Grant Agreement Number **ECP 2006 DILI 510049**

**ENRICH**

# Description of the standards used by the partners, definition of collaboration principles, data and metadata standards

| | |
|---|---|
| **Deliverable number** | *D2.2* |
| **Dissemination level** | *Restricted* |
| **Delivery date** | *8 September 2008* |
| **Status** | *Final* |
| **Author(s)** | *Tomas Psohlavec, Zdeněk Uhlíř* |

***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) |
|---|---|---|---|
| 0.0 | 1/07/2008 | Discussion draft developed by AIP | TP |
| 0.1 | 14/07/2008 | Revised by NKP | ZU |
| | | | |

## Document Review

| Reviewer | Institution | Date and result of the review |
|---|---|---|
| Zdeněk Uhlíř | NKP | 14 July 2008 (sent back with comments) |
| Gabriella Lovasz | CCP | 8 August 2008 (language check) |
| Stanislav Psohlavec | AIP | 29 August (approved for submission to EC) |
| | | |

| **Approved By (signature)** | **Date** |
|---|---|
| | 8 September 2008 |

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

# 1 Executive Summary

This document links to the deliverable D2.1 (Survey results and their interpretation) and provides a comprehensive overview of metadata standards as applied by particular Content partners and defines final methods and principles of collaboration. As such it provides one of the most important inputs to the implementation of WP5 tasks.

# CONTENT

# 2 Resources available for aggregation

## *2.1 Metadata standards*

The metadata standards are described for each Content Partner's collection individually below. The reasons to the individual approach were described in the deliverable D2.1. In short summary these reasons are:

- different types of original documents
- different cataloguing practices, tools and formats
- different approaches to digitisation

## *2.2 Methods of collaboration*

There were 4 (5) typical ways of cooperation defined in the D2.1. The ongoing collaboration with Content Partners and the deeper analysis realised of the particular sources of the content led us into a re-definition of these ways.

Finaly for the Content Partners there are the follwoing possible methods of collaboration:

1. M-Tool
   (creating of the structural and descriptive metadata for individual documents)
2. Off-line connector
   (converting existing structural and descriptive metadata content off-line)
3. On-line connector
   (converting existing structural and descriptive metadata content using OAI-PMH interface)
4. Offline automated generation of **structural** metadata and connector of existting **descriptive** metadata (see below for details)

**Note:** the other ways of cooperation as originaly defined in the D2.1 may be still applicable to possible cooperating Associated Partners.

# 3 Content partners in detail

The ENRICH quality objectives are to set quality measures, to provide support to consortium partners to achieve these and monitor adherence to the Quality Plan throughout the project's lifecycle. The Quality Plan is designed to provide for the assurance of quality, according to the main ENRICH Project characteristics.

## *3.1 BNCF*

### 3.1.1 Metadata standards

| All available collections | |
|---|---|
| **Descriptive metadata** | **UNIMARC slim** in the form of XML<br>UNIMARC slim format conforms regular cataloguing guides and will be converted to TEI P5 using a dedicated connector.<br>Possible derivations from the cataloguing manual guides and content of possible national specifics fields will be discussed with the responsible BNCF cataloguers.<br>UNIMARC slim profile is agreed to be included later during ENRICH progress. |
| **Structural metadata** | **MAG**<br>The MAG format contains not only a physical structure information but also the valuable applicable foliation/application (which is appreciated by the end-users as the researchers can refer to or can be referred to a particular page/folio both in the original AND the digitised document. |
| **Additional metadata related notes** | BNCF uses a DC format to share bibliographic descriptions (DC based profile is mandatory for OAI-PMH interface). The records contain only a substraction of the original structure, therefore will not be processed. |

### 3.1.2 Method of collaboration

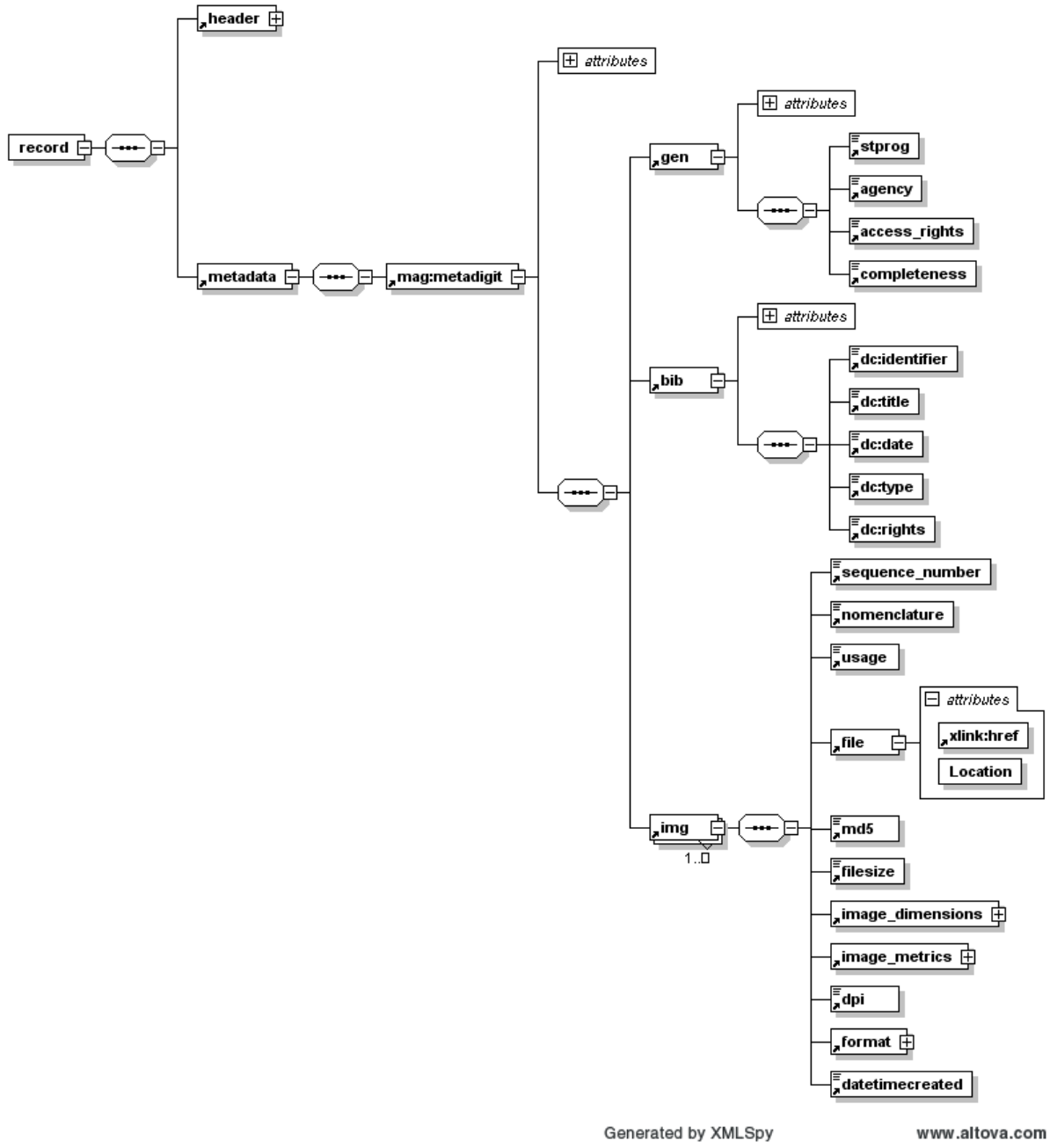
OAI-PMH based harvesting of both UNIMARC slim bibliographic records and the MAG structural metadata with subsequent processing and conversions within a BNCF dedicated connector.

Data elements are available via a HTTP by calling a script http://teca.bncf.firenze.sbn.it/TecaFrontEnd/servlet/readImg
with appropriate parameters, for instance
http://teca.bncf.firenze.sbn.it/TecaFrontEnd/servlet/readImg?RisIdr=BNCF0003471607&usage=3
where `RisIdr` is a particular image ID and `usage` is a requested image quality level indicator.

