ECP-2008-DILI-518001

BHL-Europe

Plan for managing interoperability issues, data harmonisation and the integration of the content into BHL-Europe, EUROPEANA and the BHL

Deliverable number	*D3.3*
Dissemination level	*Public*
Delivery date	*24 November 2009*
Status	*Final*
Author(s)	*Alexander Herzog, Adrian Smales*

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.

Table of contents

1 Document History

This chapter describes the document's creation events and contributors.

1.1 Contributors

This document is based on the meetings of the technical team in Leiden, Graz and Berlin with the following members contributing to this document.

Person	Partner
Alexander Herzog	AIT
Walter Koch	AIT
Gerda Koch	AIT
Adrian Smales	NHM

1.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
20 Nov 2009	Alexander Herzog	0.1	First draft for internal review
24 Nov 2009	Adrian Smales	1.0	Final version after internal review

1.3 Reviewers and Approvals

This document requires the following approvals.

Name	Position	Date	Version
Adrian Smales	Deliverable responsible	24 Nov 2009	1.0
Henning Scholz	Project coordinator	24 Nov 2009	1.0

1.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
Core development team	20 November 2009	0.1
BHL-Europe consortium	27 November 2009	1.0

2 Purpose and document structure

This document depicts the management of interoperability issues, the vision for data harmonization and the basic requirements for integrating content into BHL-Europe, the EUROPEANA Portal (www.europeana.eu) and the BHL portal (www.biodiversitylibrary.org).

3 METS Profile as base for interoperability issues

The OAIS reference model differentiates between three kinds of information objects. The SIP, Submission Information Package, is being sent in by the data producers, the AIP, Archive Information Package, is preserved in the archive, and the DIP, Dissemination Information Package, is provided for the consumers of the archive.

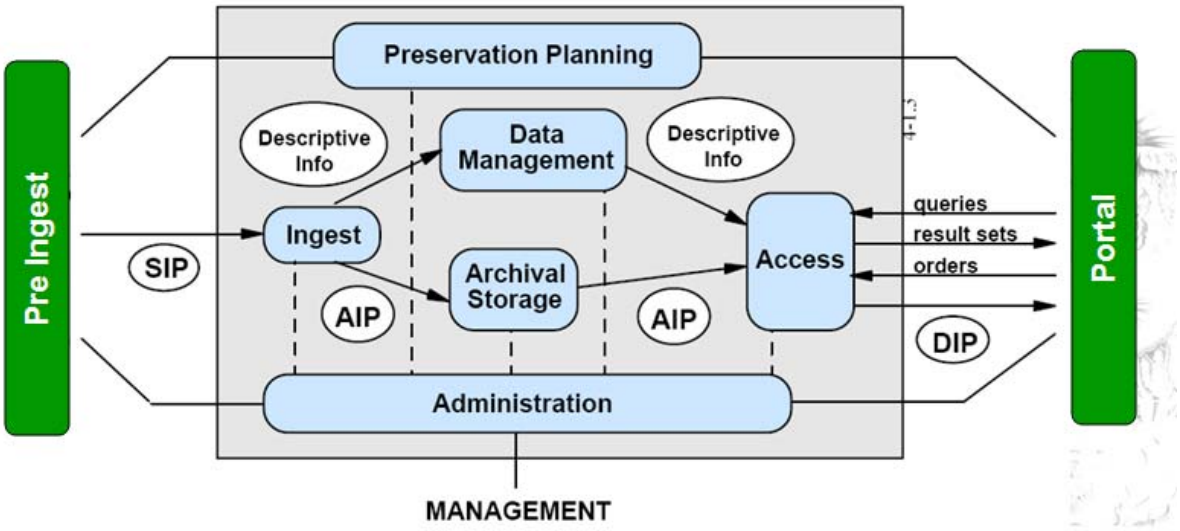

Figure 3-1: The OAIS components relevant for BHL-E

3.1 General METS Profile introduction

In order to build our work upon best practice and standards it was decided to use the METS standard (Metadata Encoding & Transmission Standard, http://www.loc.gov/standards/mets/) for creating the BHL-Europe Information Packages. METS promotes interoperability of descriptive, administrative and technical metadata. It uses XML, which ensures that it is interchangeable. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.

