D3.4 Final Technical & Logical Architecture and future work recommendations

Europeana v1.0

Deliverable number	*D3.4*
Dissemination level	*Public*
Delivery date	*June 16, 2010, Update October 5, 2011*
Status	*Final*
Author(s)	*Makx Dekkers, Stefan Gradmann, Jan Molendijk*

Grant Agreement Number: 558001

Table of Contents

1 Introduction

This deliverable has three tasks:

- To characterise the technical and logical architecture of Europeana as a system in its state at the end of the project (that is to say by the time of the 'Danube' release)
- To list the features under development or planned in Europeana V2.0, Assets and EuropeanaConnect.
- To outline the additional future work recommendations that can reasonably be made at that moment.

This also provides a straightforward and logical structure to the document: characterisation comes first followed by the plans and recommendations for future work.

2 Technical and Logical Architecture

2.1 Current situation

From an architectural point of view, Europeana.eu is best characterized as a search engine and a database. It loads metadata delivered by providers and aggregators into a database, and uses that database to allow users to search for cultural heritage objects, and to find links to those objects. Various methods of searching and browsing the objects are offered, including a simple and advanced search form, a timeline, a map Search, an openSearch API. The data is also made available as Linked Open Data.

It is also important to describe what Europeana.eu does not do, even though people sometimes expect it to. It does not store the actual digital objects. Only a thumbnail representation of the objects is cached locally. It does not (yet) index the content of those objects (e.g., the full text of digitized books), just the metadata.

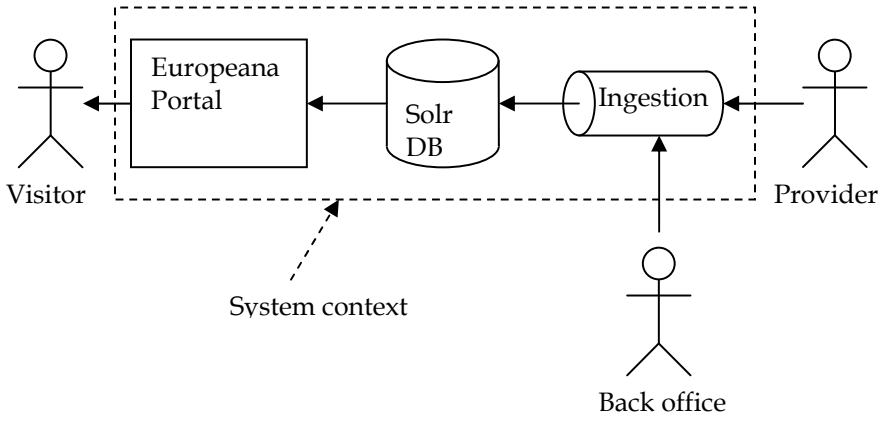
Because Europeana only has the metadata to work with, which is typically less than 50 words per object, brute force approaches to indexing, searching and providing multilingual access do not work very well. We have to use any structural information we can extract out of the metadata records our providers give us. Currently that structure is delivered to us in the form of ESE 3.3 records, which can be characterized as "Dublin Core plus a few project-specific elements".

The Europeana.eu ingestion process reads this structure and puts it in a Lucene/Solr search index. Lucene/Solr is strongly optimized to search through large datasets of both structured and unstructured information. Because it handles both structured and unstructured information equally well, we can implement searches base on specific fields (e.g. dc:title or dc:creator) as well as searches throughout the whole record, and still maintain a close control over the weighting of various fields in the search result etc. More traditional databases excel in either fielded or general search, but never both.

From a technical perspective the implementation has been highly optimized and modularized. Web servers, solr (database) servers and image servers all run on separate machines, allowing optimum configurations to be selected for each of these various functions. This brings both vertical and horizontal scalability, and allows for redundancy: the Europeana infrastructure is divided over two physical locations, each capable of functioning independently should the other fail.

Note however that we also offer an Open Source version of all Europeana software, allowing our software to be used by other institutions. That also means that the architecture was designed under the assumption that this separation is not a strict requirement. It is possible to run all processes on a single machine, and this may be an appropriate choice for a smaller library or museum that wants to run a cultural heritage object search engine for a medium sized collection.

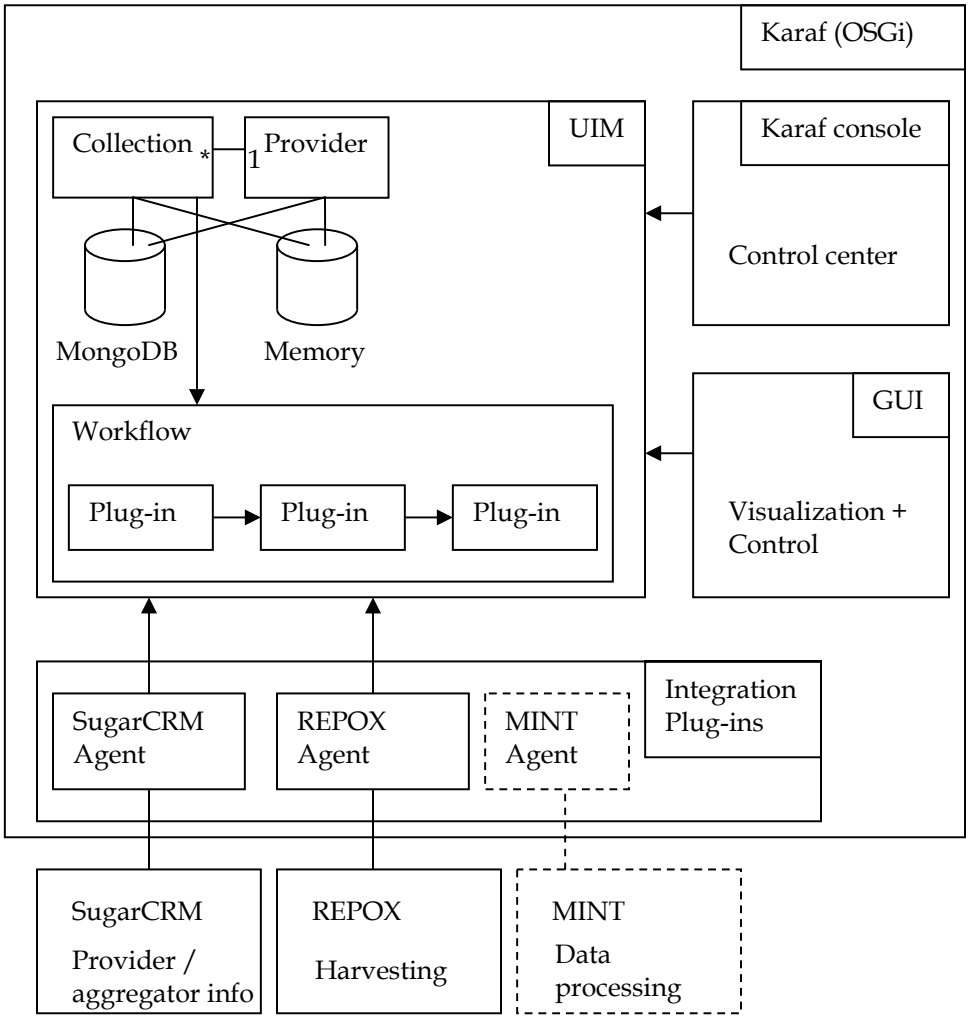
A closer look at the Europeana Architecture



