COMPONENT REPORT

Project Acronym:	**OpenUp!**
Grant Agreement No:	**270890**
Project Title:	**Opening up the Natural History Heritage for Europeana**

C2.6.1 Draft data flow documentation and storage concept

Revision: V0.3

Authors:

Gavin Malarky
with specific input from Anton Güntsch, Simon Kennedy, Gerda Koch and Walter Koch

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
V0.1	23 May 2011	Gavin Malarky	NHM	1st draft
V0.2	27 May 2011	TMG		Draft updated following review by TMG members
V0.3	7 June 2011	Coordination Office/ Coordinator	BGBM	Editorial changes (adoption of template)/ minor changes to text.

Statement of Originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Distribution

Recipient	Date	Version	Accepted YES/NO
Gavin Malarky	27 May 2011	0.2	YES
TMG	27 May 2011	0.2	YES
Work Package Leader (Gavin Malarky)	27 May 2011	0.2	YES
Project Coordinator (W. Berendsohn)	8 June 2011	0.3	YES

Table of Contents

1 PURPOSE AND DOCUMENT STRUCTURE

The aim of this document is to provide a draft description of the technical architecture, data flows and data storage for the technical implementation of deliverable D16 (Distributed metadatabase). Overall this document will cover the initial concepts and outline the key components that will be built to deliver the technical implementation of the distributed metadatabase. This document will evolve as components are built and questions answered.
The documentation for the final implementation will be a separate document due in M35.

2 OVERVIEW

OpenUp! Technical Architecture
OpenUP! will make available Natural History high quality images, movies, animal sound files, and artwork through EUROPEANA. Access will be based on the established technical infrastructure of the Global Biodiversity Information Facility (GBIF) including the BioCASE network (Biological Collection Access System for Europe). OpenUp! will develop a pathway and data flow from content publishers in the BioCASE network and GBIF to provide a steady stream of multimedia objects to EUROPEANA.
All content providers will be expected to present metadata in the ABCD standard and it is therefore necessary for all the content providers to utilise the BioCASE Provider Software (http://www.biocase.org/products/provider_software/).

The OpenUp! data workflow will utilise services (Ontology services) being developed as part of the project that allow each content provider to enhance and validate their metadata by:

- Enrichment of metadata towards compliance with EUROPEANA standards and compliance with ABCD.
- Quality control of species names
- Incorporation of multilingual metadata, in particular vernacular names of organisms
- Incorporation of metadata that will allow semantic linking of content with other domains, particularly scientific organism names

Content providers will be expected to utilise these services and be responsible for the quality of their metadata and content. Access to these services will be through the Data Quality Toolkit being developed in WP2 by FUB-BGBM. Content providers will choose when to have their data analysed and the toolkit will return information in a human readable format to allow content providers to update their own records.

When content providers have sets of records that they wish to be harvested they will request a harvest. Metadata in the ABCD standard will then be harvested to a central aggregation point using the Harvesting and Indexing Toolkit (HIT, http://code.google.com/p/gbif-indexingtoolkit/) developed by GBIF. Content providers will provide records in sets which are yet to be defined but are likely to follow those used by GBIF. Content providers can ask for a set to be re-harvested when the content has changed.

Services developed by AIT and hosted on the central aggregator will map between ABCD data standards and the EUROPEANA metadata scheme ESE. At the same time the metadata will be checked against the Ontology Services for further enhancement. It will be expected that integrity of metadata will have been checked by the content providers but could optionally be rechecked at this stage.

Harvesting to the central aggregator will be an on-going process and the data in ESE standard will be stored in a central metadata database. A link checker will be developed by NHM to check that the content that has been harvested remains available through the supplied URLs. Where supplied URLs do not respond the providers will be informed and provided with advice; but if the problem persists the content will be flagged as unavailable and will not be included in EUROPEANA. It will be the provider’s responsibility to make sure the content is available. Providers should use persistent identifiers for their content but OpenUp will not provide a mechanism to do this but rather point providers at best practice.

Harvesting of metadata to EUROPEANA will occur during harvesting cycles which are expected to occur approximately every six months but there is currently no timetable available. Harvesting will occur through the OIA-PMH service which is a component of the central aggregator developed by AIT. If thumbnails are not provided then during harvesting EUROPEANA will generate thumbnails from downloadable images if a URL for this purpose is provided or else a default image will be used.

Once content has been harvested by EUROPEANA there will be an on-going process of link checking to ensure providers are informed if their content goes off line.