The full link to file can be read in the
`//mag:img/mag:file[@Location='URL']/xlink:href`, e.g.

```
<mag:file Location="URL"
xlink:href="http://teca.bncf.firenze.sbn.it/TecaFrontEnd/servlet/readImg?RisIdr=BNCF0003465109&amp;usage=3"/>
```

The applicable foliation/pagination can be found in the element `//mag:img/mag:nomenclature`.




### 3.1.3 Current results

The MAG profile has been harvested in the testing mode and processed with the most basic descriptive information it contains (`mag:bib`). The resulting documents were presented in the clone at enrichdata.manuscriptorium.com.

### 3.1.4 Remaining steps

To add the UNIMARC slim profile, set up conversions, include into the connector, bring the harvesting and the connector into the routine service.

### *3.2 Quality System*

The quality system applied on the project is described in the present Quality Plan and subsequent revisions to it.

### *3.3 BNE*

## 3.3.1 Metadata standards

| **Manuscripts, maps** | |
|---|---|
| **Descriptive metadata** | **MARC 21** in the form of XML<br>The MARC 21 format conforms regular cataloguing guides and will be converted to TEI P5 using a dedicated connector.<br>Possible derivations from the cataloguing manual guides and content of possible national specifics fields will be discussed with the responsible BNE cataloguers. |
| **Structural metadata** | **none**<br>There are no structural metadata that would enable the construction of documents according the basic principles of cooperation within the ENRICH project.<br>The metadata are being prepared during the project to enable the collaboration. |
| **Additional metadata related notes** | BNE uses also a DC format to present the bibliographic descriptions (DC based profile is mandatory for OAI-PMH interface). The records contain only a substraction of the original structure, therefore will not be processed. |

## 3.3.2

| **Incunabulas, old printed books** | |
|---|---|
| **Descriptive metadata** | **MARC 21** in the form of XML<br>(see above) |
| **Structural metadata** | **none**<br>The documents have a form of PDF (one PDF = one physical document). See below for additional information. |

## 3.3.3 Method of collaboration

### 3.3.3.1 Manuscripts, maps

Off-line metadata passing (via FTP or other method) with subsequent processing and conversions within a BNE/manuscrpits and BNE/maps dedicated connectors.

The missing structural metadata are created via an automated process by reading names of images above the complete folder structure of the appropriate repository. The output of such process does have a form of and XLS file with a table filled by folder and filenames.

During subsequent processing of such a XLS file an XML with necessary structural metadata is created. The filenames contain often an applicable foliation in some forms – therefore it is possible to create high quality structural metadata (including the applicable foliation to enable efficient folio referring).

The process is fully automated and the metadata can be created and/or updated for the complete repository in a batch process. The initial step of the batch is prepared by the BNE (the output is the XLS), the last step is performed in the BNE dedicated connector.

The input looks like (note the applicable foliation information in the filename):

| | | | | | |
|---|---|---|---|---|---|
| 3532 | 1082684_Res_000045-02_365 | | jpg | 0 MB | -a--- |
| 3533 | u:\ENRICH\1096513\ | 2 | | 1 MB | |
| 3534 | 1096513_Vitr_000004-021_800ppp | | jpg | 1 MB | -a--- |
| 3535 | 1096513 | | xml | 0 MB | -a--- |
| 3536 | u:\ENRICH\1096516\ | 2 | | 5 MB | |
| 3537 | 1096516_Vitr_000005-011_001 | | jpg | 5 MB | -a--- |
| 3538 | 1096516 | | xml | 0 MB | -a--- |
| 3539 | u:\ENRICH\1096519\ | 7 | | 28 MB | |
| 3540 | 1096521_Vitr_000006-006_195r | | jpg | 5 MB | -a--- |
| 3541 | 1096522_Vitr_000006-006_195v | | jpg | 5 MB | -a--- |
| 3542 | 1096523_Vitr_000006-006_196r | | jpg | 5 MB | -a--- |
| 3543 | 1096524_Vitr_000006-006_196v | | jpg | 5 MB | -a--- |
| 3544 | 1096525_Vitr_000006-006_197r | | jpg | 5 MB | -a--- |
| 3545 | 1096526_Vitr_000006-006_198r | | jpg | 4 MB | -a--- |
| 3546 | 1096519 | | xml | 0 MB | -a--- |
| 3547 | u:\ENRICH\169842\ | 72 | | 14 MB | |
| 3548 | 169849_Mss_009990_0010 | | jpg | 0 MB | -a--- |
| 3549 | 169853_Mss_009990_0012 | | jpg | 0 MB | -a--- |
| 3550 | 169858_Mss_009990_0013 | | jpg | 0 MB | -a--- |
| 3551 | 169862_Mss_009990_0014 | | jpg | 0 MB | -a--- |

and an appropriate current result of such a process is:

```
...
  <page>
    <pgFoliation>195r</pgFoliation>
    <pgImage                                        id="ID1096521"
href="http://www2.bne.es:81/Enrich/ENRICH/1096519/1096521_Vitr
_000006-006_195r.jpg" quality="Normal"/>
  </page>
  <page>
    <pgFoliation>195v</pgFoliation>
    <pgImage                                        id="ID1096522"
href="http://www2.bne.es:81/Enrich/ENRICH/1096519/1096522_Vitr
_000006-006_195v.jpg" quality="Normal"/>
  </page>
  <page>
    <pgFoliation>196r</pgFoliation>
    <pgImage                                        id="ID1096523"
href="http://www2.bne.es:81/Enrich/ENRICH/1096519/1096523_Vitr
_000006-006_196r.jpg" quality="Normal"/>
  </page>
  <page>
    <pgFoliation>196v</pgFoliation>
    <pgImage                                        id="ID1096524"
href="http://www2.bne.es:81/Enrich/ENRICH/1096519/1096524_Vitr
_000006-006_196v.jpg" quality="Normal"/>
  </page>
```

. . .

Data elements are available directly at http://www2.bne.es:81/Enrich/ENRICH/ via a HTTP.

#### 3.3.3.2 Incunabulas, old prints

The available documents are searchable PDF documents. The PDF format is (despite its other advantages) characterised by lower interoperability which makes it unusable within ENRICH. The dependency on specialised browsing tools (such as Adobe Acrobat reader and its variations) makes it hardly possible to achieve the main ENRICH goal of aggregation under one single homogenous user interface.

Therefore the PDF files will be extracted, and the resulting JPEGs will be placed on the webserver of BNE (the estimated size of the full collection should be aproximately3 GB only). During the extraction the necessary structural metadata will be produced. Then the documents will be processed in a way similar to the Manuscript, maps collections.

There is a possibility to produce TEI P5 based fulltext transcription during the extraction as there is the textual layer in the PDF as a result of OCR. The final decision depends on further discussion about the OCR text usability.

### 3.3.4 Current results

#### 3.3.4.1 Manuscripts, maps

The structural metadata have been created and processed, the initial bibl. conversion routine has been placed into position (the full conversion will wait for results of cooperation within T3.1) and the collections have been processed making the documents available in the clone at enrichdata.manuscriptorium.com.

#### 3.3.4.2 Incunabulas, old prints

Analysis of PDF content, sample selections for the first tests of extraction (already passed successfully).

### 3.3.5 Remaining steps

#### 3.3.5.1 Manuscripts, maps

Prepare MARC 21 conversions and set up them in the connector.

#### 3.3.5.2 Incunabulas, old prints

Confirm the proposed process and details of extraction, generate and process structural metadata, prepare MARC 21 conversions and set up them in the connector.

