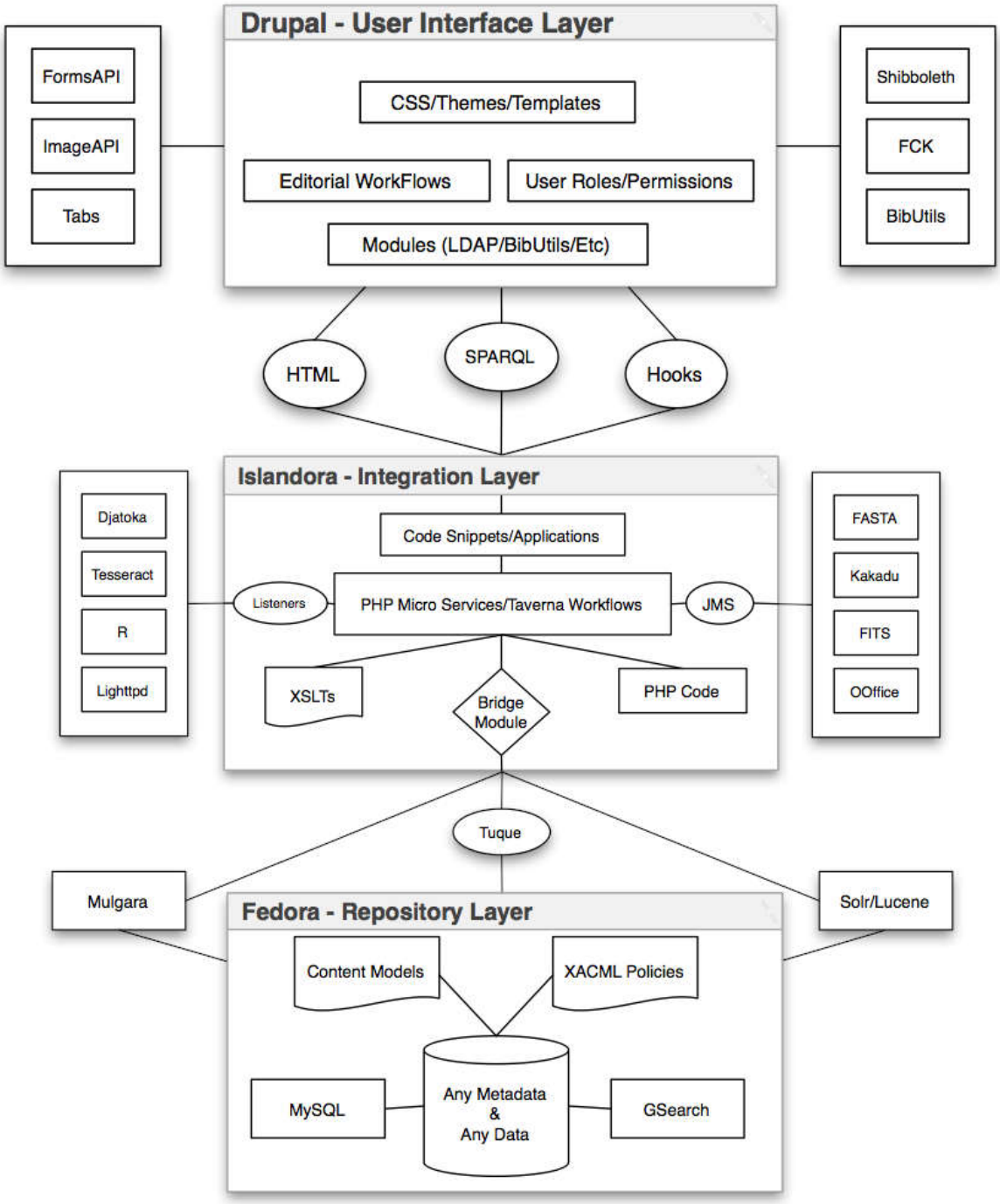

Figure 1. Islandora diagram[12]

4. Islandora Solution Packs

"Islandora Solution Packs (ISP) provide the framework for the ingestion, organization, and display of digital assets in a Fedora repository through the frontend Drupal web interface"[20], they are made as Drupal modules. ISP combine prechosen Content Models, Metadata Forms, and Viewers with appropriate software dependencies. There are 12 Solution Packs:

1) Audio Solution Pack
2) Basic Collection Solution Pack
3) Basic Image Solution Pack
4) Book Solution Pack
5) Compound Solution Pack
6) Disk Image Solution Pack
7) Entities Solution Pack
8) Large Image Solution Pack
9) Newspaper Solution Pack
10) PDF Solution Pack
11) Video Solution Pack
12) Web Archive Solution Pack

Object is made of data streams. There are three types of objects:

- Content Model Objects. **Content Model Object** is a model to work with a specific content type and servers to build an object through an ISP.
- Collection Objects. **Collection Objects** are Fedora objects which gather objects made with ISP.
- Data Objects. **Data objects** are files ingested into the repository and any associated metadata, derivatives, or related files that should be managed as a single digital asset in the repository.

Objects have a relationships, states and unique, persistent identifiers (PID). Relationships of objects are stored in RELS-EXT data stream, written in RDF/XML file. Object can be a part of a collection, shared with different collections, part of one particular object, etc.[21]. The states of objects are “Active”, “Inactive” and “Deleted”. Unique, persistent identifiers (PID) is assigned to every object and can be managed by user according to needs, there are no two identical PIDs.

RELS-EXT	Fedora Object to Object Relationship Metadata.
MODS	MODS Record
DC	Dublin Core record
TN	Thumbnail image
PDF	PDF derivative created by ImageMagick
OCR	Consolidated OCR

Figure 2. List of data streams in a digital object made by Book Solution Pack

Book Solution Pack can be distinguished among packages. Book Solution Pack can be used to create a digital object from books with possibility to OCR, add and remove pages, preview in appropriate viewer, and/or create downloadable PDF file. Book SP depends on following mandatory and optional modules:

- Islandora (mandatory)
- Tuque (mandatory)
- Islandora Paged Content (mandatory)
- Large Image Solution Pack (mandatory, required for creating thumbnail and JPEG data streams)
- Islandora OCR (mandatory, required for creating OCR data streams)

- Islandora Internet Archive Bookreader (optional, can be used to view and search text in book)
- OpenSeadragon (optional, can be used to view pages)
- ImageMagick (optional, required for creating PDF data streams).

The process of creating an object using Book Solution Pack has the following steps:

1. Choose a Collection and add a new object in that collection (make relation of object and collection)
2. Import MARCXML metadata or create a metadata in a predefined form
3. Import a pdf file or import a zipped image files.
4. Create Page derivative PDF, thumbnail, jpeg, jp2 and hocr.

Ingested book can be viewed either with Internet Archive Bookreader and pdf.js reader. Pages can be seen with OpenSeadragon for more detailed preview. Book prepared in this process has full text searching possibility, option to download it as a pdf version, and a choose of a format for displaying metadata, MODS record translated to DC or MODS in defined Solr fields.

5. Documentation and support

Islandora has an excellent Documentation which can be found on address https://wiki.duraspace.org/display/ISLANDORA/Start. All elements are published on Github under the GPL-3.0 license and can be found here https://github.com/Islandora. Support is organized in two Google Groups by community:

- Users Google Group (https://groups.google.com/forum/?hl=en#!forum/islandora)
- Developers Google Group (https://groups.google.com/forum/#!forum/islandora-dev)

Also, there is a JIRA Ticketing System to report a problem in a specific module or system.

6. Digital repository of "Milutin Bojić" Library

In "Milutin Bojić" Library repository there are three collections:

1. Poet Milutin Bojić collection of books, manuscripts, articles and photographs
2. Collection of an edition of Serbian writers library published in 1920s
3. Complete edition of "Umetnički Prelgled" (Art Preview) magazine

For Milutin Bojić collection appropriate ISP is searched regarding the type of material, for published books and articles a Book Solution Pack is used, for photographs Large Image Solution Pack, but for manuscripts appropriate solution pack is missed.

OpenSeadragon image viewer supports both single and multi-image view, the task was to use this multi-image property of the viewer to present manuscripts. For manuscript preview [22], Manuscript Solution Pack is created to get a Solution Pack needed to appropriately display this type of digitized material (Figure 3).

