**ECP 2006 DILI 510049**

**ENRICH**

# Report on pilot full integration and publication of selected partner's metadata and externally stored data in Manuscriptorium

| | |
|---|---|
| **Deliverable number** | *D 5.3* |
| **Dissemination level** | *Public* |
| **Delivery date** | *1.6.2009* |
| **Status** | *Final* |
| **Author(s)** | *Tomáš Psohlavec, AIP* |

***e*Content*plus***

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

## Document Version Control

| Version | Date | Change Made (and if appropriate reason for change) | Initials of Commentator(s) or Author(s) |
|---|---|---|---|
| 0.1 | 20/05/2009 | Draft version. | TP, AIP |
| 0.2 | 28/05/2009 | Links added | TP, PSNC |
| 1.0 | 1/06/2009 | Final version, formatting done | JH, CCP |
| | | | |
| | | | |
| | | | |

## Document Review

| Reviewer | Institution | Date and result of the review |
|---|---|---|
| Tomasz Parkola | PSNC | 28.5.2009, document approved with small amendments |
| Jakub Heller | CCP | 28.5.2009, document approved |

| **Approved By (signature)** | **Date** |
|---|---|
| | |

| **Accepted by at European Commission (signature)** | **Date** |
|---|---|
| | |

# 1 Executive Summary

Deliverable 5.3 summarizes the progress of work done and the results achieved during the process of integration of selected partner's sources into the Manuscriptorium platform.

It reports the full integration and publication of selected partner's metadata and data. As full metadata and data integration of two of the partners is already reported in the previous D5.2 report we selected one additional interesting use case - the hereby reported partners' integration approach is interesting from the technical point of view and also is important for future cooperation with DjVu based sources.

## CONTENT

# 2 Introduction

This deliverable links to the previous deliverables mainly those prepared within WP2 and the subsequent deliverables from WP5:

- D2.1: Survey results and their interpretation (Month 3), responsible partner: NKP
- D2.2: Description of the standards used by the partners, definition of collaboration principles, data and metadata standards (Month 8), responsible partner: NKP
- D5.1: Definition of basic conditions for sharing of large data sets in the frame of Manuscriptorium (Month 10), responsible partner AIP
- D5.2: Report on pilot full integration and publication of selected partner's metadata in Manuscriptorium

The partner reported here is the University Library Wroclaw (ULW) (http://www.bu.uni.wroc.pl/en/) which provides documents for aggregation via the dLibra Digital Library Framework (http://dlibra.psnc.pl/) operated by Poznań Supercomputing and Networking Center (PSNC, http://www.man.poznan.pl/).

The way of cooperation is based on OAI-PMH harvesting and subsequent large-extent batch processing of metadata within the Manuscriptorium input tools ("Connectors").

This way of cooperation is described in previous deliverables.

# 3 Integration of partners' metadata

## *3.1 Short collection description*

The digital collection of ULW provides access to manuscripts, music manuscripts, old drawings and early printed books. It is a selection of the most valuable physical collections of ULW that is the largest historical manuscript library in Poland.

At present the following amount of data is fully integrated:

- 850 documents
- 67 310 images

### 3.1.1 Metadata

The University Library Wroclaw (ULW) provides metadata via OAI-PMH interface at http://bibliotekacyfrowa.pl/dlibra/oai-pmh-repository.xml?verb=Identify. Uses the Open Archives Initiative Protocol for Metadata Harvesting, protocol version 2.0 (see documentation at http://www.openarchives.org/OAI/openarchivesprotocol.html).

The library provided a specific set of documents for the harvesting purposes. The set is based on a dynamic query which is supplied to the interface using the set argument. This approach is fully compatible with usual usage of the set argument and enables efficient modification of the content of harvested set of documents.

The available profile works with METS format which is used to wrap all the descriptive and structural metadata (see http://www.loc.gov/standards/mets/ for METS details).

The structural metadata is carried by the METS structural maps. To carry the other types of metadata the following formats are used within the METS:

- Local metadata format (dlibra_avs)
- DC (http://dublincore.org/documents/dces/)

See Cooperation details chapter for further information. Also following schemas are provided here as an example demonstrating the extent of information provided via the OAI interface - see the Appendix A – Source metadata for detail on partner's metadata.

### 3.1.2 Data

As stated in the previous deliverables - the main condition of seamless integration of partner's documents into the single homogenous end-user interface requires usage of image formats commonly supported by web browsers (without the need to install/use additional plug-ins ), in other words: use of JPEGs, GIFs, PNGs for image data is required at present.