## *3.4 BUTE*

### 3.4.1 Metadata standards

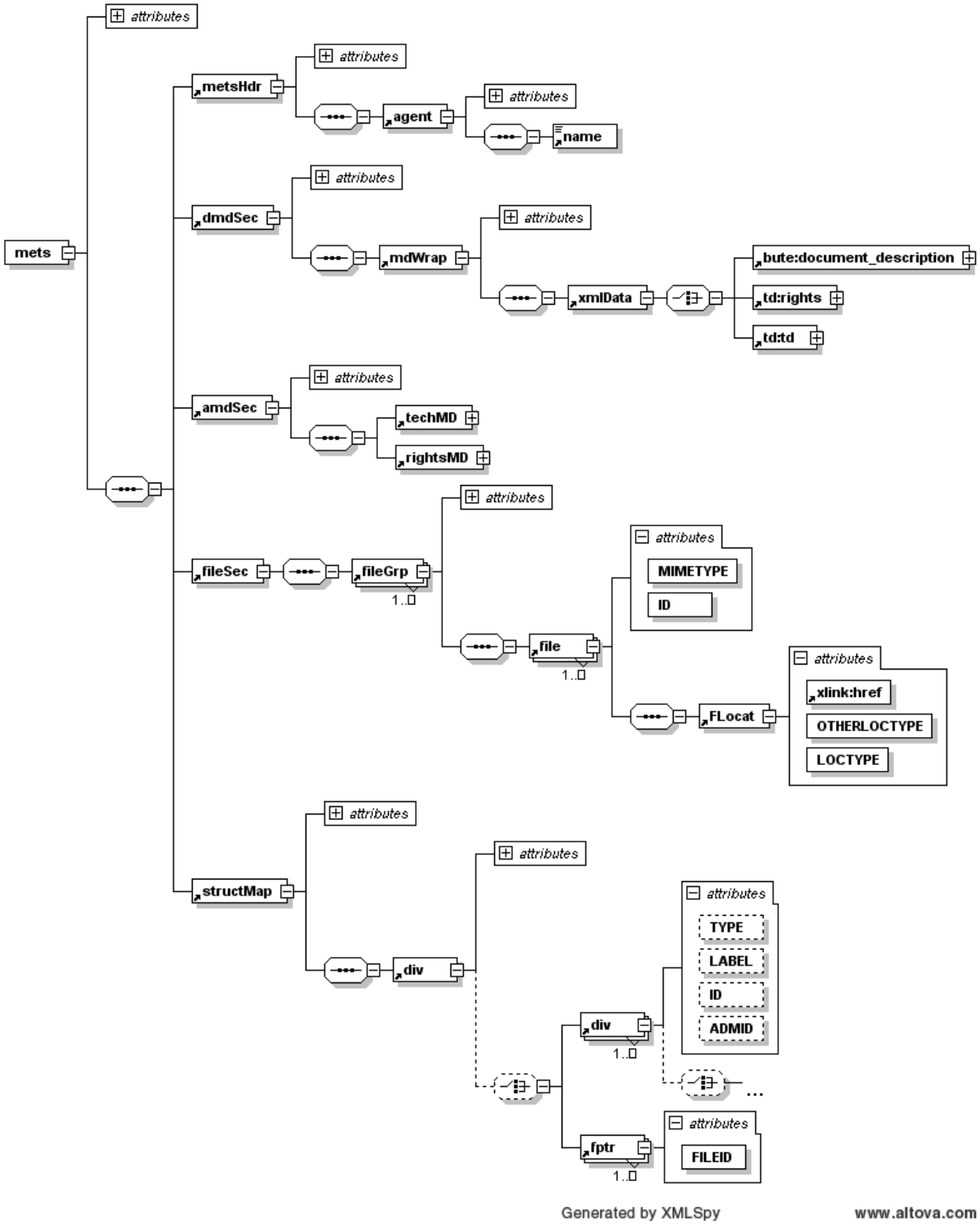
| **All documents** | |
|---|---|
| **Descriptive metadata** | **local format** in the form of XML wrapped in the METS format |

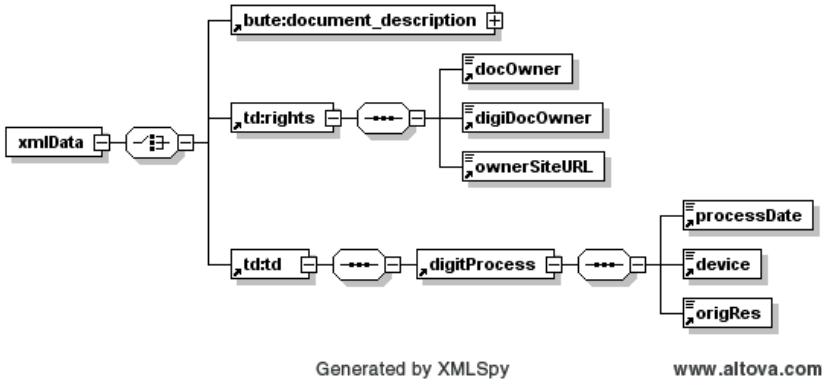
| | |
|---|---|
| | The local format structure was analysed and it will be converted to TEI P5 using a dedicated connector. |
| **Structural metadata** | **METS structural maps**<br>According to the common METS format practice. |

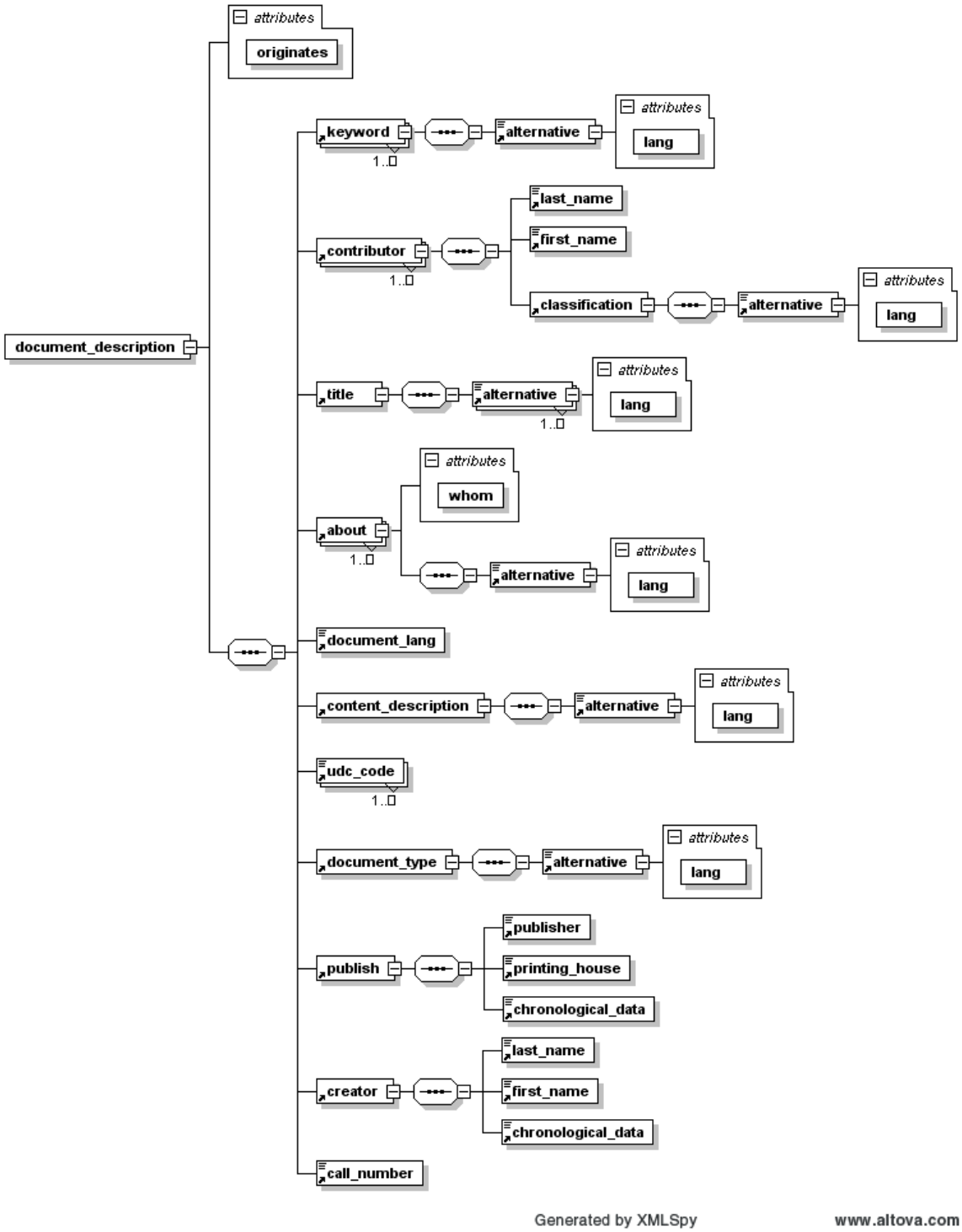
### 3.4.2 Method of collaboration

Off-line metadata passing (via FTP or other method) with subsequent processing and conversions within a BUTE dedicated connectors.

