DELIVERABLE

Project Acronym: **Europeana v3.0**

Grant Agreement number: **620484**

D5.2 - Review of Europeana's logical and technical architectures

Revision: 1 (April 2015)

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

Authors:

Yorgos Mamakis (Europeana Foundation)
Pavel Kats (Europeana Foundation)
Maike Dulk (Europeana Foundation)

Revision	*Date*	*Author*	*Organisation*	*Description*
0.1	*April 2015*	*Yorgos Mamakis* *Pavel Kats* *Maike Dulk*	*EF* *EF* *EF*	*Initial Version*

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Introduction

Over the past few years Europeana has been transitioning from a family of interrelated projects of aggregation and dissemination of European cultural heritage data to the pan-European platform and infrastructure for storing, managing and sharing this data in many ways. As of 2015, Europeana will be funded by the European Commision as a Digital Service Infrastructure (DSI), reflecting the role of Europeana as the cultural digital infrastructure of Europe, similarly to network and transport infrastructures.

These changes have had a profound effect on how Europeana designs, develops and maintains its technical architecture. From standalone tasks and modules, serving specific needs of a project during its limited timeframe, we move towards an integrated platform for our data partners to stream large amounts of metadata and content from their local repositories towards global dissemination hubs following the COPE vision, *Create Once Publish Everywhere*. Implementing this new vision for Europeana requires a fundamental overhaul of the technical architecture. This work has already started under the projects Europeana v3.0, Europeana Creative, and Europeana Cloud and will continue into the next funding round of Europeana as DSI.

The goal of this deliverable is to provide a snapshot of this process by the moment the project Europeana v3.0 is close to its end. It is intended for a reader with a modest

technical background but should be fairly comprehensible even to an unprepared one. The deliverable describes the important components of the Europeana architecture and how they support processes of data aggregation and dissemination.

Logical Architecture

Data stores

Europeana keeps data in many forms for various internal and external needs. For this, it uses several types of data stores, implemented by available open-source software technologies. In this section these types of data stores, essentially basic building blocks of the Europeana infrastructure, are introduced.

Document Database - MongoDB

Document databases are a popular technology for storing semi-structured data, like EDM, used by Europeana to store all its records. We use one of the leading document database technologies, the open-source MongoDB database for storing EDM records and some other kinds of internally used data.

Search Index - Apache Solr

To allow efficient search across its dataset, Europeana uses the mature open-source Apache Solr solution for its search index. The index allows searching for records using each one of the EDM fields which were defined searchable.

Relational Database - PostgreSQL

Relational databases are the traditional technology to store structured data. Europeana uses PostgreSQL, a popular open-source relational database, to store information about MyEuropeana, private space for users to tag and annotate Europeana records.

Graph Database - Neo4J

EDM entities can form complex hierarchies for various needs. For example, in the archival domain it is common to see independent records, each one having its own representation as EDM, to be linked together by a meaningful relation. To efficiently store hierarchical structures like this, graph databases are used. Europeana employs one of the most popular open-source solutions, Neo4j.

Overview

Diagram 1 shows the schematic overview of the entire Europeana architecture.