METS attempts to provide an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users). It contains references to scanned digital data, OCR data, Structural information for displaying a table of content and related items. The structural information is also used to created composed items.

A METS document consists of seven major sections:

1. METS Header (metadata describing the METS document itself)
2. Descriptive Metadata (may contain external, like MARC, or internally embedded descriptive metadata, or both)
3. Administrative Metadata (provides information regarding how the files were created and stored, intellectual property rights, etc.; may be either external to the METS document, or encoded internally.)
4. File Section (lists all files containing content which comprise the electronic versions of the digital object.)
5. Structural Map (outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element.)
6. Structural Links (allows METS creators to record the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map.)
7. Behavior (this section can be used to associate executable behaviors with content in the METS object.)

3.2 Description of the proposed BHL-Europe METS profile for the SIP

This section represents the profile documentation as it will be sent to the METS consortium for validation.

3.2.1 URI

[not registered yet] http://www.bhleurope.eu/ingest/sip/20091005.xml.

3.2.2 Title

BHL-Europe METS Document Profile for Submission Information Package (SIP).

3.2.3 Abstract

BHL-Europe is building an Open Archival Informaion System (OAIS) compliant web application. This implementation has the need for a Submission Information Package (SIP) for the ingest module. METS profiles can contain all information needed for the Use Cases described by the consortium.

Due to various pre-ingest tool providers and the need for a standard, this profile is intended to provide all information needed by the content providers to get the most features out of BHL-Europe's web application for their items.

This profile describes the SIP and may also be used as Dissemination Information Packages (DIPs) or Archival Information Packages (AIPs).

3.2.4 Creation Date

2009-10-05 --:-- CET

3.2.5 Contact Information

- Institution: Biodiversity Heritage Library for Europe
- Address:
- Phone: 0
- Mail:

3.2.6 Related Profile

No related profiles.

3.2.7 Extension Schema

The following schemas will be used for BHL-Europe SIP.

MARCXML: The MARC 21 XML Schema

- URI: http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd
- Context: Descriptive Metadata Set ./*[@LABEL='marcxml']
- Note: Prepared by Corey Keith, May 21, 2002 - Version 1.0 - Initial Release

DjVuXML-s

- URI: http://djvu.cvs.sourceforge.net/viewvc/*checkout*/djvu/djvulibre-3.5/share/djvu/pubtext/DjVuXML-s.dtd
- Context: Descriptive Metadata Set ./*[@LABEL='djvuxml']
- Note: Part of the djvutoxml tool (http://sourceforge.net/projects/djvu/). It is a simplified HTML DTD.

3.2.8 Rules of Description

MARCXML

The following fields are required for BHL-Items as these describe the BHL Deduplication Subset (BDS)[2].

- Marc Leader Pos7 - Category: Monography or serial component

- 245ac - Title: The title of the BHL-Item
- 100, 700 - Creator: the creator of this BHL-Item
- 260$b - Publisher: The Person or organisation responsible for publishing the original object
- 250ab - Edition of monographs: The edition of the book beeing scanned
- Date Created: Date of the creation of the digital object (part of METS:ROOT element)
- Date Last Modified: Date digital object was öast updated (part of METS:ROOT element)

DjVuXML

The OCR output of DjVuXML will be used for highlighting words in the bookreader and as source for fulltext search. Therefore, the DjVuXML has to provide word coordinates.

3.2.9 Controlled Vocabularies

- the consortium did not decide yet which controlled vocabularies will be used in the future.
- nevertheless, there will be several controlled vocabularies used within the MARC records, eg. MARC21 relator codes, ISO 639-2.

