ECP-2008-DILI-518001

BHL-Europe

Technical architecture status and progress report with particular focus on the development of the German prototype

Deliverable number	*D3.5*
Dissemination level	*Public*
Delivery date	*30 April 2010*
Status	*Version 1.0*
Author(s)	*AIT, ATOS, MfN, NHM, NHMW*

*e*Content*plus*

---

[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
Gerda Koch	AIT
Lee Namba	ATOS
Henning Scholz	MfN
Adrian Smales	NHM
Wolfgang Koller	NHMW

1.2 Revision History

Revision Date	Author	Version	Change Reference & Summary
15 April 2010	A. Herzog	0.1	1st draft
21 April 2010	A. Herzog	0.2	Draft taking comments of the BHLE-tech group into account
22 April 2010	W. Koller	0.3	Draft adding information on SMT
27 April 2010	H. Scholz	0.4	Draft incorporating partner feedback
30 April 2010	H. Scholz	1.0	Final version

1.3 Reviewers and Approvals

This document requires the following reviews and approvals.

Name	Position	Date	Version
Adrian Smales	Work Package Leader	30 April 2010	1.0
BHL-Europe tech group	TMB	30 April 2010	1.0
Henning Scholz	Project Coordinator	30 April 2010	1.0

1.4 Distribution

This document has been distributed to:

Group	Date of issue	Version
BHL-Europe tech group	15 April 2010	0.1
BHL-Europe tech group and consortium	21 April 2010	0.2
BHL-Europe consortium	22 April 2010	0.3
BHL-Europe tech group	29 April 2010	0.4
BHL-Europe consortium	3 May 2010	1.0

2   Purpose and document structure

This document is publicly available and describes the concepts of the BHL-Europe system. It will introduce the components needed within the system without technical background or details. It will cover what has been done so far within the project and will provide a milestone plan for the final development.

The document will discuss the current technical architecture status by explaining the available components and their key aspects. Each component serves a specific need within the BHL-Europe system. These needs are either serving user and system interaction or scalability and preservation.

Chapter 4 is describing the development on the German prototype. All components are discussed in relation to the German prototype to be delivered in Month 18. This prototype is not scalable nor does it provide a preservation or archival storage component. The prototype will be used to collect feedback for the final implementation.

The document finishes with a milestone plan for the final development and will outline the availability of functionality through the following 24 month and a certain period after.

3 Technical architecture status

In addition to the BHL Europe System architecture described in D3.4 this chapter describes the extension of the BHL-Europe System architecture to a global scale taking into account other regional BHL repository systems, global identifiers between systems, and replication procedures between them. The current diagram is shown in Figure 3-1.
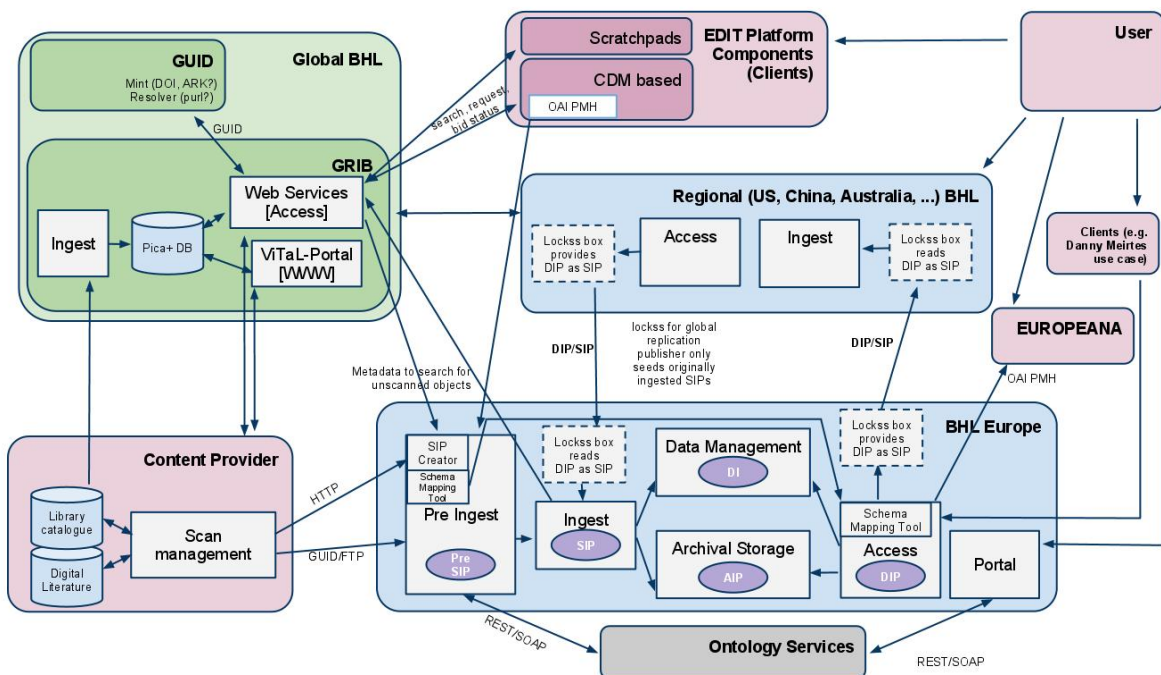