3 KEY COMPONENTS

3.1 Data Quality Toolkit

The data quality toolkit will be an application (local or web) developed by FUB-BGBM which will query the Ontology services being developed in WP3/4/5 to check data quality and provide enrichment. The toolkit will be managed by the content provider and they will use it to ask for data quality reports which will be provided in a human readable format so they can review and apply to their data if they choose. The same web based ontology services being developed for the toolkit to query will also be available central aggregator for further enhancement of the data before submission to EUROPENA.

Ontology Services will include:

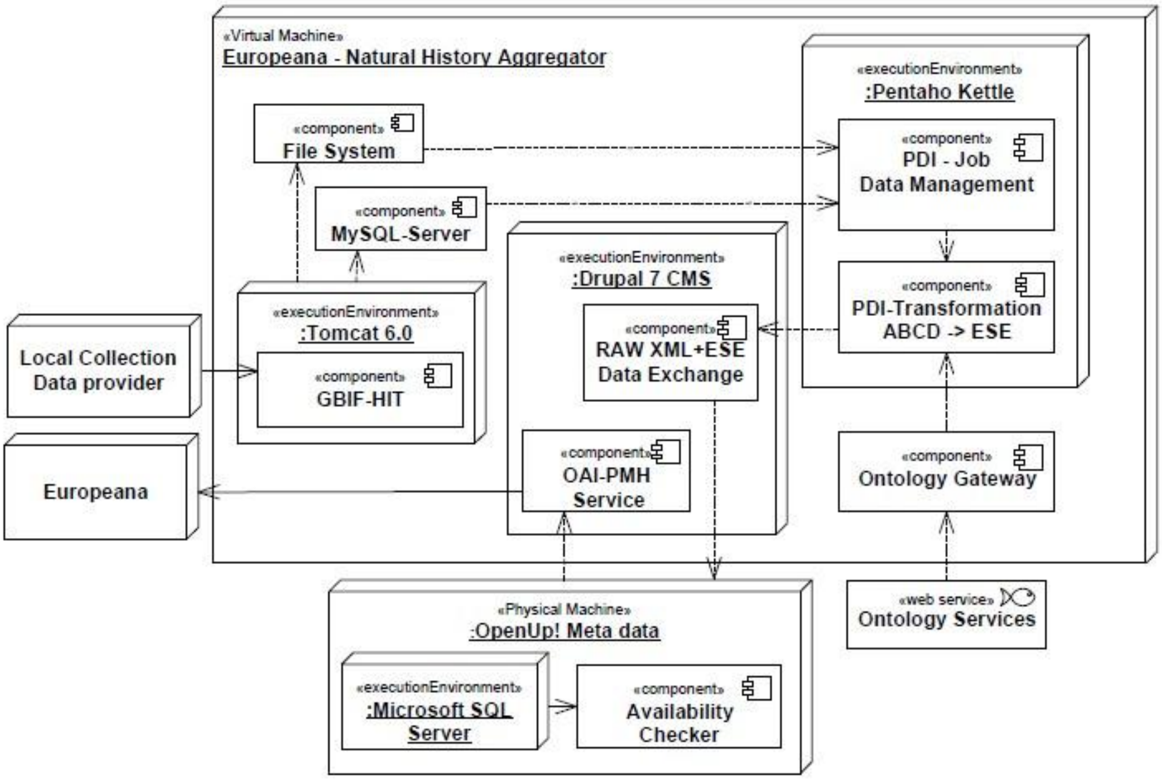
- Botanical data quality service
- Zoological data quality service

- Common Names Service
- OpenUp! Data Integrity Service

A description of these services as well as simple mockups will be delivered in June 2011.

3.2 Natural History Aggregator

Figure 1 includes a first iteration of a UML deployment diagram showing the nodes and components to be implemented when setting up the "Europeana - Natural History Aggregator". A further refinement was elaborated after the installation and test of the GBIF-HIT aggregator at AIT's test and staging servers.



The test server (test111) is set up as VMWare Image including an Ubuntu 10.04 Linux System.

Data (eg. ABCD records) generated by the GBIF-HIT component is stored in the Filesystem (/opt/hit) and in different tables of a connected MySQL data base.

For preparing (PDI-Job) ABCD-XML records for transformation (PDI-Transformation) and input into the RAW-XML component the Pentaho Data Integration environment will be used.

Transformed (ESE) and original ABCD-records will be imported into nodes of a Drupal7 CMS. The Drupal nodes are following the METS structure which allows further integration of an archival system in case content (images, etc) has to be preserved (not part of the OpenUp! DoW). A Drupal7 Test Environment will be available at: http://test111.ait.co.at

Drupal will prepare the export of the records to be collected by the Europeana Harvester and unload data into a directory which can be accessed by the OpenUp! Meta Data Database. This data store is located at and managed by a SQL-Server RDBMS.