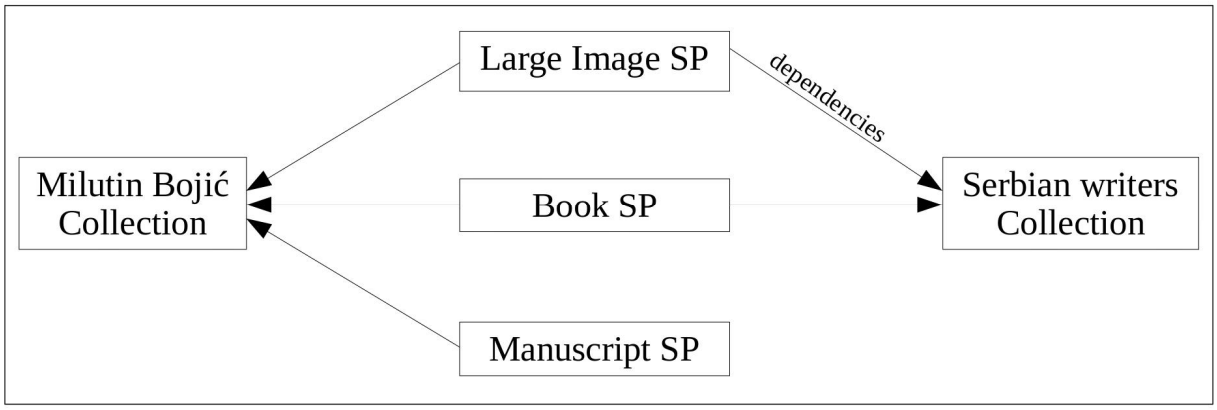

Figure 3. Islandora Solution Packs used for Collections

With included additional modules a better presentation, administration and metadata view is available.

The following modules were used for search:

- Solr modules Islandora Solr Metadata
- Islandora Solr Facet Pages,
- Islandora Solr Views
- Islandora Solr.
- For metadata creation is used:
- Islandora XML Forms
- Islandora MARCXML modules.

For OAI-PMH metadata sharing, Islandora OAI module is used, also it is verified and registered as data provider[23].

Available viewers are Internet Archive BookReader v2 and OpenSeadreagon.

Drupal theme for repository design is Bootstrap which has a good responsive feature.