Figure 3-1 Architecture diagram

The technical architecture is based on the Open Archival Information System (OAIS) reference model. In addition to the core elements, BHL-Europe developed two modules that fulfil the roles Provider and Consumer. The provider role is covered by a Pre-Ingest module and the Consumer is represented by a Portal module.

The **Portal** is the user interface for all BHL-Europe users. Search and retrieval will be done within this component. The user will be able to search for books via a simple, one field search box and an advanced search based on various metadata fields like author, publication date or title. Details of the search functionality will also come from the use cases we currently develop and the results of the user requirement survey open from 15 March to 30 April 2010. Additional ontology services will help the users to find books by browsing multilingual taxons. The retrieval will provide the possibility to browse through the books of the result set

with a book reader application and to download whole books, serials or just parts of a book like an article. The portal will be based on Drupal[2].

The **Access** component is the API that the portal and 3rd party applications use to pull information from the BHL-Europe system. It handles the creation of result sets and the provision of dissemination information packages (DIPs). DIPs are the downloadable items displayed at the Portal. Access will be used to distribute the system's stored items throughout the web to other OAIS systems. This mechanism is used to support distributed storage and replication. One distribution mechanism can be Lockss[3]. The DIPs can be created using the Schema Mapping Tool[4]. The evaluation will be carried out in M12.

Data Management is used to store search and retrieval information about the items stored in Archival Storage. Data Management can be seen as the index Access will work with, when a search request is conducted. The information stored in Data Management is called Descriptive Information (DI). The main search index will be based on Solr[5].

Archival Storage is the component where all digital representations of the books are stored. This includes the metadata and all images of the scanned pages. The data stored in Archival Storage is called Archival Information Package (AIP). By using only AIPs it must be possible to recreate the whole system. Existing projects and software packages like Fedora[6], DuraSpace[7], ArchiveMatica[8], iRods[9] or HoneyCom[10] may be used. The evaluation will be finished in M16.

The **Ingest** component is responsible for taking Submission Information Packages (SIP), extracting the DIs and creating AIPs from the SIPs and passing them to Data Management and Archival Storage, respectively. It is also responsible for validating the SIP and monitoring the status of the Ingest procedures.

---

[2] Drupal: http://drupal.org/ [2010-04-14]

[3] Locks: http://lockss.stanford.edu/lockss/Home [2010-04-14]

[4] Schema Mapping Tool: http://code.google.com/p/schema-mapping-tool/ [2010-04-22]

[5] Solr: http://lucene.apache.org/solr/ [2010-04-14]

[6] Fedora: http://www.fedora-commons.org/ [2010-04-14]

[7] DuraSpace: http://duraspace.org/index.php [2010-04-14]

[8] ArchiveMatica: http://archivematica.org/wiki/index.php?title=Main_Page [2010-04-14]

[9] iRods: https://www.irods.org/ [2010-04-14]

[10] HoneyCom: http://www.honeycom.com.au/ [2010-04-14]

Pre-Ingest represents a provider of the Open Archival Information System (OAIS). It will help the content providers (libraries, digitization centres, etc.) to harmonize the data before it will be ingested. This will be done on a directory basis and the usage of ontology services to enrich the metadata. Pre-Ingest will push the SIP into Ingest once it is created. Within the SIP Creator the Schema Mapping Tool can be used to map the local library data to BHL Europe. The entire Pre-Ingest process will be managed by experienced BHL-Europe team members to assist content providers to understand the process and to successfully finish the harmonization process.