The Availability Checker finally determines what records are handed over to the OAI-PMH service (Data Provider) component.

The OAI Data Provider located at the Drupal7 execution environment will be based on a refactored PHP application which is already used in different Europeana projects (eConnect/DISMARC, EuropeanaLocal).

During the Transformation Process at the Pentaho execution environment data can be added by looking up vocabularies and consuming webservices providing access to (multilingual) vocabularies. This process has been successfully implemented in the aggregation platform described above as well tested in other projects like BHL-Europe (publication , presentation) and in collection management systems (attached). The interaction with uBios Taxon Finder is outlined in a BHLE presentation (August 2009).

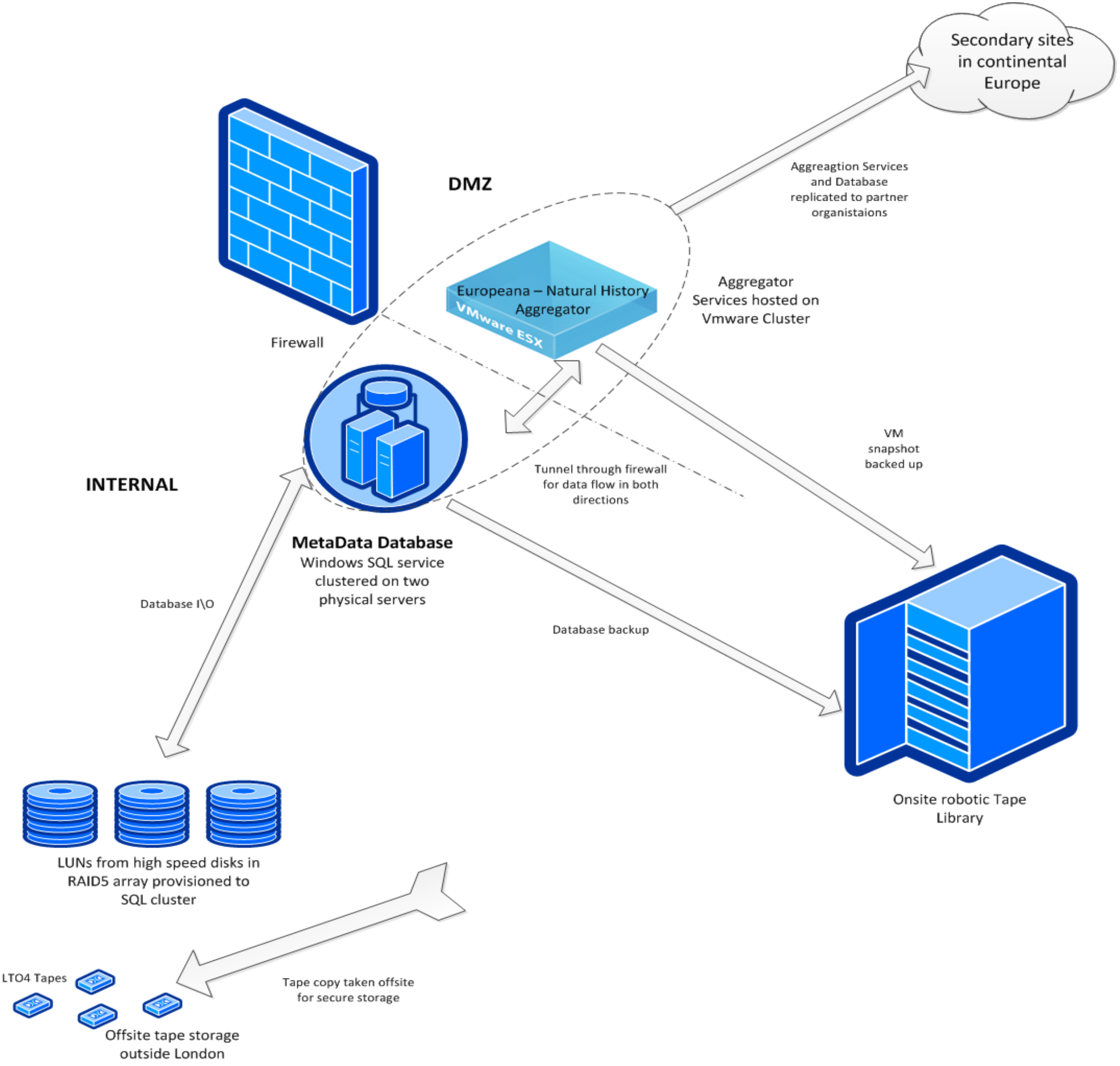
To harmonize the input coming from different "vocabulary WebServices" an "Ontology Gateway" will be installed in front of the transformation environment. As reference for such a component AIT's TGN-Vocabulary WebService can be used.