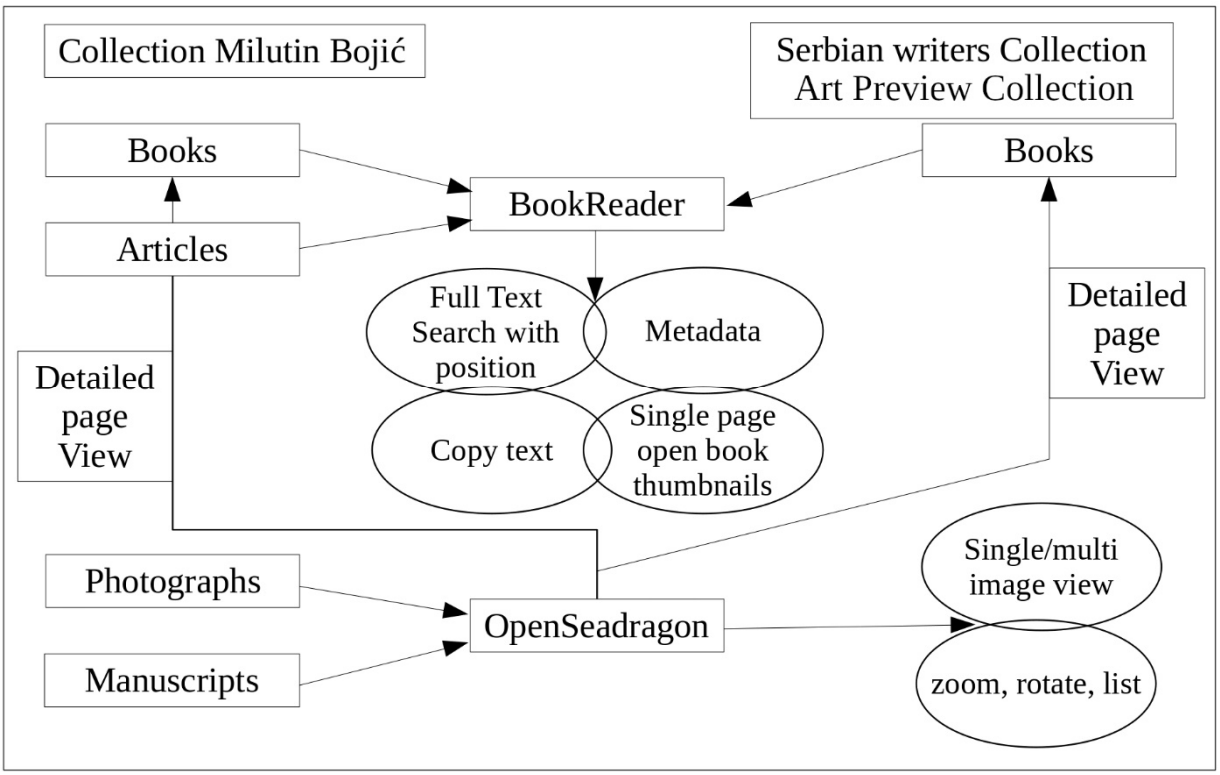

Figure 4. Collection viewers

Figure 4 shows presentation of objects in collections, what viewers are used and shows a feature that viewers have. There are only two viewers:

- BookReader v2
- OpenSeadragon

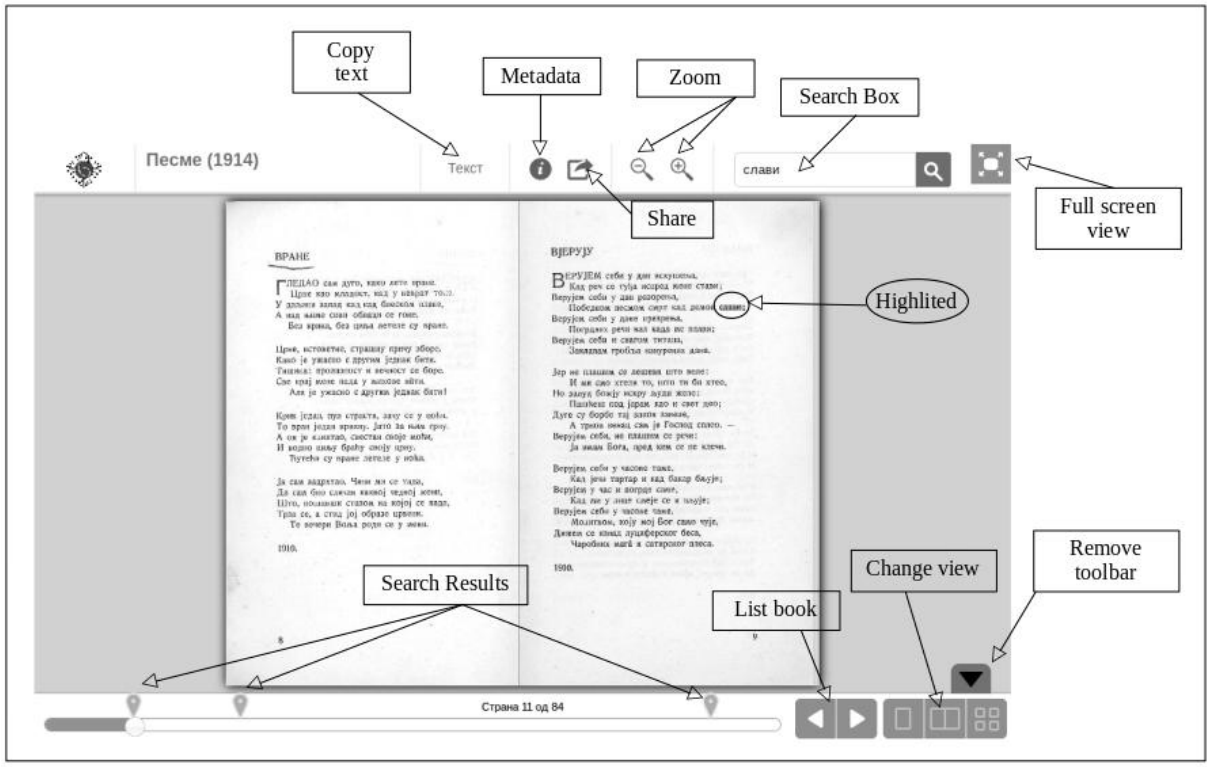

Figure 5. Internet Archive BookReader

BookReader (Figure 5) serves for display Books and Articles and includes all needed derivatives (images and OCR) for a functional presentation to the user. Results of full text searching, using a search box of all repository or using a Book Reader search box (Search Box), give an exact position of searched words in pages (Search Results) and highlight it on a page (Highlighted). Textual page content can be copied to word processor or a note (Copy text). Metadata table shows fields defined with Solr. There are different configurations for previewing one page, two pages or thumbnails (Change view). Zoom option is available. Field for sharing an object via e-mail and social networks, either a specific page(s) or object (Share). Full screen view allows better display (Full screen view). To hide/show toolbar use Remove toolbar. All derivatives and data streams are delivered through BookReader. Images used in display are in JP2 format.