3.2.10 Structural Requirements

mets Root Element

1. Each SIP describes one and only one Record that will show up in BHL-Europe (=BHL-Item). In case you need to describe multiple items for series, create one SIP per sub item and link them within the logical structure map to a compound item.
2. Each SIP must have a mets:objid with an URI like oais:<INSTITUTION>:<INTERNALIDENTIFIER>. These identifiers will be used for deduplication and will be replaced by a BHL internal URI during ingest. The original mets:objid will be moved to the mets:metsHdr.

mets Header

-

[2] still under discussion

mets Descriptive Meta Data

1. Each SIP must have one mets:dmdSec/(mets:dmWrap|mets:dmRef)[@type="MARC" and @label="marcxml"] node. These information will be used for indexing the metadata. If multiple nodes of this type are present, then the first or the one with the status current will be used.

```
<mets:dmdSec ID="metadata">
  <mets:mdWrap MDTYPE="MARC">
    <mets:xmlData>
      <marc:record>
        <marc:leader>00854nam a2200265u  4500</marc:leader>
        <marc:controlfield tag="001">000001</marc:controlfield>
        <marc:controlfield tag="005">20090716064900.0</marc:controlfield>
        <marc:controlfield tag="008">090716s1901 gw a 000 0 ger </marc:controlfield>
        <marc:datafield tag="040" ind1=" " ind2=" ">
```

Figure 3-2 Excerpt from BHL METS descriptive metadata section MARC

2. For fulltext search and GnuBook highlighting a mets:dmdSec/(mets:dmWrap|mets:dmRef)[@type="OTHER" and @use="djvuxml"] must be available. If multiple nodes of this type are present, then the first or the one with the status current will be used.

mets Administrative Meta Data

1. the consortium did not decided yet on rights and image preservation metadata

mets File Section

1. For GnuBook support a mets:fileSec/mets:fileGrp with binary or reference data of images (tiff, jpg, jp2) must be available. Depending on the use attribute, master, reference and thumbnail mets:fileGrp will be generated.
2. For preservation a mets:fileSec/mets:fileGrp[@use="master"] with binary or reference data of images (tiff) must be available.

```
<mets:file ID="tmb0290" SEQ="0290">
  <mets:FLocat LOCTYPE="URL"
    xlink:href="http://dev.ait.co.at/textupl/Books/bilderatlasmw4/img/thumbs/bilderatlasmw4_0290.jpg" />
</mets:file>
<mets:file ID="tmb0291" SEQ="0291">
  <mets:FLocat LOCTYPE="URL"
    xlink:href="http://dev.ait.co.at/textupl/Books/bilderatlasmw4/img/thumbs/bilderatlasmw4_0291.jpg" />
</mets:file>
```

Figure 3-3 Excerpt from BHL METS file section

mets Structural Map

1. For page numbers and page names support a mets:structMap[@type="physical"] must be supplied with a root div containing all leafs. The containing leafs must be named by the orderlabel attribute and put in sequence with the order attribute starting at 0.
2. For logical item structure navigation support a mets:structMap[@type="logical"] must be supplied.
3. The mets:structMap types should use one of the following values: frontcover, halftitlepage, titlepage, imprint, dedication, inspiration, foreword, preface, toc, lot, lof, introduction, chapter, part, afterword, bibliography, references, appendix, glossary, index, colophon, promotion, backcover

```
<mets:structMap TYPE="logical">
  <mets:div TYPE="book" LABEL="Bilder Atlas des Pflanzenreichs">
    <mets:div TYPE="cover" ORDER="0" LABEL="Cover (front)">
      <mets:fptr FILEID="tif0001" />
    </mets:div>
    <mets:div TYPE="titlepage" ORDER="4" LABEL="Title Page">
      <mets:fptr FILEID="tif0005" />
    </mets:div>
    <mets:div TYPE="section" ORDER="6" LABEL="Vorwort zur dritten Auflage">
      <mets:fptr FILEID="tif0007" />
    </mets:div>
```

Figure 3-4 Excerpt from BHL METS structural map

mets Structural Link

-

mets Behaviour

1. [DRAFT] The behavioural elements are used to do automated image transformation
2. [DRAFT] The behavioural elements are used to control incremental updates only

3.2.11 Technical Requirements of Content, Behavior and Metadata Files

Images

For the mets:fileGrp the following file types are accepted:

- Bitonal images must be 300dpi-600dpi TIFF.
- Grayscale images must have 300dpi - 600dpi as uncompressed TIFF or lossless compressed images like LZW or JP2000) in 8bit color depth.