3.3 Metadata Database

Following the transformation process in the central aggregator, records will be stored in the metadata database which will be designed to provide the link checking tool access to relevant URLs and also content provider contact details. Extra fields will be included to allow tracking of link checking progress and also to flag records that have broken links and will not be made available to EUROPEANA.
Suitable records in the metadata base will be made available to OAI-PMH service component of the central aggregator for harvesting.

3.4 Hosting of the primary copy of the Metadata and Aggregator at the NHM.

Figure 2 shows a representation of the key physical infrastructure hosting the primary metadata and the aggregator at the NHM.


Secondary sites in continental Europe
DMZ
Aggreagtion Services and Database replicated to partner organistaions
Europeana – Natural History Aggregator
VMware ESX
Aggregator Services hosted on Vmware Cluster
Firewall
VM snapshot backed up
Tunnel through firewall for data flow in both directions
INTERNAL
MetaData Database
Windows SQL service clustered on two physical servers
Database I\O
Database backup
Onsite robotic Tape Library
LUNs from high speed disks in RAID5 array provisioned to SQL cluster
LTO4 Tapes
Tape copy taken offsite for secure storage
Offsite tape storage outside London

The primary metadata store will be hosted internal to the NHM firewall protected from attack from external IP addresses. The aggregator will sit in the DMZ and tunnels will be created to allow data to flow from the aggregator to the metadata store.

The primary metadata store will be held on a Microsoft SQL database hosted on clustered Microsoft SQL 2008 running on two dedicated IBM blade servers. Storage for the cluster is provisioned from high speed SAS disks in a RAID5 array on an IBMds4300 disk subsystem purchased to host BHL-e infrastructure.

The database will be backed up nightly to a robotic tape library using IBM TSM with copies of the last 10 backups kept. A snapshot of the VM hosting the aggregator will also be backed up to tape. Copy sets of the backup will be made and the tapes sent offsite to a specialist data handling company with storage facilities outside London. In the event of loss of the disks and onsite tape backups these offsite tapes can be used to rebuild the database and restore the aggregator

3.5 Replication of the Metadata

Mirror sites for the metadata database and aggregator will be hosted at UCPH, NBGB, and BGBM. This means that in the unlikely event that the NHM site is down then both the data and the service can be brought online at a secondary host site.

3.6 Link Checking

It is important that the URLs OpenUp! delivers to EUROPEANA work and do not result in errors.

The NHM will review open source tools and choose the best to create a service that will check if URLs in either ESE fields europeana:isShownAt, europeana:isShownBy in records held in the metadatabase are working links. Where a URL is provided for thumbnail creation it will be provided in the europeana:object element and it would be good if we could also test it. Once in EUROPEANA the links in europeana:isShownAt and europeana:isShownBy will be checked occasionally for valid responses.

Until a tool is chosen and tested we do not know what the performance will be. It is likely that once the number of records grow in future years the link checker will take considerable time to check all the links in the metadata base and this will have to be taken in to account when we decide how regularly a link will be checked or whether records that have not been checked can be submitted to EUROPEANA.

The tool will be limited to check for broken links and will not be able to validate that the content at the supplied URL is correct. If a broken link is found then the content provider will be informed and advice provided by email.

The link checker will not be used to check any links on content provider’s webpages as this would significantly slow down the process. It is the content provider’s responsibility to maintain their own webpages and make sure any links on them are functional.

The link checker tool will utilise the metadatabase to identify URLs that require checking, flag records with broken links and to get contact details of the relevant content providers to email if their links are broken.

3.7 Thumbnail creation

Thumbnails used by EUROPEANA will be generated by EUROPEANA during harvesting as described in appendix IIa (Europeana Portal Image Policy). EUROPEANA will use the URL provided in the ESE europeana:object element. The URL must point at a downloadable Image object (EUROPEANA cannot currently create thumbnails for sound or video objects) that the content provider wishes to be used for thumbnail creation and which must adhere to the requirements in the Europeana Portal Image Policy.

The links for the downloadable image to be used for thumbnail creation can be separate from links (URL supplied in Europeana:isShownAt or europeana:isShownBy) which EUROPEANA will use to link back to either published images or a webpage with full information content containing a number of images or other multimedia files. We will need to work with providers to identify which object URL is used to generate the thumbnail for a page and how to accommodate situations where providers already have their own thumbnails.

