D16 – DELIVERABLE 2.3.5

Project Acronym: **OpenUp!**

Grant Agreement No: **270890**

Project Title: **Opening up the Natural History Heritage for Europeana**

Distributed metadatabase fully operational

D16 – Deliverable 2.3.5

Revision: 4a

Authors:

Anton Güntsch   BGBM

Gavin Malarky   NHM

Simon Kennedy   NHM

0 REVISION AND DISTRIBUTION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
1	2012-05-30	Anton Güntsch (BGBM), Gavin Malarky (NHM), Simon Kennedy (NHM), Dominik Röpert (BGBM)	BGBM, NHM	BGBM/NHM planning mirror system planning meeting in London and first system draft.
2	2013-02-20	Anton Güntsch (BGBM), Gavin Malarky (NHM)	BGBM, NHM	Draft report.
3&4	2013-02-26	Anton Güntsch (BGBM), Gavin Malarky (NHM)	BGBM, NHM	Final report after incorporating comments from TMG
4a	2013-02-26	A. Michel and P. Böttinger (Coordination Team)	BGBM	Minor editing

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This deliverable contains original unpublished work except where clearly indicated otherwise.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
Discussion of mirroring concept in WP2/WP3 meeting in Tervuren	2011-12-15	0	YES
Presentation in TMG meeting in Paris	2012-03-19	0	YES
Report distributed to TMG (AIT, BGBM, GBIF, IBSAS, MFN, MRAC, NHM, NHMW, RBGK, UH)	2013-02-20	2 & 3	YES
Work Package Leader (Gavin Malarky, WP2)	2013-02-26	4	YES
Project Coordinator (W. Berendsohn, BGBM)	2013-02-26	4a	YES

Table of Contents

1 DESCRIPTION OF WORK

The OpenUp! technical infrastructure consists of a set of distributed data sources providing multimedia data (images and sounds) and their associated meta information using established GBIF and BioCASE technologies, a central aggregator harvesting and transforming the information in a format which can be processed by Europeana (fig.1), and Europeana itself harvesting the data from the aggregator and providing them via the Europeana portal.
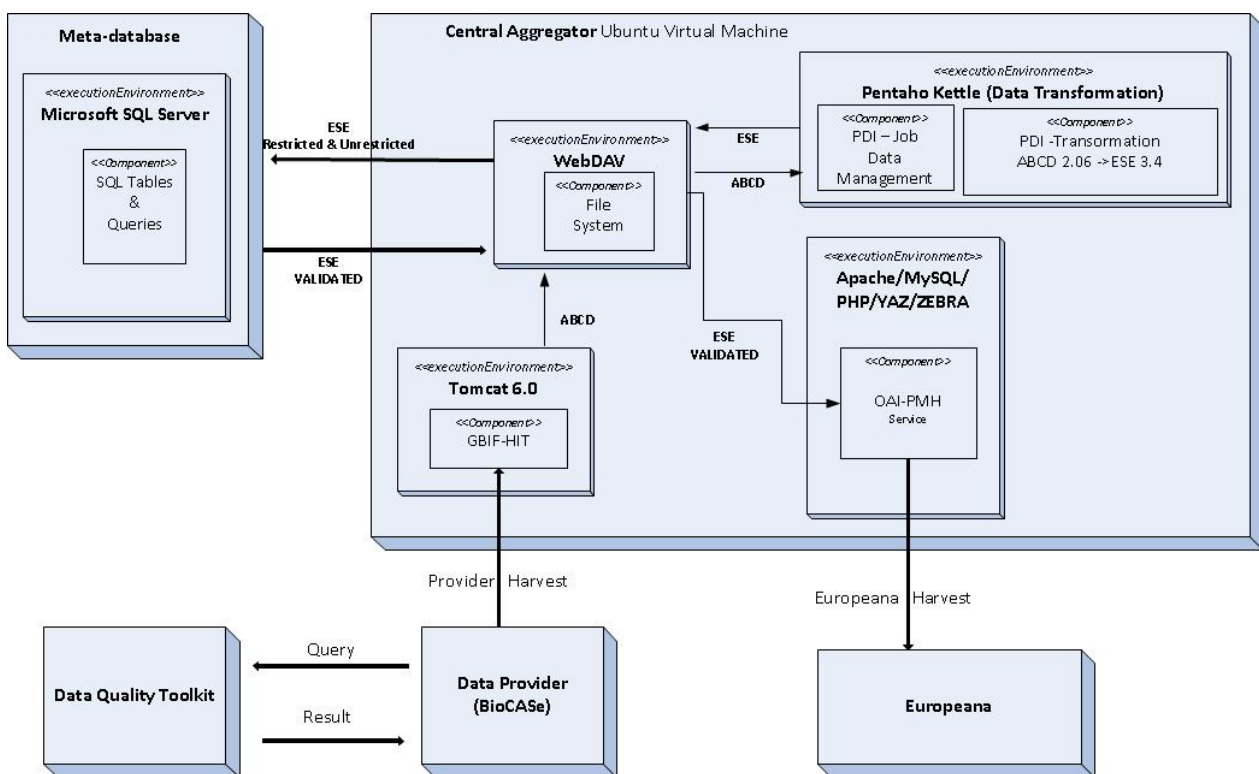