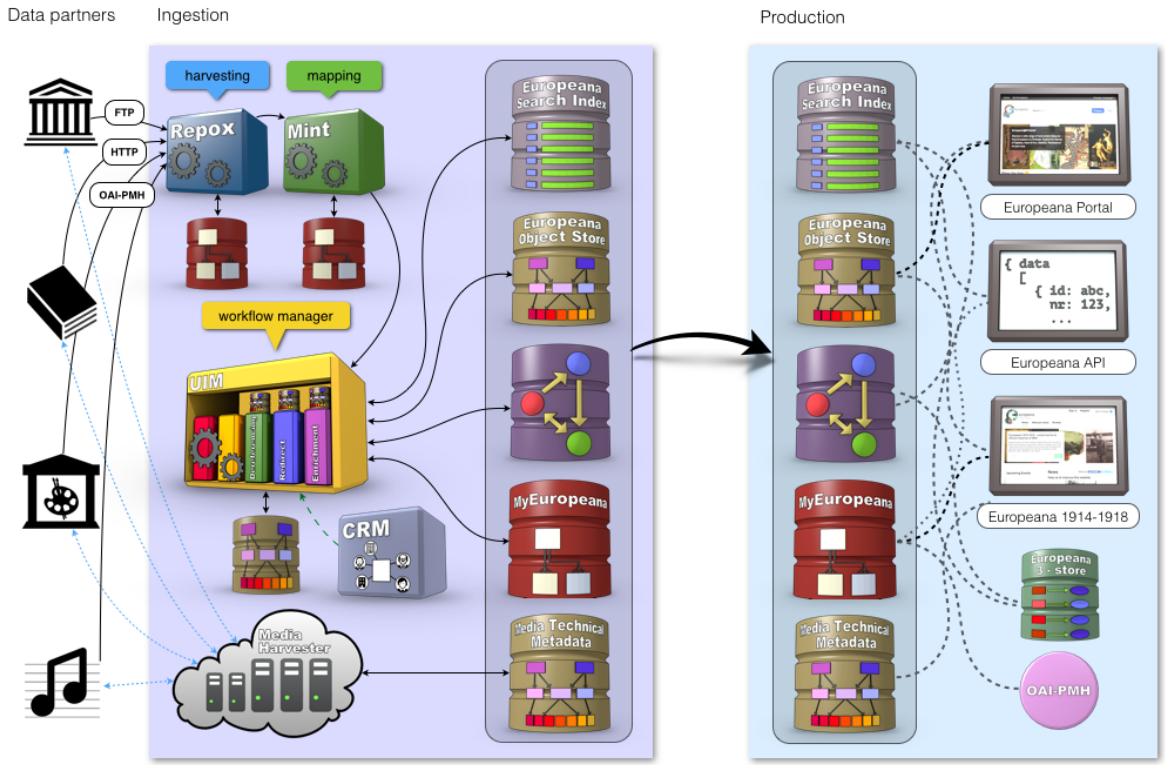

Diagram 1: Overview of Europeana Software Architecture

From data partners' repositories records flow into the complex ingestion system. In this system the records are mapped to EDM, assigned new identifiers, and passed through a series of processing flows which increase their discoverability. The Media Harvester engine, running in parallel, analyses media resources referenced by the records and extracts their technical metadata. Ultimately, the ingestion system produces a number of data stores which altogether represent the entire Europeana dataset at a given point in time and can be published. The publication process is the actual transition of this dataset to the production environment, from which it is served further on to consumers.

Ingestion

The ingestion process of Europeana, depicted on Diagram 2, is a complex sequence of ingestion workflows orchestrated by the Unified Ingestion Manager (UIM) application framework. Each workflow is implemented by a UIM plugin that either performs the workflow tasks directly or communicates with an external application to perform it.