All these listed components are part of the OAIS. To support the work of the content providers, an international book index, the **Global References Index to Biodiversity** (GRIB) was introduced. It will assign a **Global Unique Identifier** (GUID) to each item (e.g. book) that is already scanned or will be scanned. The communication between the OAIS components is defined within the OAIS reference specification however the GRIB provides additional information and systems that are detailed here.

Content providers transfer all their metadata to the GRIB for deduplication, digitisation management and to receive a GUID for each item that they intend to or have already scanned. BHL-Europe will then ingest all GRIB records to create a Functional Requirements for Bibliographic Records (FRBR) structure within its Data Management component. This could also be done manually, but because of the existing features the GRIB can provide, BHL-Europe will take it from there.

As soon as the providers are finished with scanning, they will upload their data to the Pre-Ingest component via FTP and use the SIP creator module to check for errors within the uploaded Pre-SIP via a web browser.

When the item is ingested, the Ingest module will update the GRIB and mark the record with the supplied GUID as already scanned.

4   Progress report on the development of the German prototype

This chapter describes the current status of the prototype and some features already available within it. The prototype is currently available at http://bhl.ait.co.at but will be moved to http://prototype.bhl-europe.eu/ as soon as the servers are available. Figure 4-1 shows the home page of the prototype.
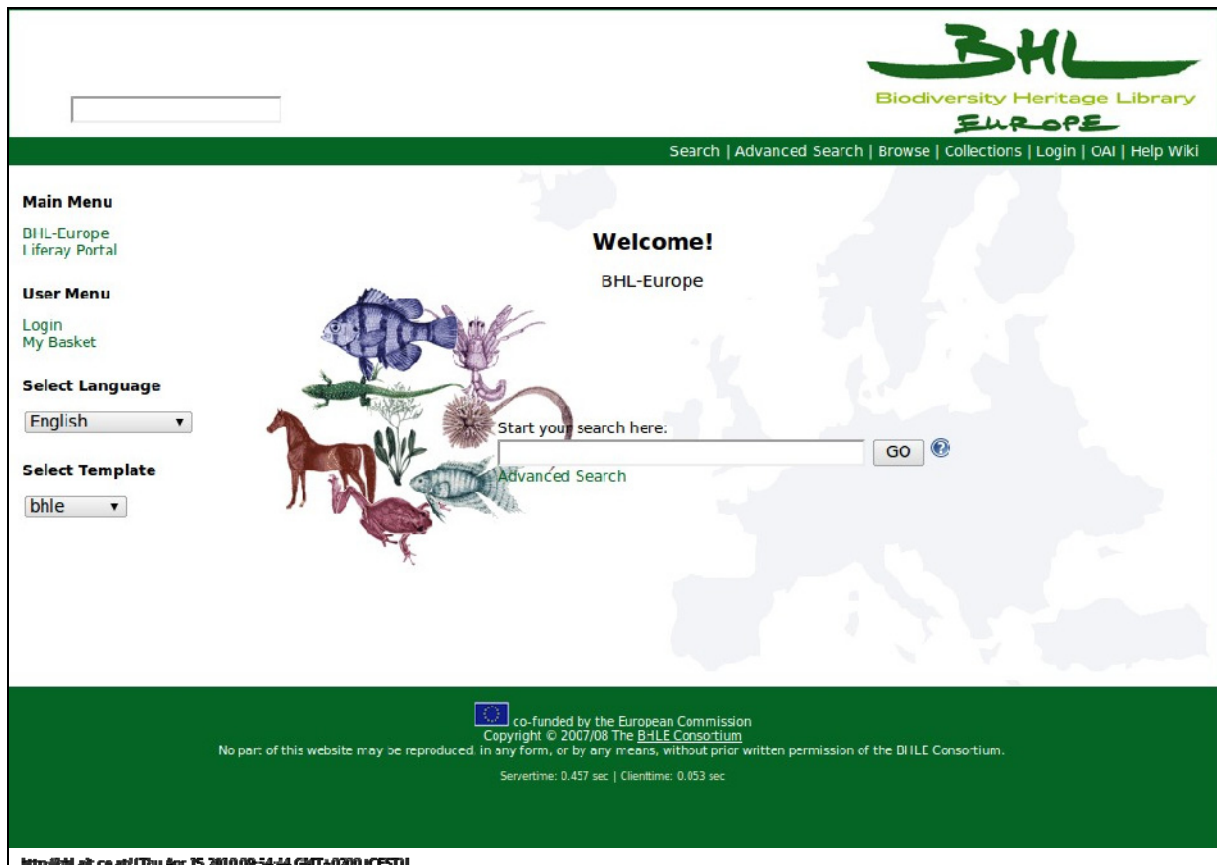
Figure 4-1 Home