Providers supply their metadata in datasets, the Europeana Back office (ingestion team) validates and enriches that data, which is

then indexed and stored in a lucene/solr search index. End users query the solr database through the Europeana Portal, or through the OpenSearch API.

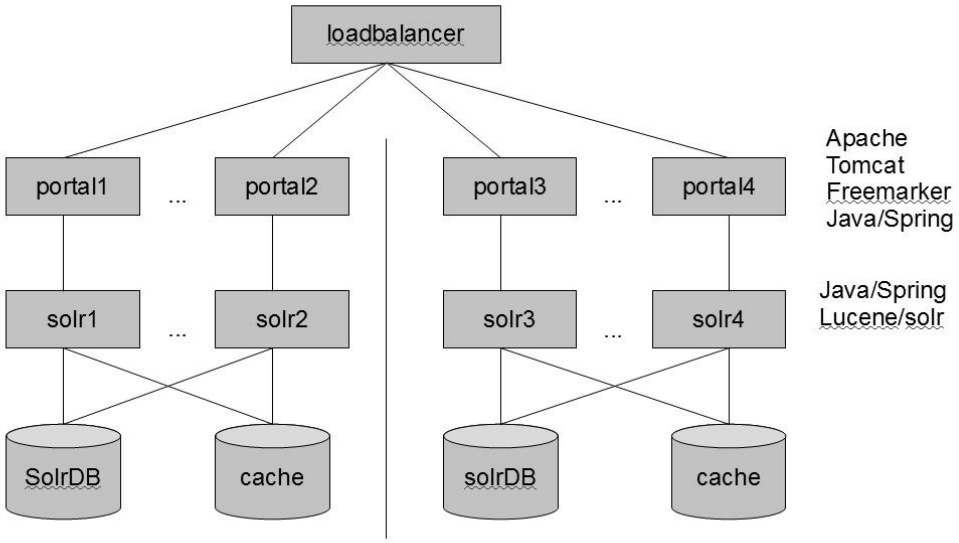
A more detailed look at the architecture of the ingestion process (Unified Ingestion Manager) shows some of the complexity involved:



Plug-ins perform the various validation, enrichment, image harvesting and indexing tasks. Plug-ins can be combined in various workflows for specific types of metadata. The UIM integrates with SugarCRM, REPOX and the MINT metadata mapping tool for more more specialized 'stand-alone' tasks such as OAI-PMH harvesting and managing provider and aggregator data. As we are adding more and more plugins, we are phasing out the old ingestion and publishing workflow consisting of standalone tools and discrete

steps in favour of using the workflow-based UIM.

The 'production system' is fully geared to high-availability, high-load continuous operations. It is a load-balanced redundant configuration, located in two separate data centers in Amsterdam and Almere. Hosting is handled through an SLA with Vancis, a private company with firm roots in Academia and therefore with excellent and relatively cheap connectivity directly to the Amsterdam Internet Exchange (AMS-IX). We use Linux as an Operating System. Apache, Tomcat and Freemarker are used to render the HTML pages. The Java code is based on the Spring framework and accesses the metadata that is stored and indexed in Lucene/solr. More information on the actual implementation, including the full source code, can be found on the EuropeanaLabs environment (www.europanalabs.eu).



Moving the European Data Model from theory to a practical implementation has proved challenging – think of it as changing the foundations of a theatre while the show is still going on. We have focussed initially on contextualizing 'who', 'where' and 'when' information and allowing data providers to supply context links for these entities. We have defined an XML schema for this first implementation of EDM, and some data providers have started to deliver according to that schema. We are currently hashing out

the issues they have encountered, so that EDM delivery may be more routinely integrated in the ingestion workflow.

2.2 Future work

Consequences of EDM

With the move to the new Europeana Data Model we are able to optimize this architecture even further, as we de-couple the ingestion process from the process of moulding the data into the structures we use to search and retrieve the data. We will also be better placed to do data enrichment and normalization.

With EDM the data providers will give us their original metadata (xml-ized if need be) together with a mapping file. The mapping file describes how the data should be mapped to EDM. Europeana stores the original metadata, and executes the mapping in the ingestion process. In that process various enrichment and normalization processes may be invoked, such as named entity recognition and linking to Geonames or VIAF records, normalization of date values etc. All enriched and normalized fields are stored in separate fields or aggregations, next to the original record. The ingestion process then again loads the mapped fields plus the enrichments and normalizations in the lucene/solr engine for indexing. At that time the indexing process will take a snapshot of the resources that the enrichment process has found links to: there is currently no system in place that would allow real-time expansion of these links to searchable data, at least not on the scale and diversity of data that Europeana offers. Links are preserved in this process, allowing the freedom to present the resources that are linked to, and potentially still allowing reasoning on these links. This is an interesting area of research that Europeana will need to consider while still maintaining an optimum search and retrieval experience for all current use cases.