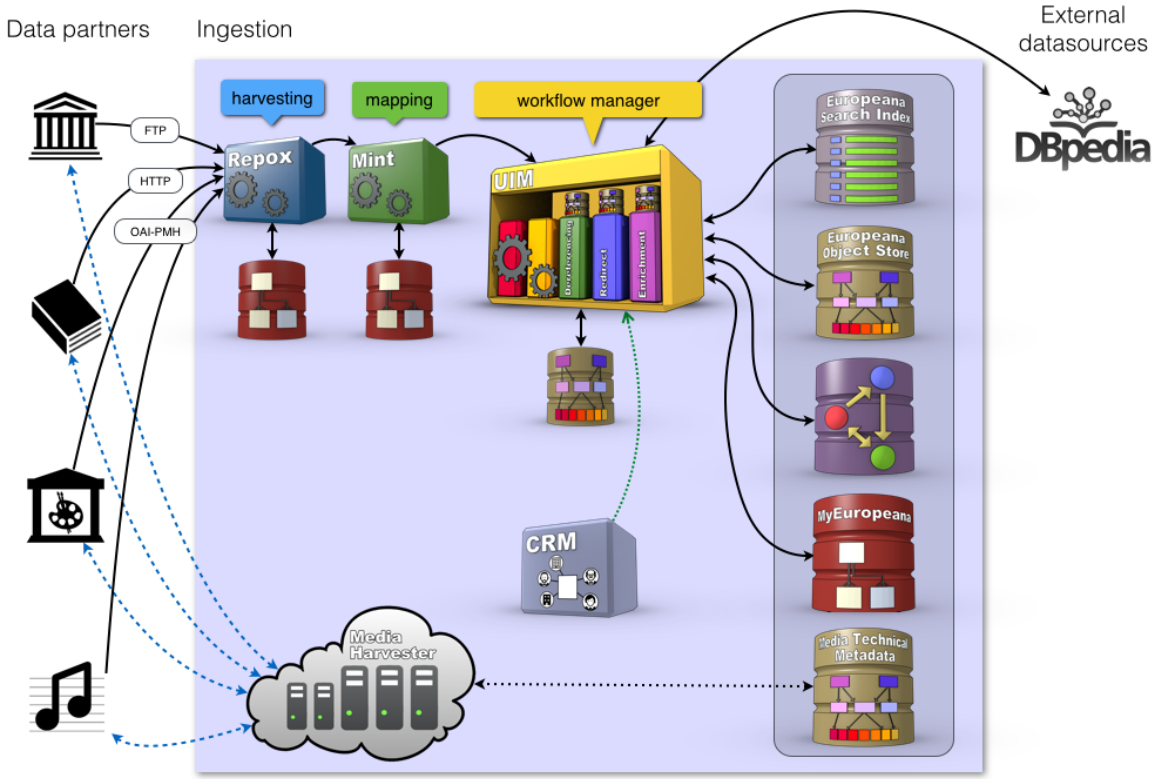

Diagram 2: Ingestion Process

CRM

Apart from cultural heritage data itself, Europeana stores information about organisations providing these data and about datasets into which this data is organised when it is consumed by Europeana. This information is stored in a SugarCRM instance, an open-source CRM solution. This instance of SugarCRM is contacted by UIM to obtain when information about organisations and datasets is necessary.

REPOX

Harvesting is the process of getting metadata records physically stored on Europeana servers. This is done using the open-source REPOX solution. REPOX can communicate with remote data repositories via a number of protocols, the ones primarily used are OAI-PMH, repository synchronisation protocol used widely in the archive domain, FTP, and HTTP.

MINT

Internally, Europeana stores metadata record in a variation of EDM called EDM- Internal. The tool used for mapping records to this format and ensuring their validity is MINT, developed by NTUA. MINT also analyses incoming data and generates statistics on the uniqueness of data fields and general data quality. There are two conceptual stages in MINT operation: first a mapping between the incoming metadata format and EDM-Internal is created; second, the records themselves are transformed using the mapping. Mappings can be re-used later.

UIM

The ingestion workflow of Europeana involves a number of steps performed either manually or automatically, orchestrated by the Unified Ingestion Manager (UIM) application framework. UIM's goal is to provide a unique and unified way of handling with the metadata: starting from the initial negotiation with the data provider through to actual generation of records and their storage as the last step before publication. Each step is implemented by a UIM plugin that communicates with a number of applications that constitute the full Europeana Ingestion framework. Below some characteristic plugins are described.

UIM Plugins: Dereferencing

Metadata records supplied by Europeana data providers often contain references to external data sources. To enhance the value of such fields for the discoverability of records, in some cases UIM connects to the external data sources, downloads from them relevant information, stores it locally for future references and embeds it into EDM records. This process is called *dereferencing* and is executed by the dereferencing plugin of UIM. Downloaded information is stored in a local MongoDB document database. To query external resources the plugin uses the SPARQL language (query language for RDF).

UIM Plugins: Record Redirect

Persistent identification of resources is a daunting task in the entire cultural sector. Knowing that, Europeana makes all the efforts to ensure that once a record is published, it will always be found under the same identifier. This task is made harder by the fact that Europeana's data partners themselves do not always guard the discipline of persistent identification. During its own lifetime, Europeana has evolved over various ingestion systems and keeping track of identifiers issued in the past is another technological challenge. The redirect plugin of UIM is responsible to address possible scenarios of changes in identifiers and to different identifiers being assigned to the same object. This information is used to redirect a request to an object even if one of its old identifiers is used.