Thus we frequently encounter partners who use DjVu format – especially for the newer documents (e.g. early printed books) - and ULW is such type of partner. Within ENRICH a new interesting solution was found which is now being tested as a pilot solution with PSNC and ULW.

The introduced solution presented here is based on on-line DjVu to JPEG conversion. Thanks to that, particular pages are now available as JPEGs - for the collection of ULW three different JPEG conversion outputs are prepared as „virtual" quality levels. The quality levels in the pilot are as follows:

- big (default quality for work with a document)
- preview (for quick orientation within a document)
- thumbnail (small thumbnails)

The JPEGs can be accessed using the information provided by METS structural maps in the harvested records.

The solution is prepared by PSNC directly within the dLibra system and it complies with the cooperation conditions and the performance is also equal to other repositories integrated within Manuscriptorium so far.

The advantage of this solution lies also in the possibility to integrate many other sources from Poland which use the same dLibra system.

### 3.1.3 Cooperation details

The collection is accessible at http://bibliotekacyfrowa.pl/dlibra/oai-pmh-repository.xml. The particular set argument value is:

```
BCUWr:criteria:dc.date%3E%3D1000%20and%20dc.date%3C%3D1850%20not%20dc.type%3Dczasop*%20not%20dc.type%3Dprog*%20and%20dc.format%3D(app*%20or%20pd*%20or%20dj*%20or%20im*)
```

Routine harvest is processed using the "`mets_exp`" metadata prefix.

The appropriate XSL transformation is applied within the connector. The XSL transformation converts the METS record to the TEI P4 based MASTER and also to the new TEI P5 ENRICH Schema.

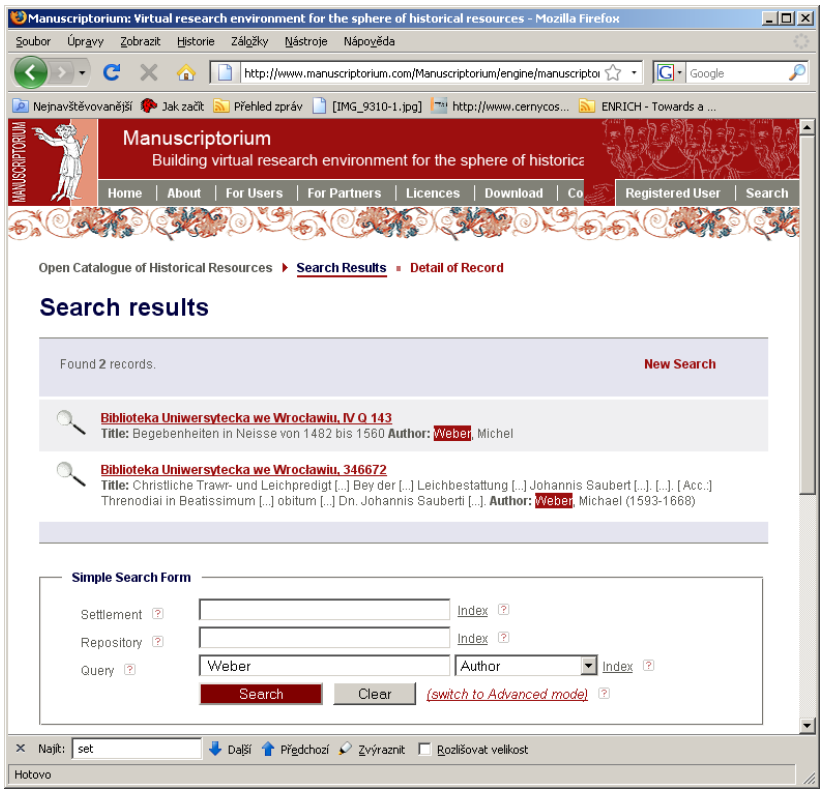
See the Appendix B - TEI based converted outputs for more information on the results of conversion.