Taking snapshots means that Europeana may have to consider regularly re-indexing all metadata to include any updated linked data. This is an added benefit of the chosen architecture: it allows Europeana to optimize the EDM data structures without having to go back to the original providers to ask them for resubmission of their data. In most cases a much simpler update of the mapping file plus a re-index will suffice.

Database / indexing engine

In the search architecture of Europeana Lucene/solr plays a major role. Even though it is a document index rather than a struc-

tured metadata database, it has served us well over the past years. This was helped by the fact the ESE data model is very flat: essentially the Europeana catalogue is a huge one-dimensional list of catalogue cards. With the move to EDM we add power of expression, and in doing so we also add complexity and more dimensions to the data – rather than one list of catalogue cards, we have an interconnected network of objects, resources, aggregations.

That means that we may run into limitations using the document index engine that is Lucene/solr.

So over the next year we will evaluate various approaches to indexing and storing the data:

- adapting our lucene/solr implementation by adding layers of caching and pre-computing search results
    - This is the approach that we have used to cope with the strong growth of traffic on the Europeana website, and based on initial experiments we expect it to bring enough performance for the first implementation of the EDM.
- relational (SQL) or document store (noSQL) databases.
    - Relational databases are good at maintaining relations between objects, which is one way of viewing the EDM model. However they also have limitations in terms of lack of flexibility, size of databases etc.
    - A noSQL database might works well in conjunction with a solr index: use the index to do fast searches on the data, then use the noSQL database to retrieve the document identified by the search. Let the document database handle the chore of storing millions of files, something that ios not easy to do with the Unix filesytem plus the tools such as rsync to handle that large volume of files.
- RDF triple stores
    - This class of 'database' is theoretically ideally suited to the EDM, but has yet to prove itself in a setting with the combination of our number of objects (we would require literally Billions of triples) and the load on our systems.

This evaluation is an interesting process in which we hope to involve also the creators/maintainers of the various indexing/data store systems: the Europeana dataset is large and of growing

complexity, and therefore potentially as interesting to them as it is to us.

The separation of portal, solr and image servers still applies and brings the same benefits in the EDM environment as it did in the ESE environment. We will however continue to monitor, evaluate and optimise performace as the Europeana database grows in number of objects, links to external resources and complexity.

Cloud computing

Cloud Computing, as a generic term, is a highly current topic. The technical evaluation that was done as part of the work in creating the New Renaissance report has as one of its recommendations that we look in to cloud computing as a way to implement Europeana services. If you look at how ´The Cloud´ is currently positioned we can see a number of potential uses of the Cloud for Europeana:

- IaaS – Infrastructure as a Service
    - quick deployment of new services: so far we have been able to support this by creating sandboxes on the Europeana development infrastructure. Services that are growing beyond the capacity of a sandbox are included in the managed infrastructure where the Europeana portal is hosted. Typical turnaround time for creating a sandbox is less than one day, making it a very flexible tool in our development environment
    - Flexibility/scalability: Already our hosting provider uses virtualization to a very high degree - most of the servers in the Europeana architecture are virtual. You could take that one step further and use not only virtual servers at the hosting provider, but also use server capacity on the cloud infrastructure providers such as Amazon. This could theoretically improve flexibility, allowing us to scale up (and down) quickly as needed. So far we have not found theneed to do this: the combination of sandboxes in our development
- PaaS – Platform as a Service
    - An example here would be to use Flickr as a storage solutions for User Generated Content projects – even though we are able to build a custom solution, we may opt to use a platform that already exists, working within the parameters of that platform. While this may not

bring us all the functionality we want, it is a lightweight way of introducing and experimenting with new services.

- SaaS – Software as a Service
  - Europeana Search Widget can in fact be viewed as an implementation of this Cloud concept: cultural heritage institutes can include a full Europeana search and display page on their site with minimal development effort.
  - Translation services: the current version of the Europeana portal uses web-based translation services provided by Google and Bing. This works very well, but has also already revealed a weakness of using web-services: you become dependent on the continued availability of 3rd party services, without the protection of a contract or ongoing business relationship.

In each case we will have to make a detailed cost/benefit analysis to determine whether Cloud alternatives are worth implementing.

Full text

Currently the Europeana search is based on metadata only. As more and more digitization projects produce full text and are willing to deliver that to Europeana (e.g., the public domain content that is part of the EuropeanaLibraries project). Integration of these full text resources in the Europeana gives the opportunity to do a much more fine-grained search and improve recall for many search queries. Maintaining precision in search is a challenge, especially when we are looking at several million objects.

Also balancing the results over several object types (texts, images, audio, video) is a real challenge: there is no obvious 'full-text' equivalent for the other content types, so why should they be penalized for the coincidence that text documents happen to be expressed in the same medium (text) in which we express the metadata. Having said that it would be a real shame to not use this rich resource, so we will experiment with ways to make a fair and balanced representation of the full text search results.

3 Work planned as Part of Europeana V2.0, Assets and EuropeanaConnect

The Europeana Cluster Advisory Board (ECAB) created as part of the ASSETS project has brought together the main projects providing technology to Europeana. In these discussions between

Europeana 1.0/2.0, ASSETS, EuropeanaConnect and Carare an number of synergetic areas were identified in areas such as

- EDM related issues,
- ingestion and enrichment,
- manual annotation tools,
- usability testing,
- 3D methods and
- processing of audio.

These areas will be monitored continuously and with the assistance of the ASSETS ECAB as long as this project is in place. However, this work will have to be continued after ASSETS funding has expired: good care should thus be taken to establish a succeeding structure!

We also need to determine the procedure for ingesting feedback from the monitoring process within the Europeana specification process.

4 Recommendations for future work

4.1 Introduction

In the description of work of Europeana v1.0, the Technology Watch is defined as an activity that will look at new developments and standards in the wider world and make recommendations on if, when and how they should be deployed in Europeana. In the previous period, the Technology Watch delivered a list of development, standards and vocabularies that were candidates for further study. In the first months of 2010, this approach has been augmented by the identification of a short list of items that support the future recommendations to be contained in D3.3 and D3.4.