UIM Plugins: Enrichment

Some fields of metadata records often contain well-known named entities, such as places, points in time, people or concepts. To increase the discoverability of such records by users, Europeana has developed its metadata enrichment process. Named entities are found in records using a set of heuristics. Contextual information, known about these entities from external data sources, is appended to the original record in the form of EDM contextual classes. This information currently includes multilingual representation of the identified contextual entity, as well as links to similar contextual resources. For the moment, contextual information is gathered from such external data sources as: Geonames, for geographical places; DBPedia for people and concepts; Gemet for concepts; and Semium for time periods. The processed output is stored locally in a MongoDB database.

The enrichment plugin, which implements the described process, is also responsible for creating the version of the record database and search index to be published later on the production environment.

Media Harvester

Under the Europeana Creative project, whose goal is to increase the potential of creative reuse of Europeana records, Europeana has introduced the Content Reuse Framework (CRF). The framework defines various degrees of reusability of media objects referred by Europeana and an important condition of reusability is the quality of these objects. To determine it, media objects are downloaded and analysed by Europeana and information about their quality (technical metadata) is stored next to the records themselves. The component responsible for this task is Media Harvester which is a heavy-duty distributed system for parallel processing of media files. It uses state-of-the-art open-source technologies for distributed computing.

Publication

The process of transferring the up-to-date content, produced by the ingestion process, to the production environment is called publication. For the moment, Europeana publishes its new content once a month, the cycle which is aligned with the needs of data providers and communicated to them regularly. During the process, the operation of UIM is paused, the content is finalised and transferred manually to the production environment. In the future, Europeana is planning to introduce the continuous publication process which will not require rigidly defined regular cycles but will allow more frequent updates of published content.

Production

Europeana's production environment contains the up-to-date Europeana dataset ready for consumption through various means. Practically, it is comprised of several artefacts produced previously by the ingestion process and transferred to production during the publication process. The dataset can be accessed in many ways: the web portal for end-users (Europeana portal), the API for developers, thematic projects, such as Europeana1914-1918 and Europeana Exhibitions, and two interfaces for specialised access.

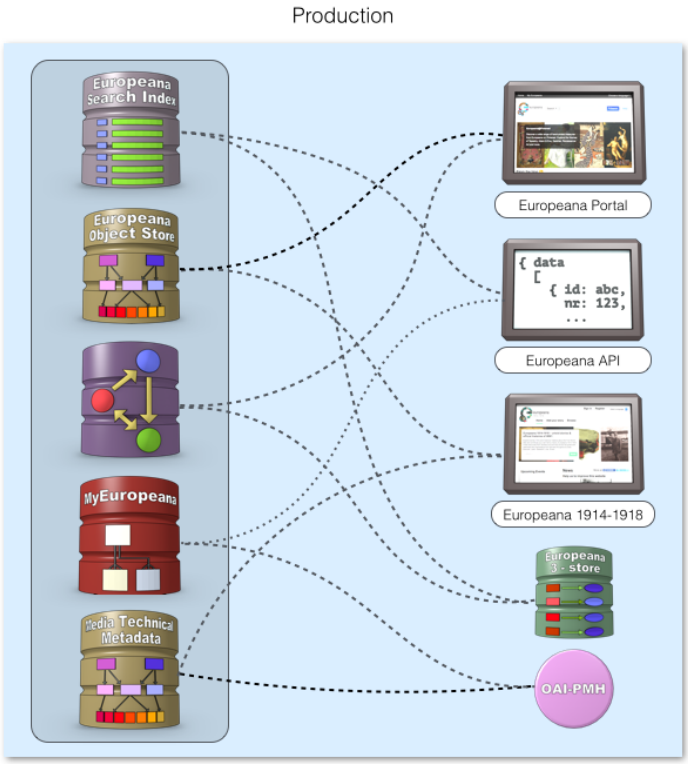

Diagram 2: Production Environment

The portal and API perform search and retrieval of EDM metadata records. The output formats supported by the API are JSON and EDM over XML. MyEuropeana is the private space a user can have at Europeana to tag favourite objects, annotate objects and save queries. Both the portal and the API support this functionality. Europeana1914-1918 and other thematic projects feature a subset of the entire dataset dedicated to a certain theme or event.