- Color images shall be between 300dpi and 400dpi. 24 bit color as uncompressed TIFF or lossless compressed JP2000 or LZW.
- NOTE: For BHL JP2000 files can be compressed by 15%.

3.2.12 Tools and Applications

- BHL-Europe Portal (under development)

3.2.13 Examples

- Examples are available for BHL and BHL-Europe members on the BHLwiki under https://bhl.wikispaces.com/file/view/sip.zip (includes the DjVU XML and a malformed SIP)

4 Harmonisation within BHL-Europe Pre-Ingest Module

In BHL-Europe the Pre-Ingest module and tools will package the various submitted content up into the BHL-Europe required METS profile and pass it on to the Ingest (as depicted in figure 1 of this chapter). Afterwards the content will be further processed and may remain in the submission format or change according to the needs of the various dissemination purposes.

- Step 1 – native schema selection
    - User indicates the native Metadata scheme (automated mapping process for standard metadata like: MARC21, Dublin Core, MODS)
    - If no standard Metadata Scheme is used the mapping tool provides field to field mapping
- Step 2 – structural augmentation
    - The user provides the Structural Information
    - Book structure
    - Types of pages
- Step 3 – data augmentation
    - The user indicates if:
    - Execute OCR reading
    - Connect to taxonomies (eg. uBio)
- Step 4 – automated validation
    - Automatic Check if SIP meets BHL-Europe standards
- Step 5 – manual validation
    - The user creates the SIP and previews the outcome (locally):
    - BHL-Europe format
    - EUROPEANA format
    - Dublin Core
    - and others
- Step 6 - ingest
    - The user submits the SIP to the Ingest functionality of BHL Europe
    - actual ingest is part of the OAIS system

5 Integration of content via BHL Europe OAIS Access Module

Integration with other projects and aggregation platforms like EUROPEANA will be handled at the BHL OAIS Access component. Access will provide a variety of different Dissemination Information Packages (DIP) as described below. Most of these DIPs will be created by XSL Transformation

5.1 OAI-PMH [standard]

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[3] is a protocol developed by the Open Archive Initiative. It is used to provide metadata description of items which will later be harvested through the simple XML over HTTP interface. The metadata will be presented in Dublin Core and additionally defined formats. For more formats see the DIP's standards. Search functionality provided by this interface reduces to collections, time frames and formats. This interface will be used by EUROPEANA[4] to get all records in their desired format.

5.2 SRU/SRW [standard]

Search/Retrieve via URL (SRU)[5] is a standard for libraries created within the Z39.50 International Next Generation (ZING). It is a modern equivalent to the Z39.50 protocol based on HTTP and XML. Therefore it is a candidate for cloud services. Search/Retrieve via Web (SRW) additionally provides a SOAP Interface. Both standards use CQL as query language.

5.3 CQL [standard]

The Contextual Query Language (CQL)[6] is a formal language for representing queries to information retrieval systems such as web indexes, bibliographic catalogues and museum collection information. The design objective is that queries be human readable and writable,

[3] http://www.openarchives.org/pmh/ [30.10.2009]

[4] http://www.europeana.eu/portal/ [30.10.2009]

[5] http://www.loc.gov/standards/sru/specs/search-retrieve.html [30.10.2009]

[6] http://www.loc.gov/standards/sru/specs/cql.html [30.10.2009]

and that the language be intuitive while maintaining the expressiveness of more complex languages.

Contextual Queries are built using an Index, a Relation and a Search Term. These elements can be used multiple times connected with Boolean operators and embraced by parentheses. Due to Relation Modifiers a wide range of queries can be built covering proximity searches as well as relevance searches.