The resulting draft then was presented to the experts meeting in Tirrenia on June 15 2010 and the opinions expressed during discussion have been integrated in the present document.

4.2 Overview of current issues

From work with the participants in WP3 and the development team in the Hague, the following items had originally been selected for further analysis.

1. FRBR/CRM harmonization: status and outlook. Extending the EDM to the FRBRoo model to take on board additional

librarian and museum aspects. The audiovisual community will benefit from such work, as well.

2. DBPedia: practical applications: Linking Europeana object representations to various Linked Open Data resources and namely to dbPedia.
3. DDC, OCLC strategy on use in linked data. Explore the systematic use of DDC as contextualisation resource also considering its pivotal potential regarding multilingual operations
4. Enable support for deductive or inductive scholarly reasoning.
5. Authentication/Identification: SAML, Shibboleth. Provide an open, standards based authorization and authentication framework based on standard components that need not be maintained (at least not entirely) by Europeana staff (OpenID and SAML based frameworks such as Shibboleth may be relevant here).

Activities 1) and 2) need not be further detailed, as they are currently already worked at (in the case of 2) or at least taken on board in WP7 of EuropeanaV2.0 (in the case of 1).

Furthermore, a need to further develop the EDM in order to enable the expression of provenance and versioning information referring to aggregations as a whole has been identified as a major issue in the meanwhile. Additionally, some advanced support for migrating legacy data to the EDM will be needed going somewhat beyond the MINT tool developed by NTUA[1] and especially for migrating data from existing library automation systems.

Finally, activity 4) has grown into a separate strand of activities targeting the so-called 'Digital Humanities' (cf. more details below in section 4.7).

4.3 EDM-FRBRoo harmonization

The relevant wikipedia article[2] makes the following statement on FRBR:

> Functional Requirements for Bibliographic Records—or FRBR, sometimes pronounced /□f□rb□r/—is a conceptual entity-relationship model developed by the International Federation

1 http://mint.image.ece.ntua.gr/

2 http://en.wikipedia.org/wiki/Functional_Requirements_for_Bibliographic_Records, 21 May 2010

of Library Associations and Institutions (IFLA) that relates user tasks of retrieval and access in online library catalogues and bibliographic databases from a user's perspective. It represents a more holistic approach to retrieval and access as the relationships between the entities provide links to navigate through the hierarchy of relationships.

FRBR comprises groups of entities:

- Group 1 entities are Work, Expression, Manifestation, and Item (WEMI). They represent the products of intellectual or artistic endeavour.
- Group 2 entities - covered by the Functional Requirements for Authority Data (FRAD) specifications - are person and corporate body, responsible for the custodianship of Group 1's intellectual or artistic endeavour.
- Group 3 entities covered by the Functional Requirements for Subject Authority Data specifications - are subjects of Group 1 or Group 2's intellectual endeavour, and include concepts, objects, events, places.

Group 1 entities are the foundation of the FRBR model:

Work is a "distinct intellectual or artistic creation." (IFLA 1998)

Expression is "the specific intellectual or artistic form that a work takes each time it is 'realized.'" (IFLA 1998)

Manifestation is "the physical embodiment of an expression of a work. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form." (IFLA 1998)

Item is "a single exemplar of a manifestation. The entity defined as item is a concrete entity." (IFLA 1998)

A related activity is FRBRoo, which is described in wikipedia[3] as follows:

The FRBRoo (FRBR-object oriented) initiative is a joint effort the CIDOC Conceptual Reference Model and Functional Requirements for Bibliographic Records international working groups to establish "a formal ontology intended to capture

[3] http://en.wikipedia.org/wiki/FRBRoo, 21 May 2010

and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information."

The idea behind this initiative is that both the library and museum communities would benefit from harmonizing the FRBR and CIDOC reference models to better share library and museum information, particularly in light of the Semantic Web and the overall need to improve the interoperability of digital libraries and museum information management systems. This led to the formation of the International Working Group on FRBR/CIDOC CRM Harmonisation in 2003 with the common goals of "expressing the IFLA FRBR reference model with the concepts, tools, mechanisms, and notation conventions provided by the CIDOC CRM…and aligning (possibly even merging) the two object-oriented models with the aim to contribute to the solution of the problem of semantic interoperability between the documentation structures used for library and museum information."

The first draft of FRBRoo was completed in 2006. It is a logically rigid model interpreting conceptualizations expressed in FRBRer [FRBR-entity relationship] and of concepts necessary to explain the intended meaning of all FRBRer attributes and relationships. The model is formulated as an extension of the CIDOC CRM. Any conflicts occurring in the harmonization process with the CIDOC CRM have been or will be resolved on the CIDOC CRM side as well. The Harmonization Group intends to continue work modelling the FRAR concepts and elaborating the application of FRBR concepts to performing arts.

A presentation by Vinod Chachra of VTLS at a TELplus FRBR workshop at the National Library of Portugal on 9 October 2008[4], outlined two ways of using FRBR: one to keep the data as they are and expose FRBRised records on the fly; the second to convert the catalogue to contain separate records for the work, expression, manifestation and item. At the same workshop, Janifer Gatenby of OCLC presented the activities of OCLC on FRBR[5],

[4] http://frbr.bnportugal.pt/ documentos/The_vision_of_software_vendor.ppt

[5] http://frbr.bnportugal.pt/documentos/The_activities_of_OCLC_on_FRBR.ppt

highlighting that OCLC WorldCat has been "FRBRised" with 110 million records representing 85 million works.

An article by Jenn Riley, Caitlin Hunter, Chris Colvard, and Alex Berry of the Indiana University Variations3 project, "Definition of a FRBR-based Metadata Model for the Indiana University Variations3 Project"[6] , an example is given of a FRBR representation of a CD with two concerts.

Relevance for Europeana:

The distinction between the work, expression, manifestation and item[7] will be relevant for the resources that are aggregated in Europeana. Functionality may be required to group results under the work level (e.g. all copies of all digital files in any format that contain all performances of a composition), under the expression level (e.g. all digital files in any format of a particular performance of a composition), or under the manifestation level (e.g. all digital files in a particular format of a particular performance of a composition).