### 3.1.4 Results

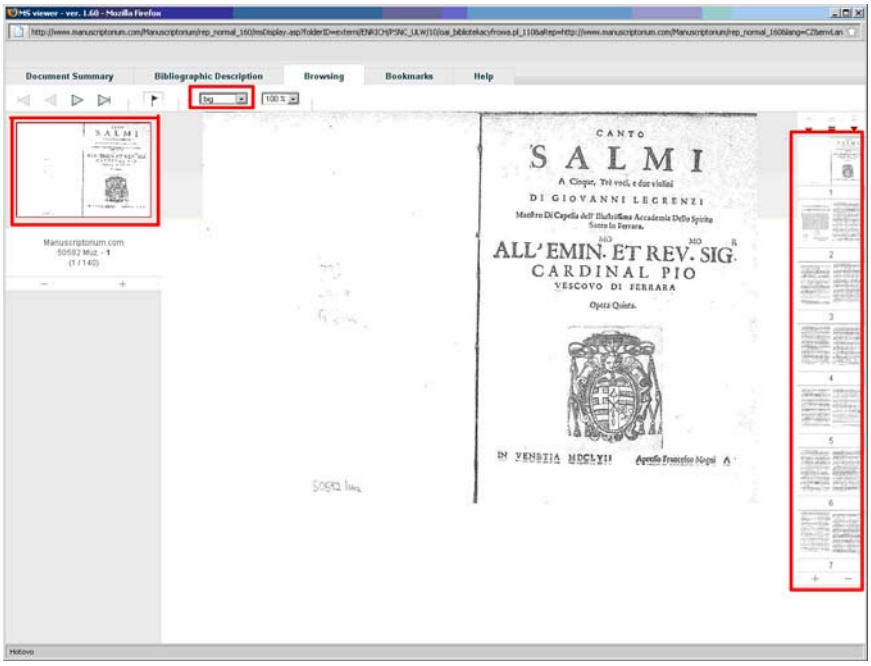
No serious problems were identified during the cooperation thanks to the availability, consistency and overall quality of the available metadata and the accessibility of the generated JPEG images from the partner's repository and above all an interesting new technical solution was implemented and verified.

The partner's documents are now fully integrated.

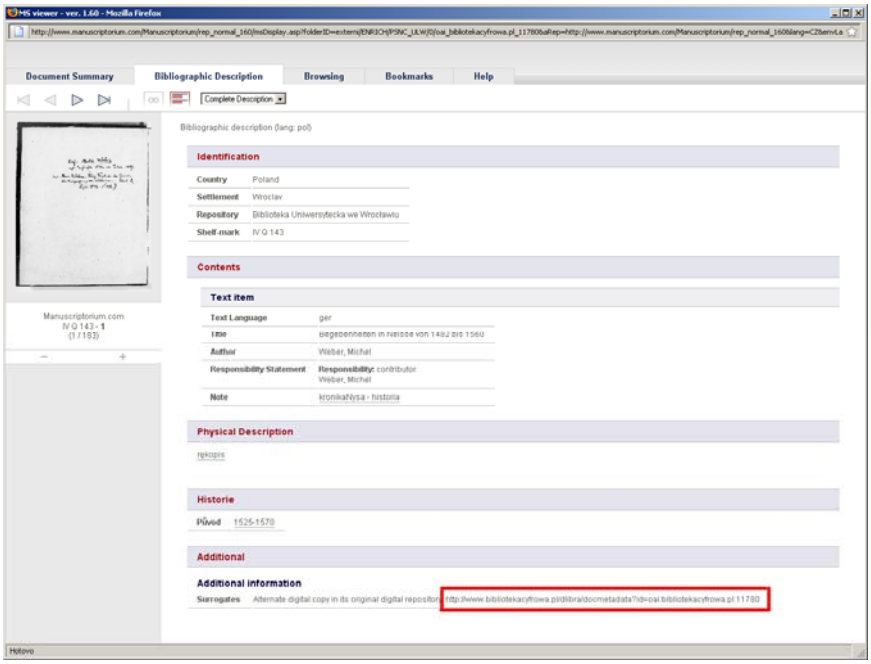
The records are searchable via the Manuscriptorium catalogue research interface. An example of search for „Weber" author:




After opening the digital document it is fully browsable and all quality levels are available to the end user. The processed metadata serve to display the document in a desired way and the end-user is neither affected by the distributed data transfers nor the DjVu->JPEG conversion process in any way:




Link to an alternate copy of the digital document is available directly from the description:



# 4 Conclusion

According to the current experience (see also D5.2 describing the integration of Heidelberg University Library and National Library of Romania) the results achieved with partners using different ways of cooperation are very similar. This indicates that the approaches are well designed for individual needs of different partners and their different digitization approaches and particular digitization projects' results.