Fig. 1: Central OpenUp! Europeana Natural History Aggregator

An essential component of the central OpenUp! Aggregator is a database running on Microsoft SQL-Server which holds an updated copy of all Metadata records processed and provided by OpenUp! By February 2013 this database holds about 14 Gigabyte of data.

The database is used by project partner AIT to store the harvested and post metadata enhanced records (ESE XML format). Links to multimedia information are extracted and used by the link checker to separate working from non-working multimedia URLs. Records with good URLs are then stored in a dedicated table and picked-up and exposed to Europeana for harvesting.

In case of a system failure at data provider or data aggregation level, Europeana will still be able to deliver the entire content from Natural History Collections based on their own caching mechanisms. However, re-

indexing of all OpenUp!-providers may take between several days and weeks (depending on the availability and responsiveness of source databases) so that distributed back-up systems of the central Metadatabase would speed-up the recovery-process significantly.

We have set up and tested a working distributed backup mechanism between NHM (central aggregation site) and the BGBM (first mirror). In a first attempt, we tested a system that synchronizes both database servers at transaction level with individual transactions on the central immediately performed on the linked mirror servers. The tests turned out that immediate synchronization of very high numbers of transactions during the OpenUp! harvesting phases lead to high network-loads and are of limited value for the overall backup-capabilities. We therefore decided to choose a slightly more asynchronous mechanism and transfer copies of the meta-database to the mirrors after major updates of the central aggregation site only.

For a secure data transfer operation we are using SSH File Transfer Protocol (SFTP). With the present setup a transfer of the entire OpenUp! Metadatabase takes about 8 hours which is sufficiently fast if we assume that a re-harvesting of OpenUp! providers takes place once every two or three months.

The mirroring system is now fully operational between NHM and BGBM. According to the performance indicator 6 (DoW, page 70), two additional continental mirror sites will have to be set up during the last project year. Both the National Botanic Garden of Belgium (NBGB) and the University of Copenhagen (UCPH) have already started to setup the necessary server infrastructure so that we expect the extension of the backup network working by mid of 2013 latest.

The Meta-database holds copies of harvested records in ESE XML format in both restricted and unrestricted formats. In addition information is maintained on availability of URL links within the records so that records with non-functioning URLs are not passed back to the OAI-PMH service for harvesting by Europeana (part of component C2.5.0 Availability Checker).

In addition we are planning to create full copies of the OpenUp! infrastructure as VMWare virtual machine files including the Operating System, OpenUp! central software components, and data as part of long-term sustainability activities in Task 2.3. NHM and BGBM system administrations are presently discussing ways of transferring large volumes of data (>250GB) and a first running prototype is expected in spring 2013.

To enable the full infrastructure to work at BGBM the replication of the infrastructure will also include the SQL code to enable the database to function within the OpenUp! infrastructure. The code will only need to be updated if any changes are made as part of ongoing development. The database copies held at NBGB and UCPH will only hold data as there is no need for the extra code without the rest of the OpenUp! infrastructure. These replication sites will hold the data in MySQL which will be delivered through SFTP having been converted to MySQL at the NHM.