The framework of reference however should not so much be the original FRBR specification (which still is too much depending on the notion of a bibliographic record!) but rather the FRBRoo approach, in which each aggregate is treated as an entity in its own right. Note that a mapping of FRBRoo and EDM is offered by CIDOC.

As a consequence, Subtask 7.3.3 (Model refinements for EDM) has been specified as part of Europeana V2.0.

References:

- Vinod Chachra. The Two Worlds of FRBR. 2008. http://frbr.bnportugal.pt/documentos/The_vision_of_software_vendor.ppt
- Talat Chaudhri. Assessing FRBR in Dublin Core Application Profiles. 2009. http://www.ariadne.ac.uk/issue58/chaudhri/
- Martin Doerr, Patrick Le Boeuf. FRBRoo introduction. 2009. http://cidoc.ics.forth.gr/frbr_inro.html

6 http://www.dlib.indiana.edu/projects/variations3/docs/v3FRBRreport.pdf

7 As well illustrated in Tiina Ison's presentation on the ACERBI project available at http://www.stks.fi/files/Ison.pdf and which is currently evolving into a formal publication.

- Chryssoula Bekiari, Martin Doerr, Patrick Le Boeuf (eds.). FRBR object-oriented definition and mapping to FRBR$_{ER}$ (version 1.0). 2009. http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBRoo_V1.0_2009_june_.pdf
- Janifer Gatenby. OCLC and FRBR. 2008. http://frbr.bnportugal.pt/documentos/The_activities_of_OCLC_on_FRBR.ppt
- Corey A. Harper. Linked Library Data and the Semantic Web. 2008. http://www.kb.se/dokument/Bibliotek/utbildning/presentationer/20080917Harpey-rev.pdf
- Corey A. Harper. Linking Library Data. 2009. http://www.lyrasis.org/Classes-and-Events/~/media/Files/Lyrasis/Classes%20and%20Events/charper%20lyrasis%2020091113.ashx
- Jenn Riley. Moving from a locally-developed data model to a standard conceptual model. 2008. http://www.dlib.indiana.edu/~jenlrile/presentations/isko2008/isko2008.ppt
- Jenn Riley, Caitlin Hunter, Chris Colvard, and Alex Berry. Definition of a FRBR-based Metadata Model for the Indiana University Variations3 Project. 2007. http://www.dlib.indiana.edu/projects/variations3/docs/v3FRBRreport.pdf
- Yin Zhang, Athena Salaba. Major Issues Facing FRBR Research and Practice Identified in a Delphi Study. Undated. http://frbr.slis.kent.edu/publications/delphi_issues.pdf
- Maja Žumer. Some outcomes of the CRM/FRBR harmonization: the definition of manifestation and a review of attributes. 2005. http://www.oclc.org/research/activities/past/orprojects/frbr/frbr-workshop/presentations/zumer/Manifestation_and_attributes.ppt
- Tiina Ison. Contextualizing the Extraction of Meaning from an Old Book into Distributed Digitization Production Processes. Presentation at the Suomen Ranskan instituutti March 18, 2011. http://www.stks.fi/files/Ison.pdf

4.4 Versioning and Provenance of Europeana Aggregations

A number of statements on the representations of objects in Europeana cannot reasonably be made referring to individual RDF triples but rather need to address aggregations of triples as a whole. This concerns essentially two types of statements absolutely vital for the take-up of Europeana as a scholarly working environment dealing with intellectual property in a responsible and efficient way:

Versioning

Unless the state of Europeana aggregations can be tracked back in time Europeana might not be adopted as a serious source for scholarly work in at least parts of the the Digital Humanities (cf. infra). This is not limited to the state of individual aggregations but also needs to take into account the linking context of such aggregations: it must be possible to tell, for instance, which aggregation was linked to which others at a given moment in time. The same applies to user supplied content (annotations and the like) pertaining to aggregations as a whole. Statements of this type cannot reasonably be made pertaining to each individual constituent triple but need to refer to an aggregation as a whole.

Provenance

Statements on provenance will be required by many content providers – and be it only to enable the identification of their contribution to Europeana. This doesn't relate so much to the content holding institutions – which are just the custodians, but not the producers of the content – but rather to the original creators of the original content item. We should clearly distinguish those from the creators of the digital representation ingested in Europeana. But here again, such statements make little sense applied to individual triples: the guiding principle of Linked Open Data is to reuse statements wherever possible instead of creating new ones. Intellectual property in Europeana therefore cannot be conceived on triple level but rather on aggregation level.

Fortunately, the use of the resource map and proxy features of the ORE specifications might provide an important building block of a solution to this issue. The EDM specifications thus need to be extended with information on how we could use ORE resource maps to enable versioning and provenance statements on aggregations as a whole should this be the appropriate way to go. Otherwise, another appropriate solution would have to be found.

RDF "named graphs" (or "quadruples") may also provide with an appropriate solution, when they become fully standardized.

However, it probably will not be possible to solve this issue on a purely technical basis: some thinking and agreements relating to the division of tasks among Europeana, the aggregators, the digitising custodians and the content producers in an overall workflow perspective will ultimately be required (cf. section 5 of this document)!

Relevance for Europeana
The issue is of high strategic importance for Europeana: provenance information clearly has strategic importance as for instance in rights clearing settings related to the data provider's agreement and versioning is key for the acceptance of our services by scientific communities and. Not dealing with both appropriately might seriously affect the overall acceptance of our endeavor both in technical and in business planning terms.