The following diagrams reflect the structure of the METS and the local descriptive metadata used.


attributes
mets
metsHdr
agent
name
dmdSec
mdWrap
xmlData
bute:document_description
td:rights
td:td
amdSec
techMD
rightsMD
fileSec
fileGrp
1..∞
file
MIMETYPE
ID
FLocat
xlink:href
OTHERLOCTYPE
LOCTYPE
structMap
div
TYPE
LABEL
ID
ADMID
fptr
FILEID
Generated by XMLSpy
www.altova.com


bute:document_description
xmlData
td:rights
docOwner
digiDocOwner
ownerSiteURL
td:td
digitProcess
processDate
device
origRes
Generated by XMLSpy
www.altova.com


attributes
originates
document_description
keyword
alternative
lang
1..∞
contributor
last_name
first_name
classification
title
about
whom
document_lang
content_description
udc_code
document_type
publish
publisher
printing_house
chronological_data
creator
call_number
Generated by XMLSpy
www.altova.com

The mets:fileGrp groups appropriate level quality files which are linked by mets:file/mets:FLocat/@xlink:href. The mets:structMap contains the structural map of the complex digital document including the map of the digitised copy. The applicable pagination is also included and processed.

The appropriate files in the file section are referred by mets:div[@TYPE=”DigitalCopy”]/mets:div[TYPE="Page"]/mets:fptr/@FILEID value which refers to corresponding value in mets:file/@ID.

Data elements are available directly via a HTTP.

### 3.4.3 Current results

First testing documents have been converted and made available for review (initially within M-Can application only).

### 3.4.4 Remaining steps

Prepare first testing batch processing, prepare MARC 21 conversions and set up them in the connector.

## *3.5 DSP*

### 3.5.1 Metadata standards

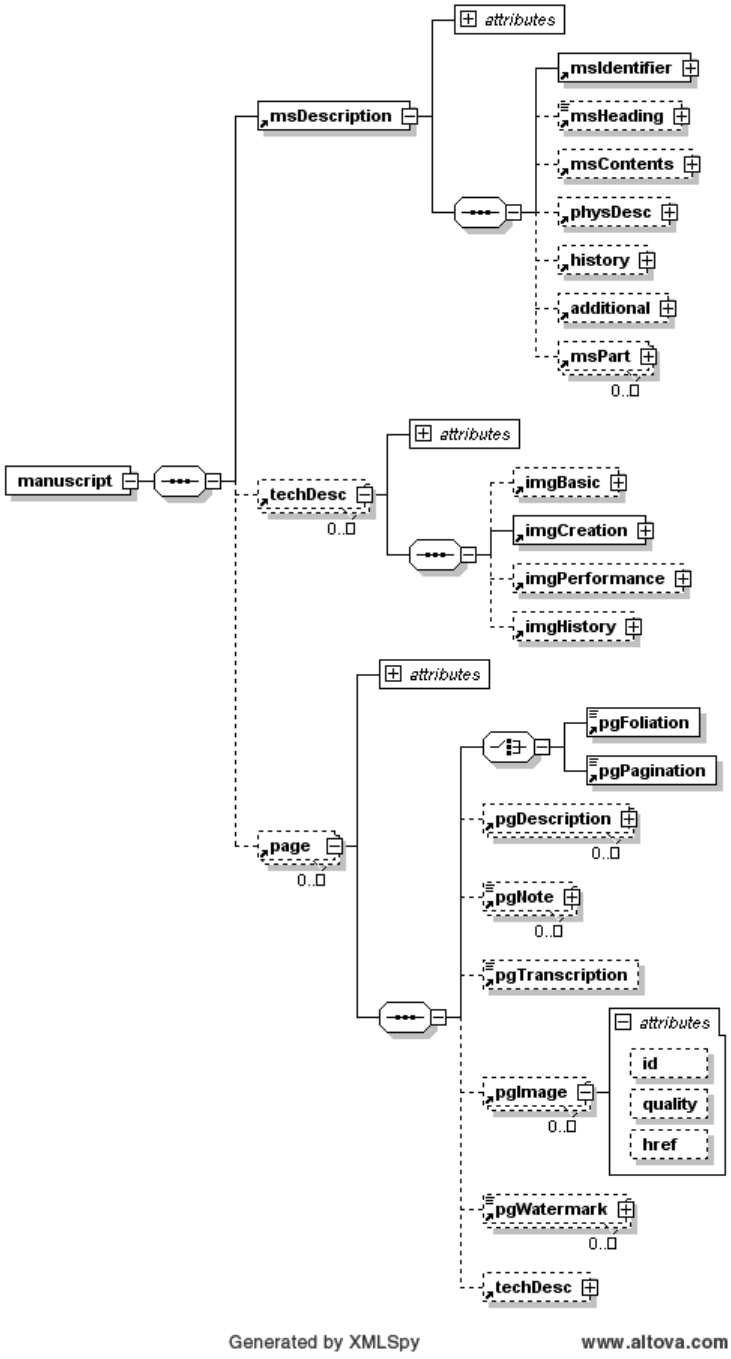
| Manuscripts, incunabulas | |
|---|---|
| **Descriptive metadata** | **MASTER+**<br>using M-Tool |
| **Structural metadata** | **MASTER+**<br>using M-Tool |

| Charters | |
|---|---|
| **Descriptive metadata** | **CEI**<br>a TEI based XML format; structure was analysed and it will be converted to TEI P5 using a dedicated connector. |
| **Structural metadata** | **CEI** |
| **Additional metadata related notes** | Some of the documents contain full texts. |

### 3.5.2 Method of collaboration

#### 3.5.2.1 Manuscripts, incunabulas

The following diagram shows the preview of structure of the MASTER+. M-Tool enables to produce the msDescription and page elements.