The query commands are case insensitive. An example query might look like
dc.title any fish or (dc.creator any sanderson and dc.identifier = "id:1234567")

5.4 RSS/ATOM feeds [standard]

Web feeds like Really Simple Syndication (RSS)[7] and Atom Syndication Format (ATOM)[8] are data formats used to provide information about frequently updated content. For BHL-Europe, new items added or changed within the BHL-Europe network will be placed onto the feed. The feed will be customizable via GET parameters in CQL/OpenSearch or the BHL-Europe Data API and will aggregate the items suitable for the query.

5.5 OpenSearch [standard]

OpenSearch[9] is a collection of simple formats for the sharing of search results. The OpenSearch formats helps people to discover and use the search engine and to syndicate search results across the web. Search engines that crawl the surface of the web are picking up only a small fraction of the great content that is available within BHL-Europe. OpenSearch provides a deeper insight on the BHL-Europe content for crawlers like Google and Yahoo. OpenSearch was created by Amazon.

[7] http://en.wikipedia.org/wiki/RSS [30.10.2009]

[8] http://en.wikipedia.org/wiki/ATOM [30.10.2009]

[9] http://www.opensearch.org/Home [30.10.2009]

5.6 OpenUrl [standard]

The OpenURL standard (Z39.88)[10] is designed to support mediated linking from information resources, such as abstracting and indexing databases (sources), to library services (targets), such as academic journals, whether online or in printed or other formats. A "link resolver", or "link-server", parses the elements of an OpenURL and provides links to appropriate targets available through a library. A source is generally a bibliographic citation or bibliographic record used to generate an OpenURL. A target is a resource or service that helps satisfy user's information needs. Examples of targets include full-text repositories, online journals, online library catalogues and other Web resources and services. A source is typically a database that indexes various types of information resources often found in libraries, such as articles, books, patents, etc. Some examples of such databases include Web of Science, Scifinder, Modern Languages Association Bibliography and Google Scholar.

5.7 SKOS [standard]

Simple Knowledge Organisation Systems (SKOS) [11]is a family of formal languages designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. SKOS is built upon RDF and RDFS, and its main objective is to enable easy publication of controlled structured vocabularies for the Semantic Web. SKOS is currently developed within the W3C framework. All vocabularies used within the BHL-Europe environment shall comply to this standard.

5.8 OAI-ORE [standard]

Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[12] defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation.

---

[10] http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=d5320409c5160be4697dc046613f71b9a773cd9e [30.10.2009]

[11] http://en.wikipedia.org/wiki/Simple_Knowledge_Organization_System

[12] http://www.openarchives.org/ore/

5.9 Data Mapping – ETL [process]

Extract, transform, and load (ETL)[13] in database usage and especially in data warehousing involves:

* Extracting data from outside sources

* Transforming it to fit operational needs (which can include quality levels)

* Loading it into the end target (database or data warehouse)

The first part of an ETL process involves extracting the data from the source systems. The transform stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources will require very little or even no manipulation of data. In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the target database. The load phase loads the data into the end target (the BHL-Europe System). The use of the open source application software Pentaho BI Project will be investigated for BHL-Europe ETL processes.

5.10 Vocabulary Webservices [service]

A web service (also webservice) is traditionally defined by the W3C as "a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically Web Services Description Language WSDL). Other systems interact with the web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other web-related standards.[14] BHL-Europe will integrate available web services where useful. It is envisaged to use the uBio[15] web services to add taxonomic intelligence to BHL-Europe. uBio provides access to NameBank and Classification via SOAP and XML REST web services.

[13] http://en.wikipedia.org/wiki/Extract,_transform,_load

[14] http://en.wikipedia.org/wiki/Web_service

[15] http://www.ubio.org/index.php?pagename=services_overview

5.11 BHLE Data API [BHL-Europe Development]

This interface will provide search and retrieval functionality especially designed for the BHL-Europe Portal component. It will just serve specific items needed for high performance portal usage without transformation processes at Portal side.

6 Figures