References:

- Herbert Van de Sompel, Carl Lagoze, Jeroen Bekaert, Xiaoming Liu, Sandy Payette, Simeon Warner: An Interoperable Fabric for Scholarly Value Chains. D-Lib Magazine Volume 12 Number 10, October 2006
- ORE Specifications and User Guides. http://www.openarchives.org/ore/1.0/toc.html
- Herbert van de Sompel, Michael Nelson, Rob Sanderson: HTTP framework for time-based access to resource states – Memento. April 2011. http://www.mementoweb.org/guide/rfc/ID/
- Memento Guide: Introduction. http://www.mementoweb.org/guide/quick-intro/

4.5 Linked Open Data Integration and Linking to DBPedia

DBpedia as the most prominent Linked Open Data resource give the following description[8] regarding their own activities:

> DBpedia is a project aiming to extract structured information from the information created as part of the Wikipedia project. This structured information is then made available on the World Wide Web. DBpedia allows users to query relationships and properties associated with Wikipedia resources, including links to other related datasets.
>
> […]

[8] http://en.wikipedia.org/wiki/DBpedia, 22 May 2010

> The dataset is interlinked on RDF level with various other Open Data datasets on the Web. This enables applications to enrich DBpedia data with data from these datasets. As of April 2010, there are more than 4.9 million interlinks between DBpedia and external datasets including: Freebase, OpenCyc, UMBEL, GeoNames, Musicbrainz, CIA World Fact Book, DBLP, Project Gutenberg, DBtune Jamendo, Eurostat, Uniprot, Bio2RDF, and US Census data. The Thomson Reuters initiative OpenCalais, the Linked Open Data project of the New York Times, and the Zemanta API also include links to DBpedia. The BBC uses DBpedia to help organize its content. Faviki uses DBpedia for semantic tagging. Amazon provides DBpedia Public Data Set that can be integrated into Amazon Web Services applications.

And further figures extracted from the same web presence read as follows:

> The DBpedia project extracts various kinds of structured information from Wikipedia editions in 92 languages and combines this information into a huge, cross-domain knowledge base.
>
> DBpedia uses the Resource Description Framework (RDF) as a flexible data model for representing extracted information and for publishing it on the Web. We use the SPARQL query language to query this data. Please refer to the Developers Guide to Semantic Web Toolkits to find a development toolkit in your preferred programming language to process DBpedia data.
>
> The DBpedia knowledge base currently describes more than 3.4 million things, out of which 1.5 million are classified in a consistent Ontology, including 312,000 persons, 413,000 places (including 310,000 populated places), 94,000 music albums, 49,000 films, 15,000 video games, 140,000 organizations (including 31,000 companies and 31,000 educational institutions), 146,000 species and 4,600 diseases. The DBpedia data set features labels and abstracts for these 3.2 million things in up to 92 different languages; 841,000 links to images and 5,081,000 links to external web pages; 9,393,000 external links into other RDF datasets, 565,000 Wikipedia categories, and 75,000 YAGO categories. The DBpedia knowledge base altogether consists of over 1 billion pieces of information (RDF triples) out of which 257 million

were extracted from the English edition of Wikipedia and 766 million were extracted from other language editions.

DBPedia usually has two URIs associated with an entity, for example http://dbpedia.org/resource/Paris for the "non-information resource" (the real-world entity, the city of Paris) and the description about that entity http://dbpedia.org/page/Paris.

Practical usage:

Tools like OpenCalais or Luxid (from Temis) use DBPedia (and additional sources like GeoNames, the Internet Movie Database IMDB and VIAF) to derive URIs to be used in metadata, thereby making it possible to unambiguously refer to entities and provide additional information about those. It may be useful to also make use of WordNet (in spite of the lack of a coherent notion of term identity) as 'glue' between vocabularies.

It would be important, in this respect, to include the Getty thesauri (AAT and others) as linked open data in this list, as they have been key resources for our work up to now. Martin Doerr / CIDOC will establish communication with Getty in this respect.

Relevance for Europeana:

To support the objective to build semantic networks around the cultural heritage resources accessible through Europeana's portal and API, there is a strong requirement to use unambiguous references to these resources. Using DBPedia URIs is one practical option to realise this. Another option is to make Europeana by itself a provider of reference URIs for cultural heritage objects.

It needs to be noted though that there are two issues related to referencing resources:

1. Persistent identification: for any service that aims to have a long-term existence, like Europeana, it is important to base itself on persistent identifiers, i.e. identifiers that will be both unambiguous (the identifier will identify only one thing) and stable (the identifier will always refer to the same thing). Neither DBPedia nor its main source Wikipedia have explicit persistence policies. Furthermore, we should distinguish two usage scenarios in this respect:
    a. Europeana "consumes" resources, for which it encourages its providers to use PIs (especially the digitized material, and the object pages on provider's site). A tool like DSNotify could be useful to palliate issues in the sources Europeana depends on.

b. Europeana provides resources, which should be as persistent as possible for others to "consume" them.

2. Co-referencing: DBPedia is just one of a number of services that provide URIs for real-world entities. For example, for people, there is VIAF. As an example, Johann Wolfgang von Goethe can be referred to with the URL http://www.viaf.org/viaf/24602065/, http://dbpedia.org/resource/Johann_Wolfgang_von_Goethe , while in addition, organisations and people may coin their own URI (e.g. http://purl.org/dc/aboutdcmi#DCMI). In general, in the Semantic Web one entity can have many identifiers, and practical approaches to equate the various URIs for the same thing need to be found.

Besides, similarity, full content search and content summarising techniques need to be considered in this context.

Needless to say in this context that all this will only be possible if Europeana does operate on the original provided digital cultural heritage object at least once, at ingest time, in order to extract from it what the Europeana Outline Functional Specification document (D2.5) had called abstractions and which is referred to above (under 2.2) as "full text".

Furthermore, we may need to establish a clear policy (in terms of a recommendation) as to what are the preferred linking targets per category of linking options. In the case of person names, for instance, we would need to decide whether the linking target with the highest preference by default would be VIAF or FOAF. Such a 'semantic policy' cannot be 100% prescriptive and will probably be restricted to a set of rules expressed as rather general statements – but still it could be valuable for creating some basic homogeneity within the Europeana linking practice.

Finally, it will be crucial to make sure that Europeana itself will integrate in the linked open data paradigm and thus be available as a contextualisation resource for others without restrictions. Not meeting this objective would seriously affect our credibility in the Linked Open Data community.

As a consequence, two subtasks (7.3.1: Tools for semantic extraction and 7.3.2: Social Semantic Web) have been specified as part of Europeana V2.0 WP7), work on a Linked Open Data Pilot is far advanced and linking with DBPedia and other Linked Data resources well under way as can be seen at http://europeana.eu/portal/search.html?start=109&query=enrichment_agent_term%3A* .