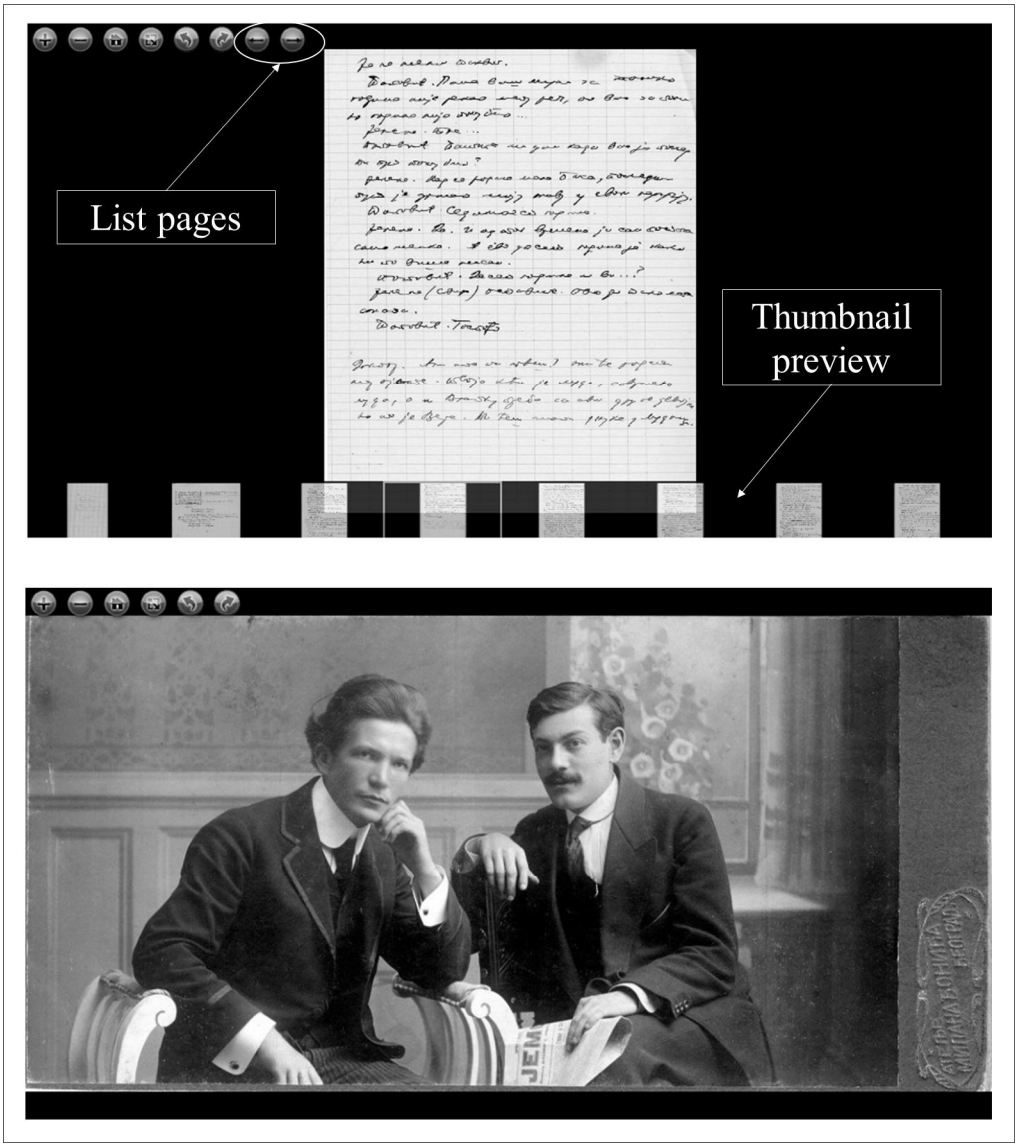

Figure 6. OpenSeadragon for manuscript and large image

OpenSeadragon delivers images fast and in good quality. As an example, in Figure 6 you can see the difference between two environments and elements used in it. Basic functions are:

- Zoom in
- Zoom out
- Go home (reset zoom)
- Toggle full page (show in full screen)
- Rotate (one click for 90 degree left or right)

The first example (up), manuscript, is displayed as multi-image, it shows additional functions:

- Thumbnails (Thumbnail preview)
- Next page and Previous page (List pages)

The second example (down) display a single image with basic functions.

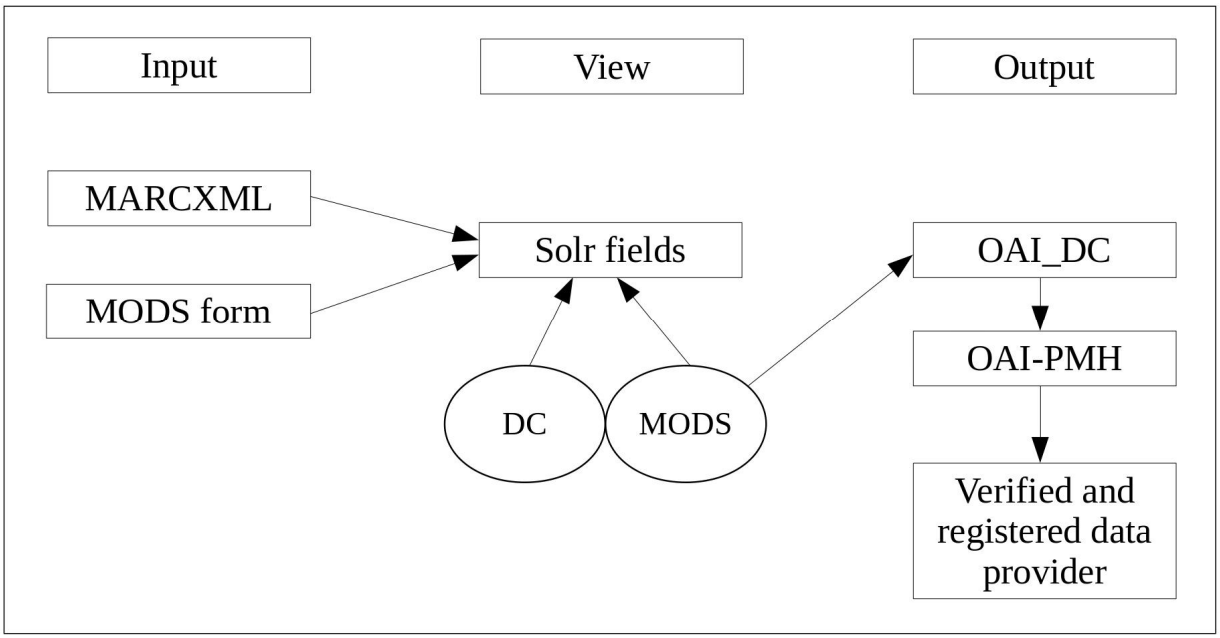

Figure 7. Distribution of metadata

Creation of metadata and metadata import is organized through modules MARCXML and XML Form Builder. MARCXML module serve to import a xml file and translate it to MODS standard. XML Form Builder[24] module serve to administer metadata scheme, define number and names of fields, which fields are obligatory and connect metadata scheme to specific content.