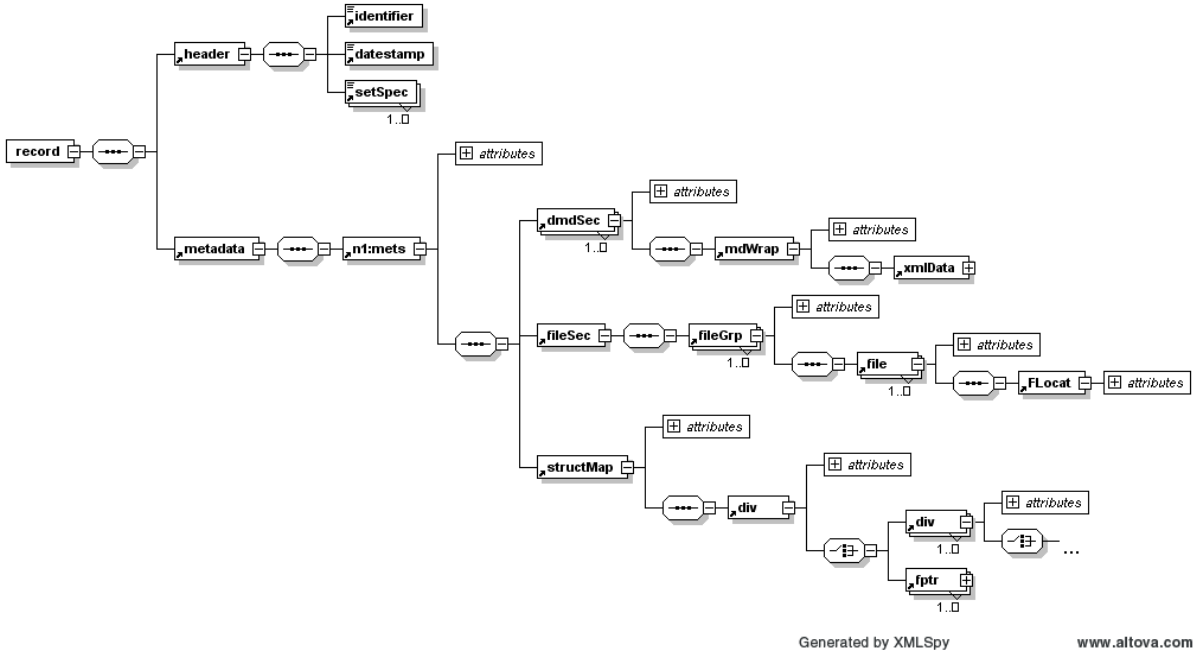
The end users of the presentation platform are provided with the documents in a homogenous interface and can work efficiently with the European digital content related to the cultural heritage. The users work without being affected by the original form of the digital documents, all the processes are hidden behind the presentation interface and the users are not being disturbed by the searching for shattered content.

The pilot solution confirmed the intended ways of cooperation as properly designed and all present and future cooperating partners will cooperate on the basis of the similar principles.

# Appendix A – Source metadata

## *University Library Wroclaw*

Record harvested via OAI with METS, structural maps and wrapped descriptive metadata:




Wrapped descriptive metadata (both DC and Local formats):


The structural metadata consist of physical structure `structMap[@TYPE="PHYSICAL"]`) and file section (`fileSec`). Different `mets:fileGrp` elements are used to group files of available qualities, appropriate `@USE` attribute is applied:

```
<fileGrp USE="thumbnail">
      <file ID="thumbnail0" MIMETYPE="image/jpeg">
            <FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="http://bibliotekacyfrowa.pl/Content/110/d2j:thumbnail,0/0001_00
01.djvu.jpg"/>
      </file>
      <file ID="thumbnail1" MIMETYPE="image/jpeg">
            <FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="http://bibliotekacyfrowa.pl/Content/110/d2j:thumbnail,1/0002_00
01.djvu.jpg"/>
      </file>
      <file ID="thumbnail2" MIMETYPE="image/jpeg">
            <FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="http://bibliotekacyfrowa.pl/Content/110/d2j:thumbnail,2/0003_00
01.djvu.jpg"/>
      </file>
..............
..............
..............
</fileGrp>
```