Under Europeana Creative project, we added two more ways to access the Europeana dataset. Through the Europeana OAI-PMH service the entire dataset of Europeana is available for download and synchronisation by applications using OAI-PMH, a popular

protocol in the archive and library domain. Europeana Linked Open Data exposes the Linked Data representation of the Europeana dataset. This service is useful for semantic researchers who can query it using the advanced SPARQL language for semantic queries. Both services will be completed and made operational in the course of 2015.

Hosting

Europeana uses services of different providers to host its environments.

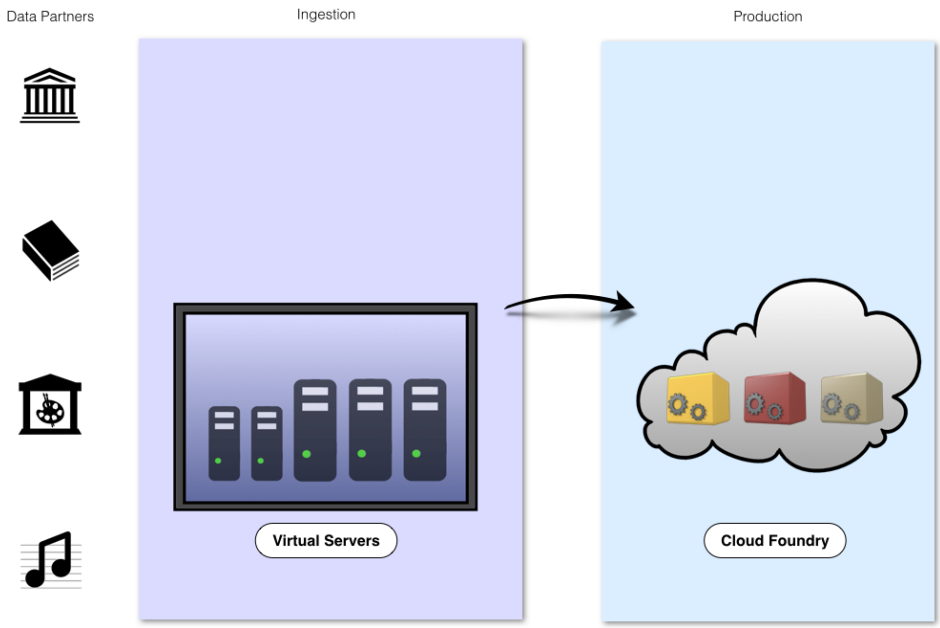

Diagram 4: Hosting Environments

The ingestion environment is hosted by ISTI, as part of our longtime partnership in the projects Europeana v2.0 and Europeana v3.0. This hosting is built of several high-performance servers and is using virtualisation technology to allocate virtual machines, also called sandboxes, for specific application needs.

The production environment is hosted using the innovative Platform-as-a-Service (PaaS) cloud approach. Another kind of virtualisation technology, it allows developers and operators to create applications directly on top of a hosting infrastructure, hiding away the underlying complexity. The software implementing this approach is Cloud Foundry, an open-source. Using PaaS was a strategic choice made by Europeana during the course of Europeana v3.0 to make itself fit for the future challenges as a network organisation delivering the European Digital Service Infrastructure for culture.

Summary

Europeana is the European Digital Library developing into a data and service infrastructure for the entire cultural sector. To deliver on this ambitious promise we need to employ a flexible technical architecture, able to serve our various needs. The deliverable above describes it from bird's eye view. The key motives in our architecture planning are

using mature open-source products able to scale to Europeana needs, state of the art hosting solutions and modular approach allowing to involve various partners and subcontracts for tasks requiring their narrow expertise.