4.1   Pre-Ingest

Pre-Ingest is used to harmonize data. Therefore a metadata mapping will be done prior to ingest. The currently manual procedure lets providers map their metadata fields to the predefined fields of BHL-Europe as shown in Figure 4-2 and Figure 4-4. These mappings are stored and configured within the system as part of Administration which can be seen at Figure 4-3. The current import is metadata only. This import will be shifted to GRIB for the final solution.

Figure 4-2 Mapping

Figure 4-3 Mapping configuration
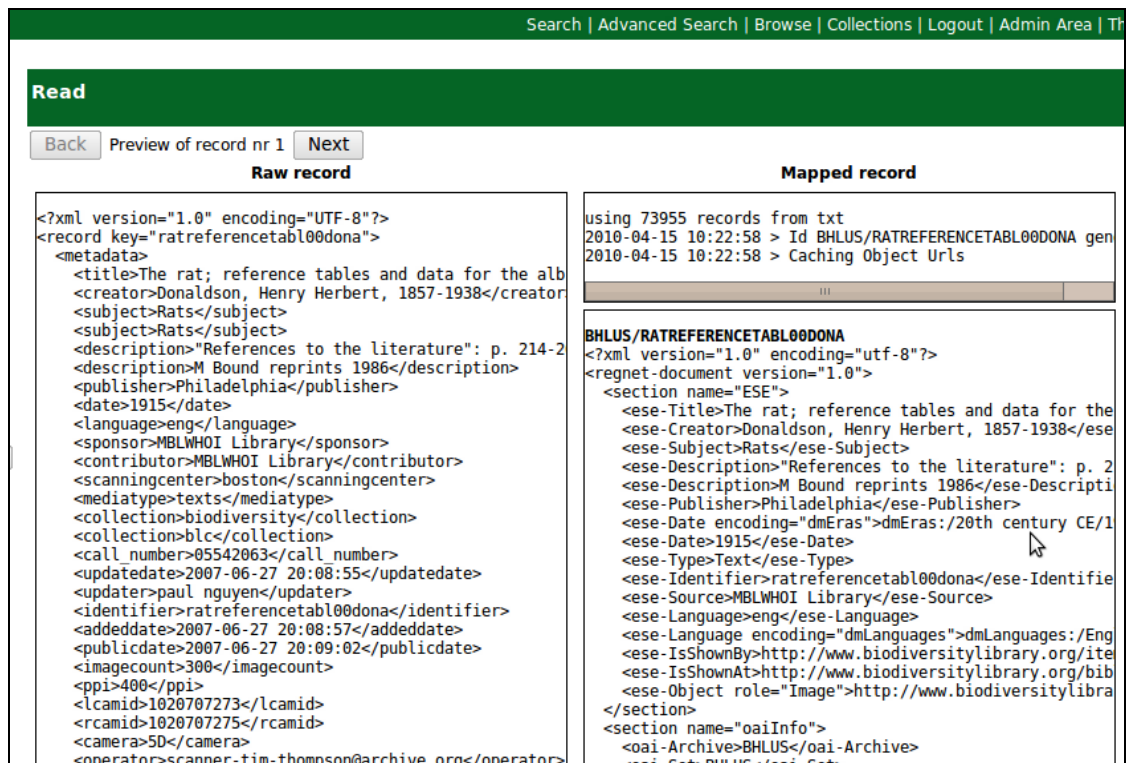
Figure 4-4 Edit configuration

4.2 Ingest

Ingest and Pre-Ingest are currently not separated. The conversion is done on demand which means there is no Pre-SIP in the prototype. Native data will be sequentially read and converted for Data Management.

The conversion can be checked before importing as shown in Figure 4-5.

Figure 4-5 Check SIPs prior to ingest

During the ingest, the data is enriched with multilingual thesauri terms. In the previous example, a language and a date thesaurus is used to unify records and make them searchable in multiple languages.