Following physical structural map is applied:

```
<structMap TYPE="PHYSICAL">
      <div ID="physicalStructure" DMDID="dmd0 dmd1" TYPE="pageSequence">
            <div ID="phys0" ORDER="0" ORDERLABEL="0" TYPE="page">
                  <fptr FILEID="original0"/>
                  <fptr FILEID="thumbnail0"/>
                  <fptr FILEID="preview0"/>
                  <fptr FILEID="big0"/>
            </div>
            <div ID="phys1" ORDER="1" ORDERLABEL="1" TYPE="page">
                  <fptr FILEID="original1"/>
                  <fptr FILEID="thumbnail1"/>
                  <fptr FILEID="preview1"/>
                  <fptr FILEID="big1"/>
            </div>
..............
..............
..............
      </div>
</structMap>
```

## Appendix B - TEI based converted output

```
<TEI>
	<teiHeader>
		<fileDesc>
			<titleStmt>
				<title/>
			</titleStmt>
			<publicationStmt>
				<p/>
			</publicationStmt>
			<notesStmt>
				<note>4ş</note>
				<note>S, B, bc, vl 1, vl 2</note>
			</notesStmt>
			<sourceDesc>
				<msDesc xml:id="oai_bibliotekacyfrowa_pl_110"
xml:lang="">
					<msIdentifier>
						<country>Poland</country>
						<settlement>Wroclav</settlement>
						<repository>Biblioteka Uniwersytecka we
Wrocławiu</repository>
						<idno>50582 Muz.</idno>
					</msIdentifier>
					<msContents>
						<textLang mainLang="lat">lat</textLang>
						<msItem>
							<title>Salmi a cinque, trč voci,
e due violini [...] opera quinta.</title>
							<author>Legrenzi, Giovanni (1626-
1690)</author>
							<respStmt>
								<resp
key="unk">contributor</resp>
								<name
type="unknown">Legrenzi, Giovanni (1626-1690)</name>
							</respStmt>
							<note>
								<index>
									<term
type="subject">17 w.</term>
									<term
type="subject">muzyka wokalno-instrumentalna</term>
									<term
type="subject">Włochy</term>
								</index>
							</note>
						</msItem>
					</msContents>
					<physDesc>
						<p>
							<index>
								<term>druk muzyczny</term>
								<term>stary druk</term>
							</index>
						</p>
					</physDesc>
					<history>
						<origin>
							<origDate>1657</origDate>
							<persName>Magni,
Francesco</persName>
```

```
                                          <placeName>Wenecja</placeName>
                                    </origin>
                              </history>
                              <additional>
                                    <surrogates>
                                          <p>Alternate digital copy in its
original digital repository: <ref
target="http://www.bibliotekacyfrowa.pl/dlibra/docmetadata?id=oai:bibliotek
acyfrowa.pl:110"/>
                                          </p>
                                    </surrogates>
                              </additional>
                        </msDesc>
                  </sourceDesc>
            </fileDesc>
      </teiHeader>
      <facsimile xml:base="http://bibliotekacyfrowa.pl/Content/110/">
            <surface xml:id="id1">
                  <desc>
                        <label>1</label>
                  </desc>
                  <graphic url="d2j:big,0/0001_0001.djvu.jpg"/>
                  <graphic url="d2j:preview,0/0001_0001.djvu.jpg"/>
                  <graphic url="d2j:thumbnail,0/0001_0001.djvu.jpg"/>
            </surface>
            <surface xml:id="id2">
                  <desc>
                        <label>2</label>
                  </desc>
                  <graphic url="d2j:big,1/0002_0001.djvu.jpg"/>
                  <graphic url="d2j:preview,1/0002_0001.djvu.jpg"/>
                  <graphic url="d2j:thumbnail,1/0002_0001.djvu.jpg"/>
            </surface>
...........
...........
...........
...........
            <surface xml:id="id139">
                  <desc>
                        <label>139</label>
                  </desc>
                  <graphic url="d2j:big,138/0139_0001.djvu.jpg"/>
                  <graphic url="d2j:preview,138/0139_0001.djvu.jpg"/>
                  <graphic url="d2j:thumbnail,138/0139_0001.djvu.jpg"/>
            </surface>
            <surface xml:id="id140">
                  <desc>
                        <label>140</label>
                  </desc>
                  <graphic url="d2j:big,139/0140_0001.djvu.jpg"/>
                  <graphic url="d2j:preview,139/0140_0001.djvu.jpg"/>
                  <graphic url="d2j:thumbnail,139/0140_0001.djvu.jpg"/>
            </surface>
      </facsimile>
</TEI>
```