References:

- About DBPedia. http://dbpedia.org/About
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Søren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. DBpedia - A Crystallization Point for the Web of Data. 2009. http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-etal-DBpedia-CrystallizationPoint-JWS-Preprint.pdf
- Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. 2009. http://www.georgikobilarov.com/publications/2009/eswc2009-bbc-dbpedia.pdf
- Tom Heath, Christian Bizer. Linked Data: Evolving the Web into a Global Data Space. 2011.

4.6 Use of DDC as contextualisation resource

DDC is described in wikipedia as follows[9]:

> The Dewey Decimal Classification (DDC, also called the Dewey Decimal System) is a proprietary system of library classification developed by Melvil Dewey in 1876; it has been greatly modified and expanded through 22 major revisions, the most recent in 2003. This system organizes books on library shelves in a specific and repeatable order that makes it easy to find any book and return it to its proper place. The system is used in 200,000 libraries in at least 135 countries.
>
> The DDC attempts to organize all knowledge into ten main classes. The ten main classes are each further subdivided into ten divisions, and each division into ten sections, giving ten main classes, 100 divisions and 1000 sections. DDC's advantage in using decimals for its categories allows it to be both purely numerical and infinitely hierarchical. It also uses some aspects of a faceted classification scheme, combining elements from different parts of the structure to construct a number representing the subject content (often combining two subject elements with linking numbers and geographical

[9] http://en.wikipedia.org/wiki/Dewey_Decimal_Classification, 22 May 2010

and temporal elements) and form of an item rather than drawing upon a list containing each class and its meaning.

The DDC is owned by OCLC and usage is subject to an annual subscription that is currently focused on use by library staff. It is not yet clear what OCLC's policies are with respect to offer Dewey as a tool for Linked Data. Summaries of the first three levels (the ten main classes, the hundreds divisions and the thousands sections) can be found at http://www.oclc.org/dewey/resources/summaries/. and at dewey.info for the linked data version of these classes.

Relevance for Europeana:

For Europeana, the use of a common classification scheme for cultural heritage resources would be a useful contribution to faceted searching on subject. However, this can only be done if the use of such a classification in an online environment with millions of items is not prohibited or prohibitively expensive.

Open issues include the following:

- It remains to be investigated to what extent DDC is actually used (and relevant) outside the library community.
- Furthermore, the DDC – LCSH mapping done by OCLC is relevant, and there are more mappings for direct reuse (such as from the CrissCross project).
- It remains to be determined whether the top 1000 classes currently available as linked open data are actually sufficient?
- We should investigate the option of harmonizing upper level domain thesauri linking these to (potentially) DDC and other resources and eventually blend the upper levels of DDC, AAT & CRM.
- We need to find out how to use LoD resources in GUI terms (cf. work done by Douglas Tudhope).

4.7 Enable Support for Scholarly Inferencing

We should evolve Europeana into a scholarly source environment enabling knowledge generation and capable of producing digital heuristics. In this respect, support for reasoning and inferencing is a key issue, but it remains to be determined what kind of inferencing is required: can we build on RDFS? Or do we need more and thus have to consider using OWL (and if so: which version, which profile)? The results of the meeting in Paris (April

2011) referred to below seem to indicate that a lightweight approach may be sufficient.

The wikipedia entry on OWL reads as follows:[10]

> The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies endorsed by the World Wide Web Consortium. They are characterised by formal semantics and RDF/XML-based serializations for the Semantic Web. OWL has attracted both academic, medical and commercial interest.
>
> In October 2007, a new W3C working group was started to extend OWL with several new features as proposed in the OWL 1.1 member submission. This new version, called OWL 2, soon found its way into semantic editors such as Protégé and semantic reasoners such as Pellet, RacerPro and FaCT++. W3C published the new version on 27 October 2009.

Relevance for Europeana:

As Europeana aims to be able to implement a certain level of reasoning over the data it manages, certain OWL properties (for value and cardinality constraints, class axioms and properties concerning individuals such as owl:sameAs) should be relevant to enable this reasoning.

In dealing with this issue it is essential for Europeana to co-operate with DARIAH and the rest of the Digital Humanities community.

The core issue is dealing with uncertainty (probability and the like).

Needs of the Digital Humanities Community have been given a closer look during a one day meeting in Paris (TGE Adonis) on April 4 2011, and the clearly dominant view was that Europeana was to provide stable resources, identified by a URI, as well as a clearly defined API – but that specialised reasoning would not have to be supported by Europeana, as this would rather take place in the Digital Humanist's specialised platforms. Smooth interaction between Europeana and these platforms thus is the core issue in this perspective and basic RDF/RDFS support may be sufficient! However, the actual consequences of this scenario need to be experimented in a prototype environment combining EDM data and the original provided digital heritage objects to es-

[10] http://en.wikipedia.org/wiki/Web_Ontology_Language, 22 May 2010

tablish strategic impact to the Europeana architecture and workflow design.

References:

- W3C OWL Working Group (eds.). OWL 2 Web Ontology Language Document Overview. 2009. http://www.w3.org/TR/owl2-overview/
- Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, Ulrike Sattler. OWL2: The Next Step forOWL. 2008. http://www.comlab.ox.ac.uk/ian.horrocks/Publications/download/2008/CHMP+08.pdf
- Stefan Decker. Who the hell needs description logics anyway? 2008. http://carbon.videolectures.net/2008/active/iswc08_karlsruhe/panel_schneider_owl/iswc08_panel_schneider_owl_01.pdf
- Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph. Knowledge Representation for the Semantic Web Part I: OWL 2. http://semantic-web-book.org/w/images/b/b0/KI09-OWL-Rules-1.pdf
- Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue, Carsten Lutz. OWL 2 Web Ontology Language Profiles. 2009. http://www.w3.org/TR/owl2-profiles/
- Nick Drummond, Matthew Horridge. A Practical Introduction to Ontologies & OWL. 2005. http://www.co-ode.org/resources/tutorials/intro/slides/ProtegeOWLPart2-v05.ppt

4.8 Authentication and authorization

The JISC Identity Management Toolkit gives the following description of identity management and related technology[11]:

> Identity management, in a general sense, includes all the processes and systems that allow the creation, retrieval, update, verification and destruction of identities and information relating to identities including any rights / authority granted to the identities. It is important to note that identities have been, and continue to be, managed using paper-based systems operated by people. In addition, many IT

[11] https://gabriel.lse.ac.uk/twiki/bin/view/Projects/IdMToolkit/Toolkit, 22 May 2010

based identity management systems are used to create artifacts (e.g. identity cards) which may be subject to visual checks and/or machine-based verification.