4.3 Data Management

Currently, the data management component can use either Zebra or Solr. The Zebra Index cannot be used due to the massive amount of records that BHL-Europe will face with the indexation of individual pages.

The Solr configuration will be tailored to fit BHL-Europe's needs.

4.4 Archival Storage

For the prototype, the Archival Storage will only cache one thumbnail for each item and store the raw and transformed metadata on a file server. This is neither safe nor meets the requirements of a backup and preservation system for the provider and therefore will be changed in the future.

4.5 Access

One of the Access methods for BHL-Europe is the OAI PMH provider for Europeana. An already working sample is shown in Figure 4-6. It uses a stylesheet transformation to present a user interface for the underlying XML.

Figure 4-6 OAI-PMH for Europeana Access

Another DIP is the RSS feed as shown in Figure 4-7. It can be used to stay up to date on certain queries.

Figure 4-7 Syndication for ingested items

4.6 Portal

The Portal is the multilingual user interface for the Access component. One of the multilingual functions is a dictionary based query expansion as shown in Figure 4-8. It will look up the search term in a predefined dictionary and return alternative search terms in the desired languages.

Figure 4-8 Query expansion with dictionary

Using this query, a multilingual result set is created at query time. An alternative is a multilingual result set based on the index. Therefore, the ontology services used to enrich the data during ingest can be used to retrieve a multilingual record set. This predefined ontology can be used as query parameter like shown in Figure 4-9.
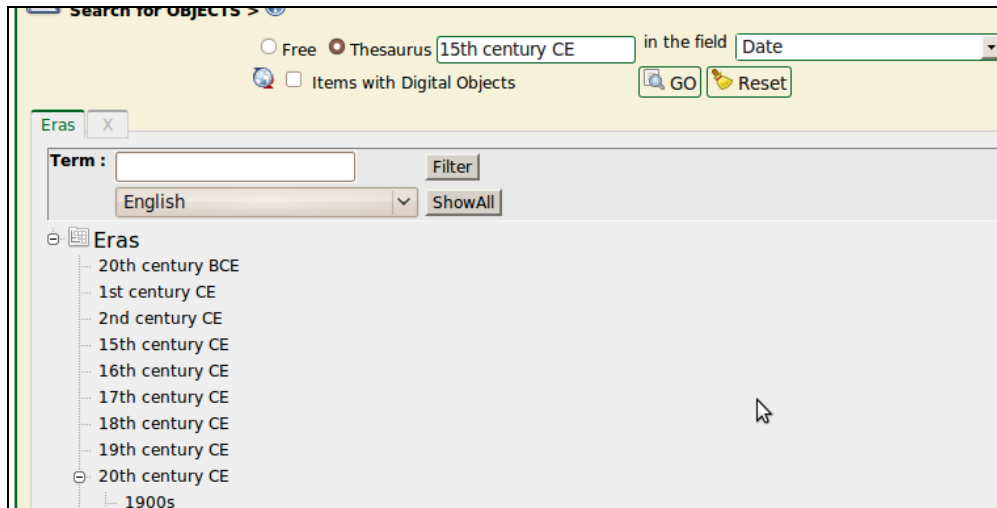
Figure 4-9 Ontology based search

The records that are enriched with ontology entries provide multilingual feedback on the search as shown in Figure 4-10.
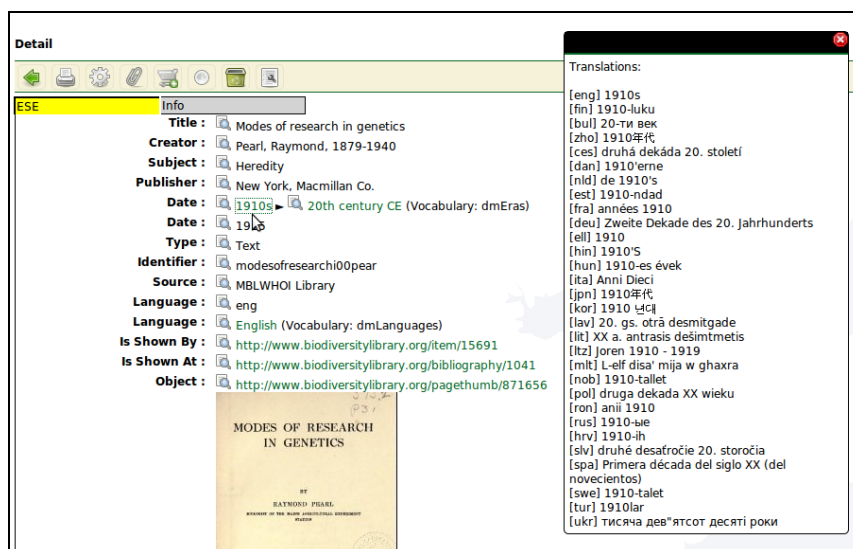
Figure 4-10 Ontology augmentation of records