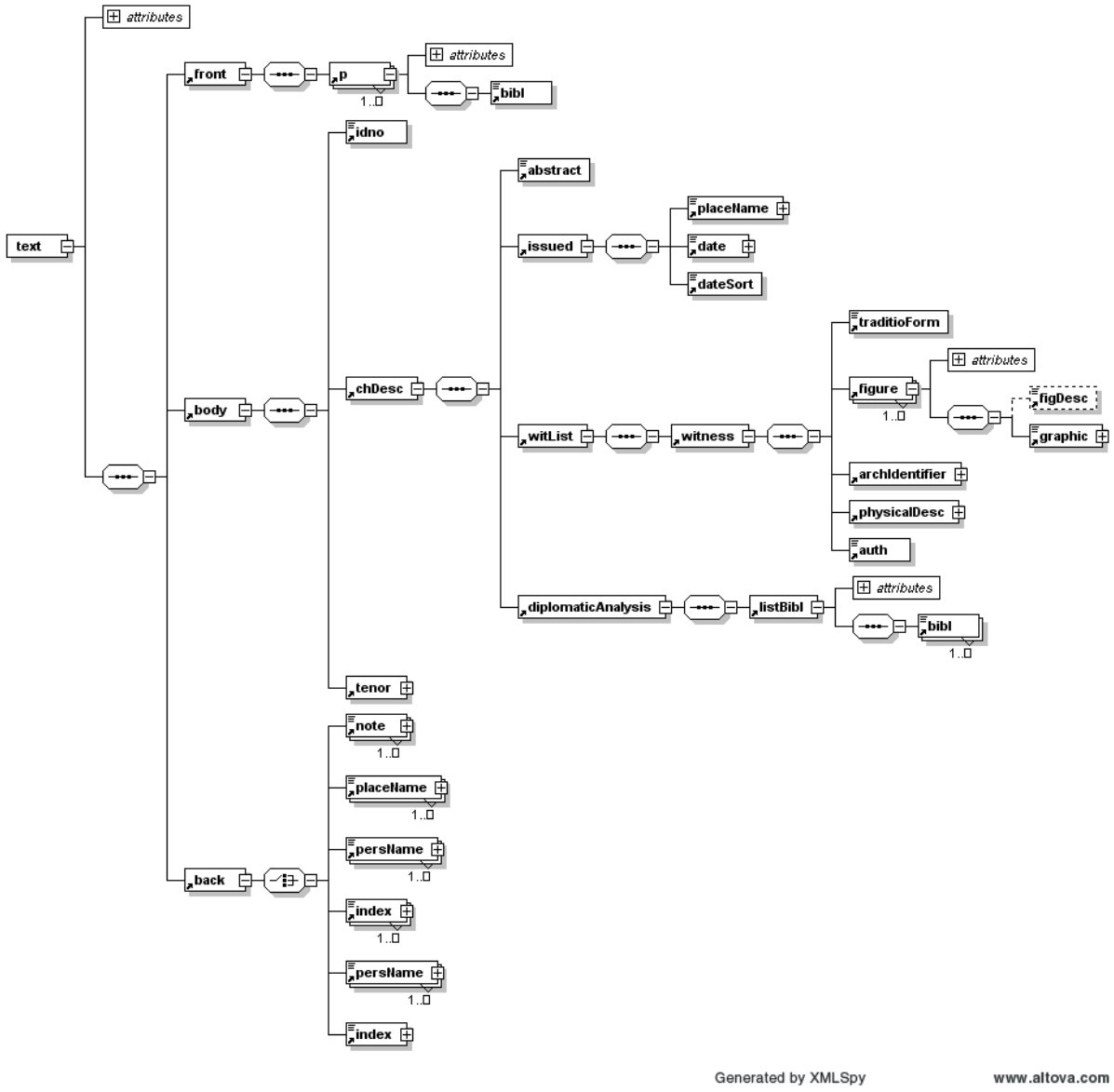



DSP will use existing and the newly created M-Tool and M-Can applications.

The multiple `pgImage` elements are allowed inside page element to enable capturing multiple image quality levels. The `pgImage/@href` contains the URL of the particular image.

### 3.5.2.2 Charters

The CEI structure as prepared based on the samples of maximally detailed records is as follows:




The images are linked directly and the information is located in `witList/witness/figure/graphic/@url`.

## 3.5.3 Current results

### 3.5.3.1 Manuscript, incunabulas

Routine production using M-Tool and M-Can applications. Some additional fields will be added in the new M-Tool forms in order to enable update of the records already created with additional information.

Current results are available in the clone at enrichdata.manuscriptorium.com.

### 3.5.3.2 Charters

The metadata content is being revised and updated by the DSP, no practical results in the clone are available yet.

### 3.5.4 Remaining steps

#### 3.5.4.1 Manuscripts, incunabula

Prepare the new M-tool and update already existing records.

#### 3.5.4.2 Charters

Analyse and process an export of metadata. Prepare a connector, set up an OAI harvesting...

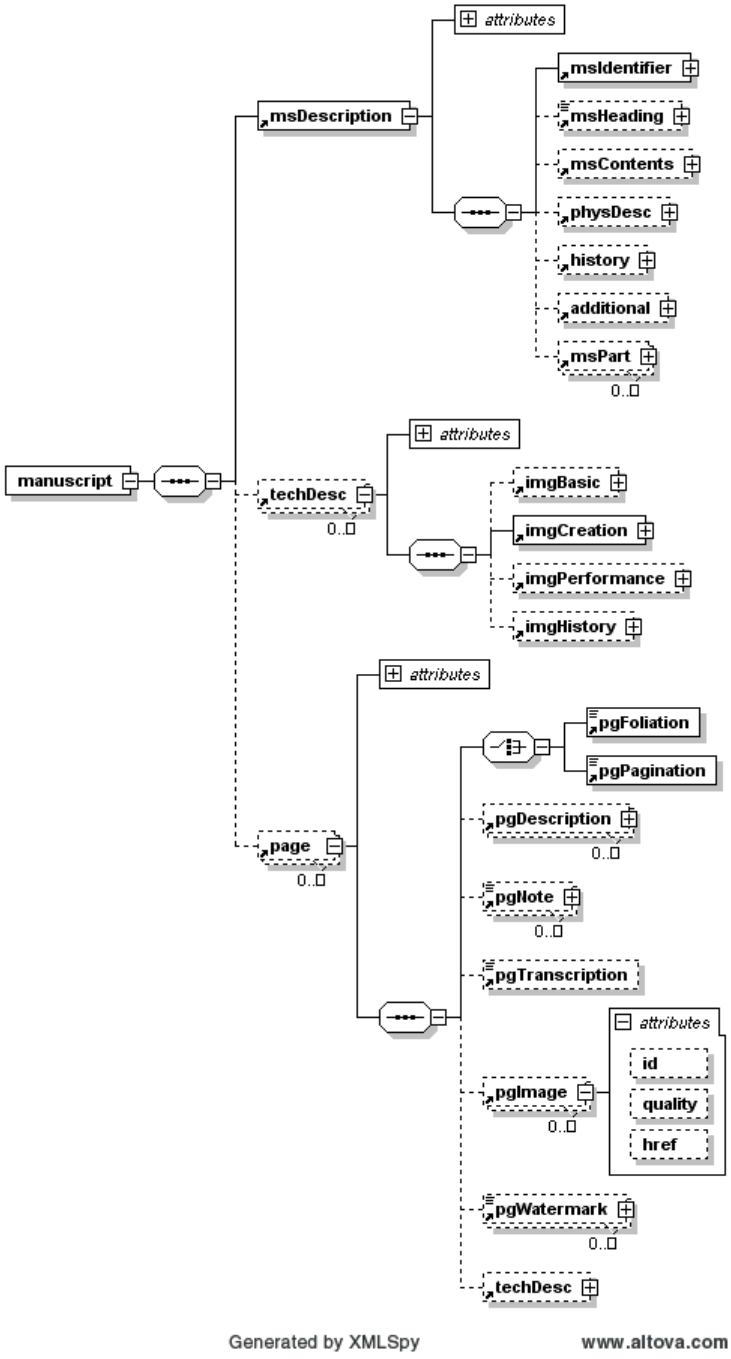
## *3.6 IMI*

### 3.6.1 Metadata standards

| Manuscripts | |
|---|---|
| **Descriptive metadata** | **MASTER+**<br>using M-Tool |
| **Structural metadata** | **MASTER+**<br>using M-Tool |

### 3.6.2 Method of collaboration

The following diagram shows the preview of structure of the MASTER+. M-Tool enables to produce the `msDescription` and page `elements`.




DSP will use existing and the newly created M-Tool and M-Can applications.

The multiple `pgImage` elements are allowed inside page element to enable capturing multiple image quality levels. The `pgImage/@href` contains the URL of the particular image.

### 3.6.3 Current results

Routine production using M-Tool and M-Can applications. Some additional fields will be added in the new M-Tool forms in order to enable update of the records already created with additional information.

More important issue regards to the possibility to create structural metadata of partially digitised documents. Manual metadata editing is necessary until now.

Current results are available in the clone at enrichdata.manuscriptorium.com.

### 3.6.4 Remaining steps

Prepare the new M-tool and update already existing records.