For metadata view Solr metadata[25] view option is used because it gives more flexibility in defining a list of fields to display. Sharing of metadata goes over registered and verified OAI-PMH protocol. Collection of poet Milutin Bojić is successfully imported in WorldCat[26] catalogue, via free OCLC service Digital Gateway, with thumbnails and links to objects. Metadata is also tested on REPOX[27].

Planning and creating a digital repository lasted two years. The main collection is digitized material from poet Milutin Bojić heritage. Thanks to excellent cooperation with Bojić family, National Library of Serbia and Archive of SANU it was possible to collect and digitize all written heritage of poet Milutin Bojić, which includes, for example, school notebooks, preparation versions of poems, manuscripts of drama, personal library, etc.

The second collection is Serbian writers library, it consist of 54 books of great Serbian writers.

The third collection is “Umetnički pregled”, editor was Milan Kašanin. This is one excellent example how institutes and researchers can use a digitized material. Institute for Literature and Arts in Belgrade organize international conference named “Visual and literary in the journal Umetnički pregled (1937–1941)”. Source for exploration is digitized material in this collection.

Digital repository of “Milutin Bojić“ Libraryconsist of 141 digital objects and over 16.000 searchable text pages. It is accessible on https://milutinbojic.digitalna.rs.

7. Conclusion

The configuration of a digital repository in Islandora is very flexible and adaptive thanks to the modules extensions made for a specific need, types of digitized material and administration a repository. There are over 50 modules developed to make a repository that will fit your needs. While working on this digital repository, Library "Milutin Bojić" became an active member of Islandora Community. The whole infrastructure, operations, additional software support and flexibility are made to ease, as much as possible, working with digitized material, administer repository and on frontend, for users, an attractive look and interactive website. This approach to build a modular repository on a Drupal CMS as an interface layer made a good balance between stable storing (objects and metadata), administration and presentation.

References

[1] Metadata Object Description Schema. http://www.loc.gov/standards/mods/
[2] The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form. https://www.loc.gov/marc/
[3] The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. http://dublincore.org/documents/dces/
[4] Provides an application-independent interoperability framework based on metadata harvesting which allow us to share metadata to third parties.
https://www.openarchives.org/OAI/openarchivesprotocol.html
[5] FOSS mean Free Open Source Software
[6] Bankier, J.G., Gleason, K. (2014). Guide to Institutional Repository Software. UNESCO.
[7] https://github.com/Islandora/islandora/wiki/Programming-Solution-Packs
[8] https://github.com/Islandora-Labs/islandora_awesome
[9] https://duraspace.org/fedora
[10] https://www.drupal.org
[11] http://lucene.apache.org/solr
[12] Images source from Islandora Documentation
[13] More on Fedora XACML policies see
https://wiki.duraspace.org/display/FEDORA38/XACML+Policy+Enforcement
[14] More on Content Model Architecture see
https://wiki.duraspace.org/display/FEDORA36/Content+Model+Architecture
[15] More on Resource Index
seehttps://wiki.duraspace.org/display/FEDORA38/Resource+Index
[16] Drupal extensions are modules and Libraries API. Libraries API brings third parties libraries which are not included in modules but can be shared in different modules. You can distribute one software on various modules.
[17] Tesseract is Open Source OCR engine which can be find on https://github.com/tesseract-ocr/tesseract also it is in many repos of Linux distributions.
[18] https://sourceforge.net/projects/djatoka
[19] https://medusa-project.github.io/cantaloupe
[20] https://wiki.duraspace.org/display/ISLANDORA/About+Islandora
[21] More on Fedora relationship see
https://wiki.duraspace.org/display/FEDORA35/Digital+Object+Relationships
[22] See it on https://milutinbojic.digitalna.rs/islandora/object/mb%3Arukopisi
[23] https://www.openarchives.org/Register/BrowseSites
[24] https://wiki.duraspace.org/display/ISLANDORA/XML+Form+Builder
[25] https://wiki.duraspace.org/display/ISLANDORA/Islandora+Solr+Metadata
[26] https://www.worldcat.org
[27] 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. https://pro.europeana.eu/data/repox

andrija.sagic@milutinbojic.org.rs