Thumbnail creation will not work with Audio or Video files and there is no capacity for EUROPEANA to utilise snippets. If no thumbnail is provided and no published image is available to use generate a thumbnail then EUROPEANA will use a default image depending on type of object as specified in europeana:type. We will need to decide if the default EUROPEANA thumbnail is acceptable or whether we should look at alternatives.

i. ACRONYMS

ABCD	The Access to Biological Collections Data (ABCD) is an XML-based common data specification for biological collection units
BHL	Biodiversity heritage Library
CMS	(Web) Content Management System
DMZ	DeMilitarised Zone. Perimeter subnet hosting services exposed to external access.
ESE	Europeana Semantic Elements
GBIF	Global biodiversity Information Facility
HIT	Harvesting and Indexing Toolkit developed by GBIF
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
PDI	Pentaho Data Integration part of the Pentaho Open Source Business Intelligence Suite
PHP	General purpose scripting language: Hypertext Pre-processor
RAID	Redundant Array of Independent Disks
RDBMS	Relational Database Management System
SAS	Serial Attached SCSI: I\O type for disks
SQL	Structured Query Language
TSM	Tivoli Storage Manager. IBM suite used to backup data
URL	Uniform Resource Locator
VM	Virtual Machine
XML	EXtensible Markup Language

ii. APPENDICES

ii.a) Europeana Portal Image Policy - January 2011

http://version1.europeana.eu/c/document_library/get_file?uuid=6b52d4be-6a4d-443a-842a-ab991bca2b1f&groupId=10602

This policy has been developed to as part of the initiative to improve the features and functionality of the Europeana portal by improving aspects of the data submitted. This document specifies the standards for the source images that should be provided for the creation of small images for use in the portal and explains how they are used. The earlier document “Thumbnails in Europeana Portal” should no longer be used for reference.

To give users a preview of the objects they have found and to make an attractive interface to the portal, Europeana creates images in two sizes from the objects providers submit. If the user is not to be disappointed these small images must be of a reasonable quality. They are generated from the source images whose links are provided in the metadata and the small images resulting are cached in the Europeana system. Note that no high quality or large format source objects are stored in the Europeana system.

What sort of images should be provided as the source?

The URL of a source image from which Europeana can generate the required small images should be provided in the europeana:object element. This may be the same URL as given in europeana:isShownBy.

This URL must link directly to an object which:

- is an image file (e.g. http://www.server.org/image.jpeg) and NOT an image embedded within a webpage.
- has a width of at least 200px
- is ideally a jpg file (or another image format supported by ImageMagick1)
- is alternatively a pdf, in which case the images will be created from the first page of this pdf. (Providers should ensure that the first page is a suitable image and not a blank page or a page containing the colour scales etc.)

Providers should not supply a link to an image that is itself already the size of a thumbnail as this produces very poor quality results when it is used in the portal functions described. Similarly, the source image should not have a watermark nor should it be a local default thumbnail image as these do not give a good result.

How are the small images used in the Europeana portal?

Two types of images are created and cached known as ***briefDoc*** and ***fullDoc***. These are then further manipulated in size to fit the places where they will appear in the portal.

briefDoc image:

Height=110 pixels

If the image provided is smaller than 110px it is scaled up to 110px

If the image provided is larger than 110px it is scaled down to 110px **fullDoc image:**

Supported file formats of ImageMagick http://www.imagemagick.org/script/formats.php

.

Width=200 pixels

If the image provided is smaller than 200px it remains unchanged

If the image provided is larger than 200px it is scaled down to 200px

What is the Europeana process for creating the images?

Using the link provided in the europeana:object metadata element the source image is accessed. ImageMagick software is used to create the two small images described previously. To enable a rapid presentation to the user the newly created images are cached. The source image is not cached or stored anywhere in the system.

Which image format is used for which function in the portal? Function	Type of cached image (brief/fullDoc)	Image size
Home page carousel		Width 70*
Search results page > brief display > gallery view	briefDoc	briefDoc size
Search results page > brief display > list view	briefDoc	Height 50*
Search results page > full display	fullDoc	Width 200
Search results page > full display > related content	briefDoc	width 25 and cropped square

What happens if a provider cannot provide an image meeting these specifications?

Where no link can be provided to a suitable source image then Europeana uses a default image corresponding to the type of object as specified in europeana:type. A results page with a large number of these default images gives a poor user experience and future policy may be to rank objects with proper images higher in the search results than objects with only a default image.