## *3.7 KU, NULI, SAM*

(partners cooperating on a common repository of manuscripts)

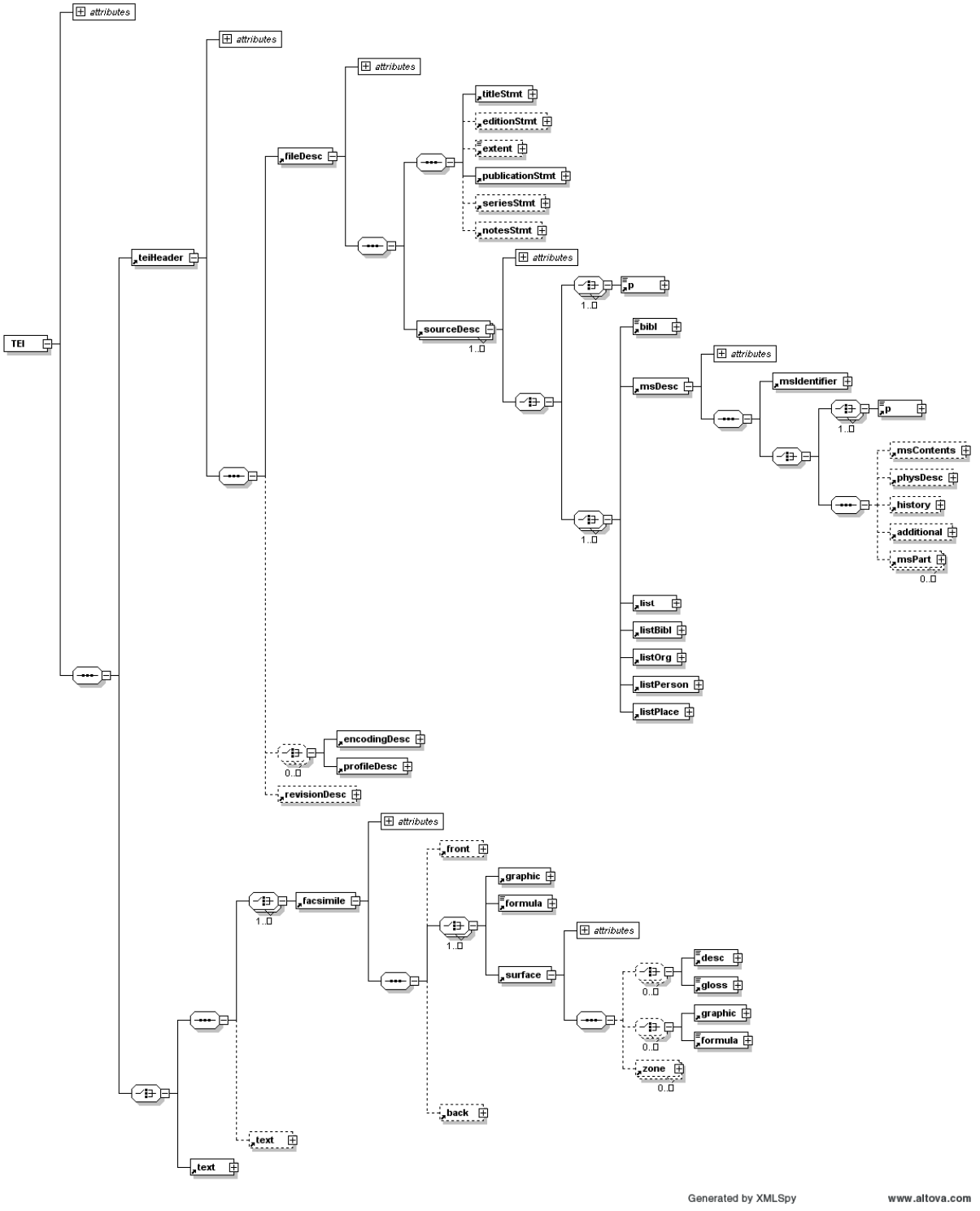
### 3.7.1 Metadata standards

| Manuscripts | |
|---|---|
| **Descriptive metadata** | **TEI P5** |
| **Structural metadata** | **TEI P5** |

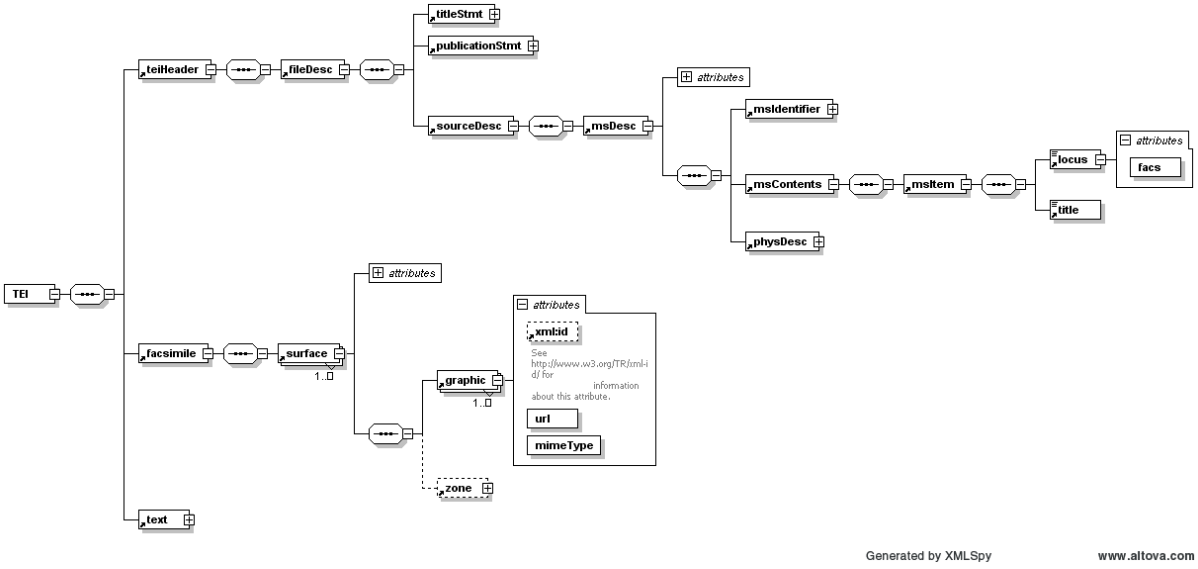
### 3.7.2 Method of collaboration

Manuscriptorium will migrate from MASTER to the TEI P5. These partners providing metadata directly in the TEI P5 will be connected automatically.

The XSD of an actual ENRICH TEI P5 scheme as a current result of T3.1 is as follows:




Structural metadata capturing will be possible by using the graphic element. The images are served directly via HTTP, the URL can be found in the `//graphic/@url`. Linking between physical structure and the logical structure (including applicable foliation for referring users to the document folios) will be available using reference from the `/locus/@facs` to the appropriate `//graphic/@xml:id`.




### 3.7.3 Remaining steps

We can continue after definite TEI P5 records are prepared.