Several other features like favourites, saved searches and a shopping basket are available as well. They are not described here in detail since this document's aim is to provide an overview on the core functionalities of the system.

5 Roadmap

This chapter drafts an outline for the tasks that need to be fulfilled in order to provide all desired future functionalities of the system.

5.1 Methodology

Since many functional requirements will be defined incrementally, the agile software development approach SCRUM was chosen for the distributed development. Scrum can be used within the PRINCE 2 methodology BHL-Europe uses.

5.2 Project backlog overview

This section provides an overview of the currently identified use cases for content providers

1. Description of the different processes
   a. Catalogue building
   b. Scanning planning
   c. GUID - Mint Report - Project Template Generation
   d. Scanning
   e. Pre-Ingest

2. Description of use cases within each process:
   a. Connecting to the GRIB
   b. (Bulk) Search within the GRIB
   c. (Bulk) Selection of literature to be digitised
   d. Indicating scanning intention
   e. GUID - Mint Report

3. Description of scenarios within each use case to define wire frames
   a. Web access
   b. Desktop application access

5.3 Sprints overview

The remaining 24 months have been split into 12 sprints each targeting a specific goal. The last two sprints are used to finalize the system without adding new features.

- M10   Sprint 0   Establish Development Team and Platform

- M12   Sprint 1   First mock-up integration GRIB + GUID

- M14   Sprint 2   Content provider - Pre-Ingest (upload files)

- M16   Sprint 3   PI – ING – AS – DM (1st ingest) + basic security

- M18   Sprint 4   DM-ACC (1st search – with a web service call not via the portal)

- M20   Sprint 5   DM - AS - ACC (1st dissemination DIP – ask for pdf / zip / …)

- M22   Sprint 6   ACC - PORTAL (basic search retrieval)

- M24   Sprint 7   All module feature improvement + integration (alpha release)

- M26   Sprint 8   Administration + Finish Security

- M28   Sprint 9   Continue development of all modules

- M30   Sprint 10   Finish all modules (Beta version)

- M32   Sprint 11   Beta Testing and Bug Fixes (Feature Freeze)

- M34   Sprint 12   Final Testing and Bug Fixes

- M36   Go live   Launch Release

6 Acronyms and Abbreviations

ACC	Access
AIP	Archival Information Package
API	Application Programming Interface
AS	Archival Storage
BHL	Biodiversity Heritage Library
CRC	Cyclical Redundancy Check
DI	Descriptive Information
DIP	Dissemination Information Package
DM	Data Management
ESE	Europeana Semantic Elements
ETL	Extract, Transform, Load
Europeana	European Digital Library
GRIB	Global References Index to Biodiversity
GUID	Global Unique Identifier
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
ING	Ingest
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LOCKSS	LOCKSS (Lots of Copies Keep Stuff Safe), based at Stanford University Libraries, is an international community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content
MARC	MAchine-Readable Cataloging
METS	Metadata Encoding and Transmission Standard
MIME type	Internet Media Type
OAI	Open Archives Initiative

OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OCR	Optical Character Recognition
ODBC	Open Database Connectivity
PDI	Preservation Description Information
PI	Pre-Ingest
RDF	Resource Description Framework
REST	Representational state transfer
RSS	Really Simple Syndication (RSS) is a lightweight XML format designed for sharing headlines and other Web content.
SIP	Submission Information Package
SOAP	Simple Object Access Protocol providing a simple and lightweight mechanism for exchanging structured and typed information between peers in a decentralised, distributed environment using XML proposed under the W3C.
SWORD	Simple Webservice Offering Repository Deposit
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	EXtensible Markup Language
XSLT	XSLT (XSL Transformations) is a declarative, XML-based language used for the transformation of XML documents into other XML documents.

7 Figures