Identity management in computing involves the mapping of real world identities to electronic identities and ensures appropriate use of IT systems.

JISC in the UK decided to implement Shibboleth as the architecture that enables organisations to build single sign-on environments that allow users to access Web-based resources using a single login.

Shibboleth in turn is described by its designers as follows[12]:

The Shibboleth® System is a standards based, open source software package for web single sign-on across or within organizational boundaries. It allows sites to make informed authorization decisions for individual access of protected online resources in a privacy-preserving manner.

The Shibboleth software implements widely used federated identity standards, principally OASIS' Security Assertion Markup Language (SAML), to provide a federated single sign-on and attribute exchange framework. Shibboleth also provides extended privacy functionality allowing the browser user and their home site to control the attributes released to each application. Using Shibboleth-enabled access simplifies management of identity and permissions for organizations supporting users and applications. Shibboleth is developed in an open and participatory environment, is freely available, and is released under the Apache Software License.

What is Shibboleth and how does it work?

A user authenticates with his or her organizational credentials. The organization (or other identity provider such as Google, Yahoo, Facebook or OpenID) passes the minimal identity information necessary to the service manager to enable an authorization decision.

There are two primary parts to the Shibboleth system:

1. Identity Provider - the software run by an organization with users wishing to access a restricted service;
2. Service Provider - the software run by the provider managing the restricted service.

[12] https://gabriel.lse.ac.uk/twiki/bin/view/Projects/IdMToolkit/Toolkit, 22 May 2010

Shibboleth leverages the organization's identity and access management system, so that the individual's relationship with the institution determines access rights to services that are hosted both on- and off-campus. For a series of technical explanations of how Shibboleth works, from easy to expert, refer to the SWITCH Federation site.

Relevance for Europeana:

In a distributed system with potentially millions of users, the handling of authentication and authorisation is a crucial aspect to make sure that access to resources is properly managed.

Work in this area should be conducted in co-operation with TERENA and JISC.

References:

- JISC. The Identity Management Toolkit Project. 2010. https://gabriel.lse.ac.uk/twiki/bin/view/Projects/IdMToolkit/WebHome
- Architecture for a Shibboleth-Protected iRODS System. http://www.jisc.ac.uk/whatwedo/programmes/einfrastructure/aspis.aspx
- Shibboleth Access to Resources on the National Grid Service. http://www.jisc.ac.uk/whatwedo/programmes/einfrastructure/sarongs.aspx
- Grouper to Support Federated Identity for Virtual Organisations. http://www.jisc.ac.uk/whatwedo/programmes/einfrastructure/gfivo.aspx
- Cardiff University Collaboration with KC-ROLO Organisational Objects. http://www.jisc.ac.uk/whatwedo/programmes/einfrastructure/cuckoo.aspx
- Shibboleth Technical Reading List. https://gabriel.lse.ac.uk/twiki/bin/view/Projects/InitialReadingList
- UK Access Management Federation for Education and Research. http://www.ukfederation.org.uk/
- Trans-European Research and Education Networking Association, http://www.terena.org/

5 Future Evolution and Revisions of Europeana Architecture

The items above are considered to have primary importance for the future developments of Europeana and more precisely affect the releases directly succeeding Danube.

This includes a mapping, matching data values & data ingestion working environment (workflow design and implementation). Some of this (GUI) is defined in ASSETS, some of it is part of the MINT tool. We should be careful to include tools, organization and communication aspects in a holistic approach.

Furthermore – and this may evolve into a strategic discussion – embedding Europeana increasingly in a Linked Open Data architecture may lead us to reconsider our data architecture in more fundamental terms: more specifically, this concerns the issue of centralised vs. distributed storage and processing of data and information once we've completed the move to RDF based operations!

In fact, we may also wish to reconsider the way Europeana, aggregators and data providers interact in technical terms: for the time being all of these actors operate within separate, autonomous workflow environments of their own and organise data streams between their storage environments: an overall expensive way of working which is highly redundant and far from efficient the way it could be once we move to truly distributed approaches. The wish to reconsider this aspect of Europeana's work may in the end simply be triggered by political questions from the funding bodies …

And finally, key recommendation 4 of section 6 in the "New Renaissance" issued by the comité des sages reads:

> For the medium term, it should be considered to give Europeana a key role in the preservation of Europe's heritage and to turn it into a European deposit site for public domain digitised cultural material and into a dark archive for in-copyright cultural material, both digitised and born digital.

To comply with this recommendation and to thus hold metadata aggregations together with the digitised / born digital cultural objects will definitely lead to another thorough step in the evolution of Europeana's data architecture – although this does not necessarily imply pulling together all digitised content in one place: such a future evolution could well into a 'yellow pages'-like directory service for preserved digital objects similar to the BHL scan

list (cf. http://www.europeana.eu/portal/thoughtlab_digitisation.html)

6 Acknowledgements

This document is the result of intense discussions among the members in WP3 of the Europeana V1.0 project – they are too numerous to be named individually.

Some members of the WP3 Core Expert group had a significant share all along this process: among these, Emmanuelle Bermès (BPI, Paris), Robina Claypham (Europeana), Martin Doerr (ICS Forth), Michael Fingerhut (Bibliomus), Antoine Isaac (Europeana/VU Amsterdam), Carlo Meghini (CNR ISTI), Daniel Pitti (University of Virginia), Sjoerd Siebinga (Delving), Vassilis Tzouvaras (NTU Athens), Herbert Van de Sompel (Los Alamos National Laboratory) and Theo van Veen (KB The Hague) deserve special thanks.

Special thanks as well to Tiina Ison (National Library of Finland, Helsinki) for substantial input in the finalising phase of this document.