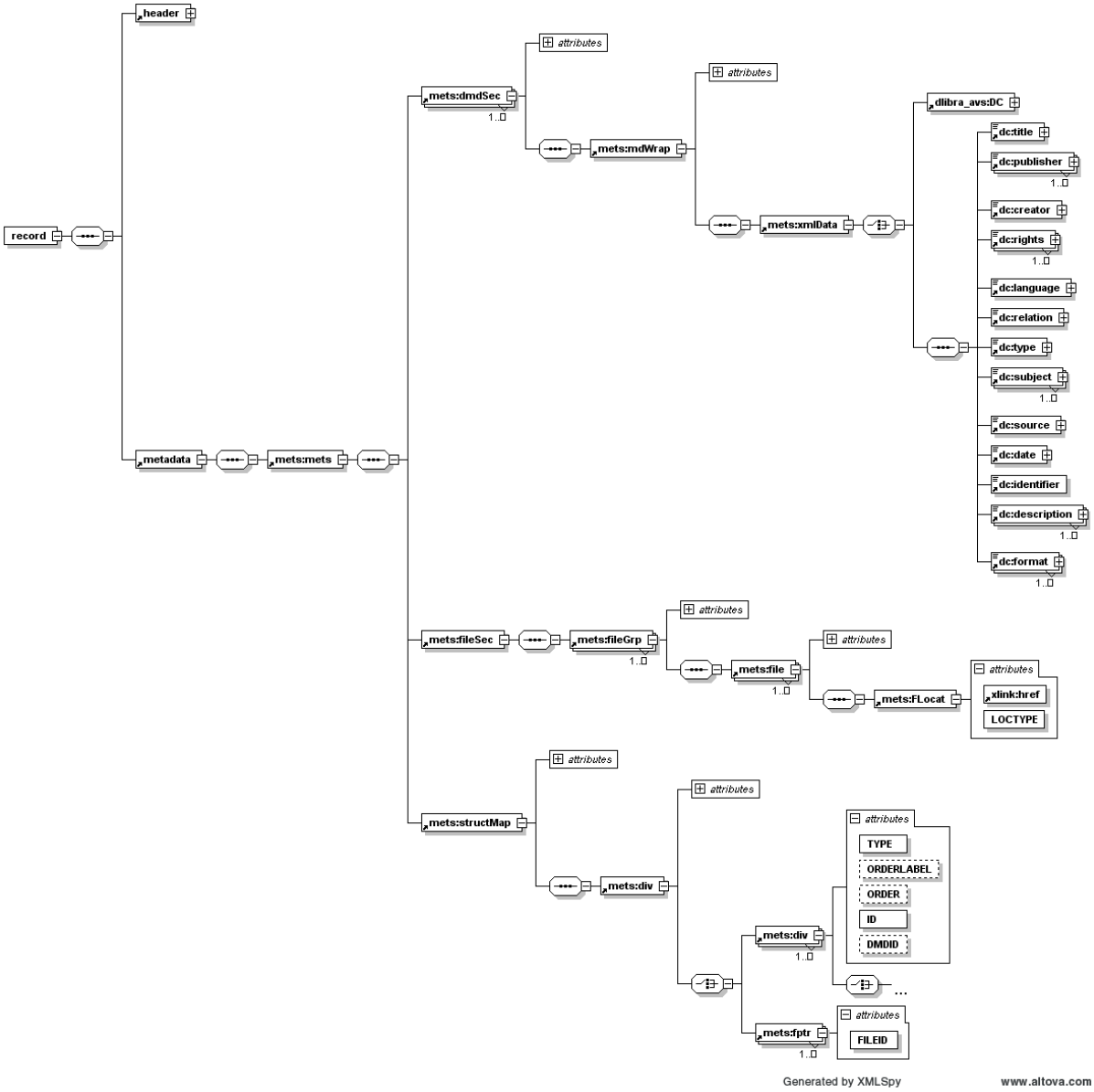
## *3.8 ULW*

### 3.8.1 Metadata standards

| Manuscripts | |
|---|---|
| **Descriptive metadata** | **DC**<br>The content of the primary local metadata format is very similar, near the sane, as the DC. Therefore the DC is processed. |
| **Structural metadata** | **METS**<br>with DC wrapped; special mets_exp profile exist in the OAI-PMH interface as the JPEG files are generated ad-hoc for the ENRICH project needs |

### 3.8.2 Method of collaboration

The following diagram shows the preview of structure of the METS sample.




The `mets:fileGrp` groups appropriate level quality files which are linked by `mets:file[@MIMETYPE="image/jpeg"]/mets:FLocat/@xlink:href`. The `mets:structMap[@TYPE="PHYSICAL"]` contains the page sequence of the digitised copy. The applicable pagination is not included.

The `mets:fileGrp[@USE="original"]` contains the original DjVu files reference.

The particular files in the file section are referred by `mets:div[@TYPE="DigitalCopy"]/mets:div[TYPE="Page"]/mets:fptr/@FILEID` value which refers to corresponding value in `mets:file/@ID`.

Data elements are available directly via a HTTP.

### 3.8.3 Current results

A special solution prepared by the PSNC enables real-time generation of the necessary JPEG files from the original DjVu. Necessary metadata are provided using a dedicated mets_exp AI-PMH profile. First sample documents were prepared and tested in the Manuscriptorium viewing interface (initially M-Can application used).

### 3.8.4 Remaining steps

After the software platform of the ULW repository will be updated by the new DjVu-JPEG functionality the OAI harvesting will be initiated, convertor will be set up and conversion routines designed.

## *3.9 UZK*

### 3.9.1 Metadata standards

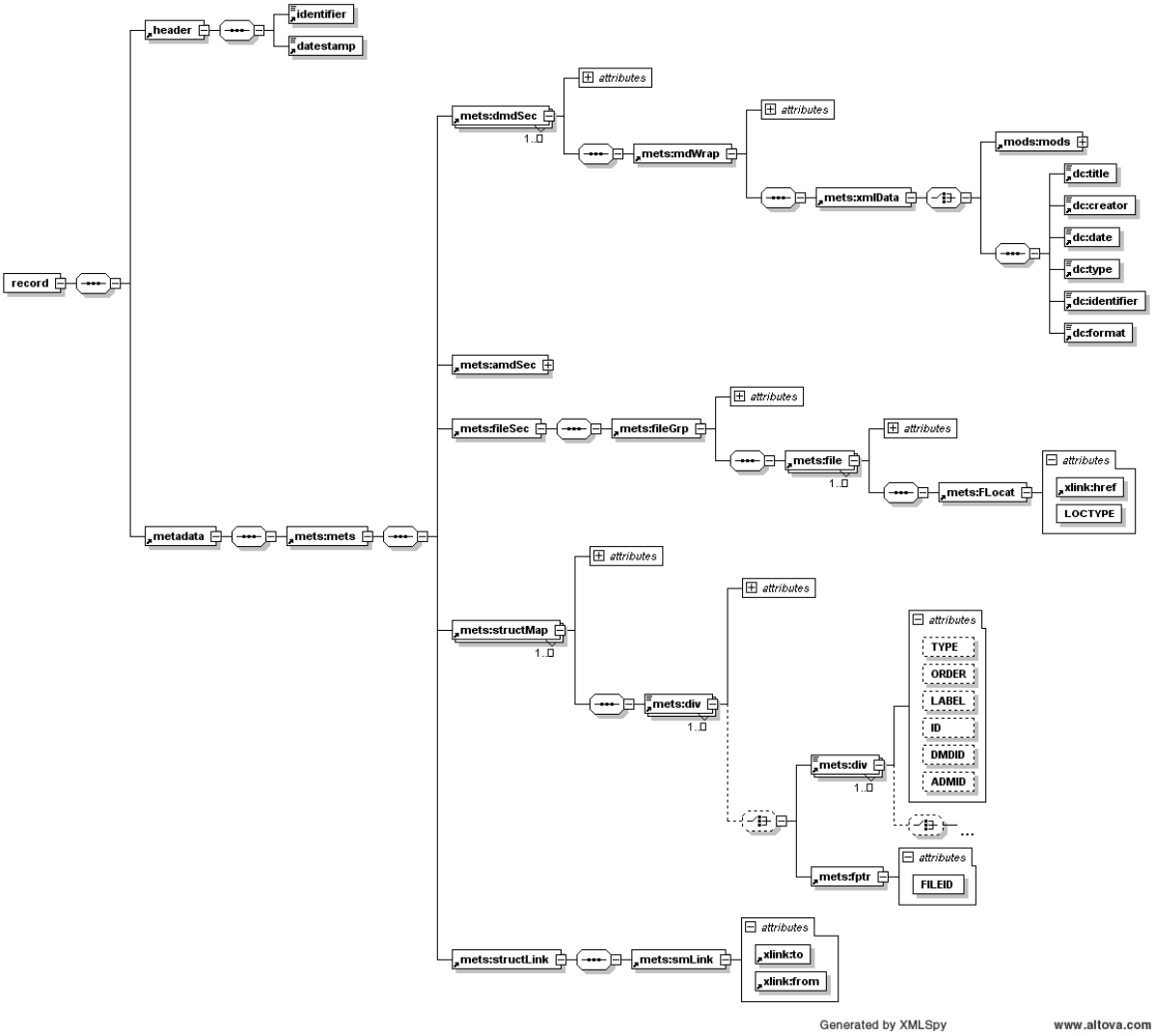
| **Incunabulas** | |
|---|---|
| **Descriptive metadata** | **ISTC**<br>The primary format will possibly be replaced in the processing by DC records in case it will be confirmed it contains the same ENRICH-useful information content as the primary metadata. |
| **Structural metadata** | **METS** structural maps |

| **Manuscripts** | |
|---|---|
| **Descriptive metadata** | **TEI P5**<br>UZK will migrate to the TEI P5 during the ENRICH project. |
| **Structural metadata** | **TEI P5** |

