ECP 2008 DILI 558001

Europeana v1.0

D3.3 Initial Technical & Logical Architecture and future work recommendations

Deliverable number	*D3.3*
Dissemination level	*Public*
Delivery date	*30 July 2010*
Status	*Final*
Author(s)	*Makx Dekkers, Stefan Gradmann, Jan Molendijk*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

1. Introduction

This deliverable has two tasks:

- To characterise the technical and logical architecture of Europeana as a system in its 1.0 state (that is to say by the time of the 'Rhine' release
- To outline the 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 recommendations for future work.

2. Technical and Logical Architecture

From a high-level 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 an advanced search form, a timeline, and an openSearch API.

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 thumbnail representations of the objects are cached locally. It does not (yet) index the content of those objects (e.g., the full text of digitized books), just the metadata.

Apart from that, Europeana is to become a platform for knowledge generation building on rich contextualisation and semantic inferencing – but this is yet to come starting with the 'Danube' release and not part of the initial technical and logical architecture.

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 to 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 optimised to search through large datasets of both structured and unstructured information. It handles both structured and unstructured information equally well, so we can implement searches based 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 optimal configurations to be selected for each of these various functions. Note however that as we also offer an Open Source version of all

Europeana software, 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.

2.1 Future work – consequences of EDM

With the move to the new Europeana Data Model, we can optimise 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-ised 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 normalisation processes may be invoked, such as named entity recognition and linking to Geonames or VIAF records, normalisation of date values, etc. All enriched and normalised 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, potentially still allowing reasoning on these links. This is an interesting area of research that Europeana will need to consider while still maintaining an optimal 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 optimise the EDM data structures without having to go back to the original providers to ask them for re-submission of their data. In most cases a much simpler update of the mapping file plus a re-index will suffice.

Europeana currently does not envisage a radically different technical architecture to support these changes. Lucene/Solr are still valid choices for this environment with their ability to handle both fielded and generic searches equally well. 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.

3. Recommendations for future work

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

3.2 Overview of current issues

From work with the participants in WP3 and the development team in the Hague, the following items have been selected for further analysis:

- FRBR/CRM harmonisation: 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.
- DBpedia: practical applications: Linking Europeana object representations to various Linked Open Data resources and namely to DBpedia.
- 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
- Enable Support for Scholarly Inferencing
- Authentication/Identification: SAML, Shibboleth. Provide an open, standards-based authorisation 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).

The above will be further described in the following sections. At the same time, they are starting points for some of the activities planned in WP7 of EuropeanaV2.0.

3.3 Further Evolution of the EDM including FRBR harmonization

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

> Functional Requirements for Bibliographic Records — or FRBR — is a conceptual entity-relationship model developed by the International Federation 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.

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

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 are person and corporate body, responsible for the custodianship of Group 1's intellectual or artistic endeavour.
- Group 3 entities are subjects of Group 1 or Group 2's intellectual endeavour, and include concepts, objects, events and 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 as follows:[3]

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

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

> 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 harmonisation 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, highlighting that OCLC WorldCat has been "FRBRised" with 110 million records representing 85 million works.[5]

An article by Jenn Riley, Caitlin Hunter, Chris Colvard, and Alex Berry of the Indiana University Variations project, "Definition of a FRBR-based Metadata Model for the Indiana University Variations 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 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).

Note that a mapping of FRBRoo and EDM is offered by CIDOC.

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*
- *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*

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

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

3.4 Linked Open Data Integration and Linking DBpedia

DBpedia, as the most prominent Linked Open Data resource, gives the following description regarding its own activities:[7]

> 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.
>
> [...]
>
> 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 organise its content. Faviki uses DBpedia for semantic tagging. Amazon provides the DBpedia Public Data Set that can be integrated into Amazon Web Services applications.

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

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.

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.
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.

Finally, it will be crucial to determine to what extent and in which way Europeana itself will integrate in the linked open data paradigm and thus be available as a contextualisation resource for others.

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*

3.5 Use of DDC as contextualisation resource

DDC is described in Wikipedia as follows:[8]

8 http://en.wikipedia.org/wiki/Dewey_Decimal_Classification, 22 May 2010

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 organises 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.

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

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/.

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 harmonising 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)

3.6 Enable Support for Scholarly Inferencing

We should evolve Europeana into a scholarly source environment enabling knowledge generation, capable of producing digital heuristics. In this respect, support for reasoning and inferencing is key, 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 Wikipedia entry on OWL reads as follows:[9]

> 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 announced 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).

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/*

9 http://en.wikipedia.org/wiki/Web_Ontology_Language, 22 May 2010

- *Nick Drummond, Matthew Horridge. A Practical Introduction to Ontologies & OWL. 2005. http://www.co-ode.org/resources/tutorials/intro/slides/ProtegeOWLPart2-v05.ppt*

3.7 Authentication and authorisation

The JISC Identity Management Toolkit gives the following description of identity management and related technology:[10]

> 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-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:[11]

> 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 the management of identity and permissions for organistions 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 organisational credentials. The organisation (or identity provider) passes the minimal identity information necessary to the service manager to enable an authorisation decision.
>
> There are two primary parts to the Shibboleth system:

---

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

11 http://shibboleth.internet2.edu/about.html, 22 May 2010

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

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/*

4. Further work

The items above are considered to have primary importance for the future development of Europeana and more precisely affect the releases directly succeeding Danube. They will be further assessed and extended in the deliverable D3.4 due in 2011, leading to recommendations for practical and research activities in the years ahead. This includes a mapping, matching data values & data ingestion working environment (workflow design and implementation). Some of this (GUI) is defined in ASSETS. We should be careful to include tools, organisation and communication aspects in a holistic approach.