### 3.9.2 Method of collaboration

#### 3.9.2.1 Incunabulas

The following diagram shows the structure of the available metadata:




Logical map of the document uses the `mets:structMap[@TYPE="LOGICAL"]` and is divided into multiple units (according to chapters and other logical units). The `mets:fptr/@FILEID` is used to link the appropriate page. Applicable pagination is included in the `mets:div[@TYPE="page"]/@LABEL`.

The `mets:fileGrp` groups appropriate level quality files which are linked by `mets:file/mets:FLocat/@xlink:href`.

### 3.9.2.2 Manuscripts

UZK will migrate to TEI P5. As soon as Manuscriptorium will migrate from MASTER to the TEI P5 the partner will be directly compatible.

The XSD of an actual ENRICH TEI P5 scheme and appropriate description is already described in the chapter KU, NULI, SAM / Method of collaboration.

### 3.9.3 Current results

#### 3.9.3.1 Incunabulas

First OAI harvest has been performed, results are available in the clone at enrichdata.manuscriptorium.com.

### 3.9.4 Remaining steps

#### 3.9.4.1 Icunabula

Decide format for descriptive metadata (DC x ISTC), correct errors found during harvest, prepare connector and conversion routines.

#### 3.9.4.2 Manuscripts

We can continue after definite TEI P5 records are prepared.

## *3.10 VUL*

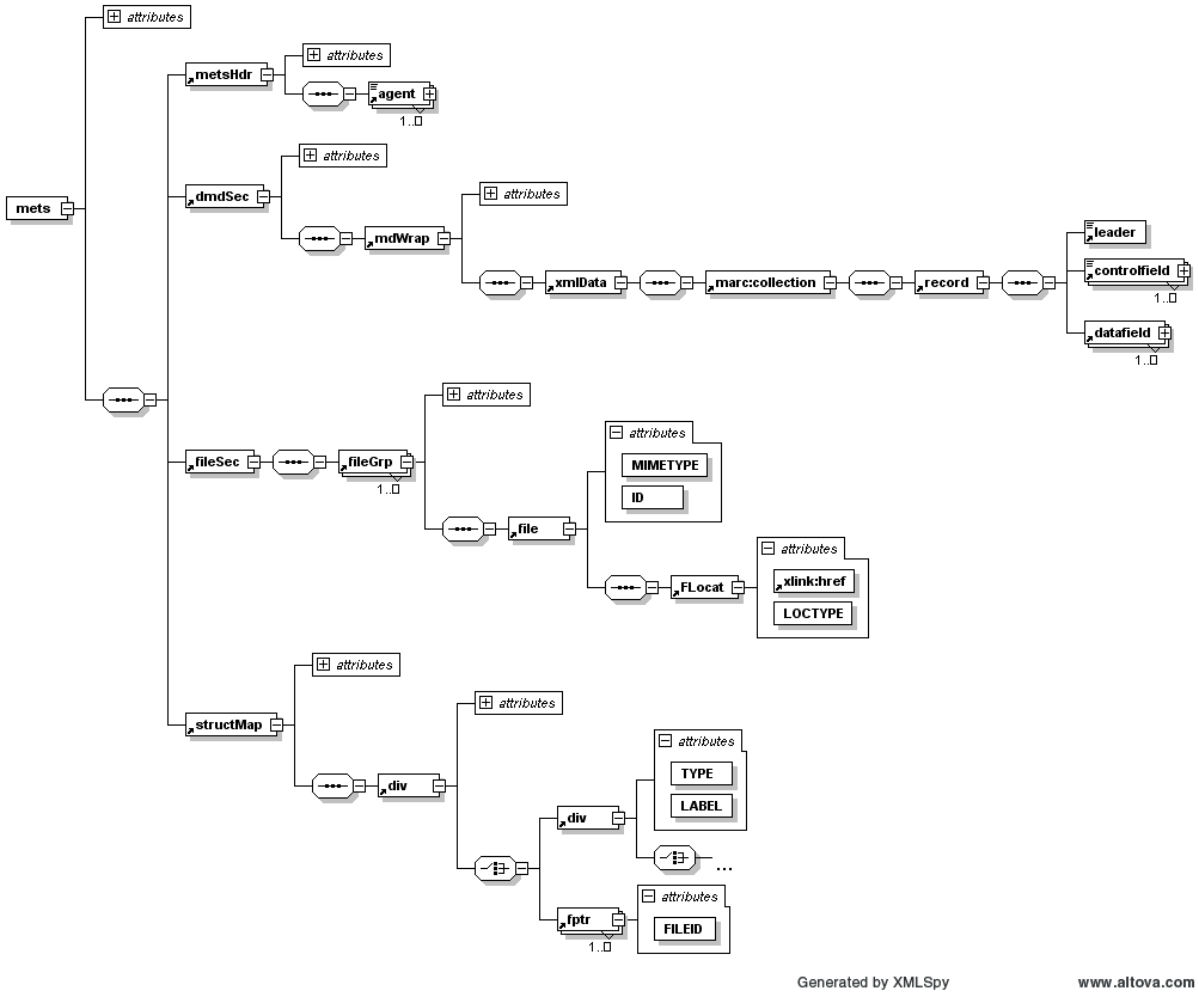
### 3.10.1 . Metadata standards

| **All documents** | |
|---|---|
| **Descriptive metadata** | **MARC 21** in the form of XML<br>The MARC 21 format conforms regular cataloguing guides and will be converted to TEI P5 using a dedicated connector.<br>Possible derivations from the cataloguing manual guides and content of possible national specifics fields will be discussed with the responsible BNE cataloguers. |
| **Structural metadata** | **METS structural maps**<br>According to the common METS format practice. |

### 3.10.2 Method of collaboration

Off-line metadata passing (via FTP or other method) with subsequent processing and conversions within a VUL dedicated connectors.

The following diagrams reflect the structure of the METS and the local descriptive metadata used:




The `mets:fileGrp` groups appropriate level quality files which are linked by `mets:file/mets:FLocat/@xlink:href`. The `mets:structMap` contains the structural map of the document.

### 3.10.3 Current results

First sample analysed and agreed timetable for preparation of real records.

### 3.10.4 Remaining steps

Prepare first testing batch processing, prepare MARC 21 conversions and set up them in the connector.

# 4 Conclusion

As stated above the information provided here is an important input for WP5 tasks. The first samples were already processed by AIP based on the information presented in this deliverable. These first results of WP5 can be seen at http://enrichdata.manuscriptorium.com.

**Attention!** This is a working and testing environment – a special clone of the real Manuscriptorium developed only to ENRICH project. It may include incomplete records, records of testing documents and possibly errors can be found which are continuously corrected. The environment can be even temporarily unavailable due testing and tuning reasons. Therefore this address should be used for the internal needs of the project, for possible monitoring of current focus and/or overall work progress. **The adress therefore is not published to the wide public!** All final converted documents and all tested and approved tools will be implemented in the real Manuscriotorium.