STYRELSEN FOR BIBLIOTEK OG MEDIER

NOTAT

17. november 2008

Leif Andresen
Specialkonsulent

LEA@bibliotekogmedier.dk
Direkte tlf.:33 73 33 54

Europeana og metadata

Indledning

Første udgave af dette notat blev lavet i april 2008 bl.a. som baggrund for forberedelserne til at levere data til prototypen af Europeana.eu. Det foreligger hermed i en opdateret udgave med bl.a. senere udgaver af to af de tre bilag.

Det er fortsat uafklaret i hvilke formater metadata vil kunne leveres i den endelige driftsversionen af *EDL - European Digital Library - Europeana* - udover Dublin Core (i den enkle form) og i Dublin Core Qualified. Men det vil være et begrænset antal og næppe nogle af de specifikke nationale formater, som p.t. anvendes i Danmark (Arkibas, Daisy, danMARC2 og Regin). Der vil derfor være behov for noget konvertering hos den enkelte institution før metadata leveres til den kommende driftsversion af Europeana.eu. Til den tid vil denne konverteringsfunktion evt. varetages af en dansk aggregator.

I forhold til de alment brugte formater i arkiver, biblioteker og museer foreligger der konverteringstabeller fra disse til det fælles præsentationsformat DKABM, hvor der enkelt vil kunne konverteres til både OAI-PMH headerinformation i Dublin Core og til Dublin Core Qualified.

Metadata til og i EDL

Der foreligger to dokumenter, som beskriver metadata i forhold til Europeana: *EDLnet D2.2 Initial Semantic and Technical Interoperability Requirements*, som er færdigt og som indeholder et afsnit, som beskriver hvordan data forventes leveret til EDL. Dette afsnit er gengivet som Annex A.

Den interne EDL datamodel bygger på Dublin Core Qualified. Dette fremgår af *EDLnet D2.5 Europeana Outline Functional Specification For development of an operational European Digital Library*, som foreligger i version 1.2. Det relevante afsnit er gengivet som Annex B og beskriver hvordan data forventes lagret internt i Europeana.

Som baggrundorientering er fra D2.5 som Annex C gengivet *Logical data model: Objects and Surrogates*, som beskriver EDL's data model.

Endvidere findes der et dokument, som beskriver den første implementeringen i den prototype, som launches den 20. november 2008: Specification *for the Metadata Elements for the European Prototype*

http://dev.europeana.eu/public_documents/Specification_for_metadata_elements_in_the_Europeana_prototype.pdf

Leverance af metadata til EDL

Data skal leveres med hjælp at OAI-PMH dels i dennes header (i Dublin Core Simple med 15 elementer) og dels som supplerende metadata leveret i XML-baserede metadataformater – jævnfør D2.2. Til den første prototype kan der forventes mulighed for at aftale f.eks. filtransport med ftp.

Som metadata leverandør kan man således vælge mellem selv at anvende DC Qualified eller et format, som bliver anerkendt af EDL kontoret.

De specifikke danske sektorformater er der ikke megen grund til at tro vil kunne håndteres af EDL. I givet fald ville al viden om hvordan dataindhold skal konverteres til det interne EDL format, skulle leveres fra dansk side. Det vil formentlig være en mere overkommelig opgave at forestå konvertering til DC Qualified selv.

For bibliotekerne kan overvejes at konvertere danMARC2 til MARC21 (som vil blive håndteret af EDL kontoret), men en mere direkte konvertering til DC Qualified vil formentlig give et bedre resultat.

Det bør overvejes at få EDL kontoret til at acceptere dkabm (se næste afsnit). På denne måde ville en række emneordssystemer blive identificeret, som senere vil kunne blive værdifulde med etablering af sammenhængende emne-ontologier (et aspekt, som der er stor opmærksomhed på ved EDLnet WP2 møderne). Endvidere kan identifikatorsystemer som ISBN genkendes.

Registreringsniveau

Hovedsigtet er at registrere hvad der svarer til en traditionel bibliografisk enhed. En stor del af bestræbelserne med EDL datamodellen handler om at give en samlet præsentation af de filer, som samlet udgør en enhed. Der er derfor også dialog med OAI-ORE (Open Archives Initiative. Object Reuse and Exchange (http://www.openarchives.org/ore/)

Men der er også opmærksomhed på at præsentere data i sammenhæng og derfor vil tilknytning til en ”collection” være relevant at lade indgå i metadata til EDL. Hvordan er ikke helt klart endnu.

Genbrug af DKABM specifikationer

Der bygges nedenfor videre på *Specifikationer for fælles præsentation af data fra arkiver, biblioteker og museer på internettet*, hvor der er mapningstabeller til dkabm fra museernes Regin, Statens Arkivers Daisy, lokalarkivernes Arkibas og bibliotekernes danMARC2.

Henvisning til *Specifikationer for fælles præsentation af data fra arkiver, biblioteker og museer på internettet:* http://www.bs.dk/standards/abm/

Mapninger kan genanvendes med ganske få justeringer for at passe til den interne EDL fortolkning af DC Qualified (som ikke anvender DC.Creator, så data skal flyttes til DC.Contributor.)

danMARC2
For Version lægges denne information i DC.Description i stedet.
En række emneords schemes droppes blot og data lægges "nøgent" i DC.Subject

Regin
Brugen af scheme SKRM skal overvejes. En udfoldning fra denne geografi-kode til tekst for sted vil være brugervenligt.

Daisy
For DC.Creator skal *actPeriod* enten droppes eller ændres til "tilføjet tekst" i DC.Contributor.
For DC.Creator skal *alternativeName* droppes og data lægges i DC.Contributor.

Arkibas
For DC.Creator skal *actPeriod* enten droppes eller ændres til "tilføjet tekst" i DC.Contributor.

DKABM faciliteter

For biblioteker er som tillæg til dkabm – men forudset i mapningstabellen for danMARC2 – lavet en konverteringsfacilitet fra danMARC2 (i MarcXchange = XML) til dkabm. Denne indeholder også en udfoldning af DK5 til DK5-tekst. D.v.s. at dkabm indeholder et antal emnebeskrivelser i form af en tekstlig beskrivelse svarende til en DK5 kode.

Genbrug af denne konverteringsfacilitet er en oplagt mulighed.

En mulig senere udbygning vil være at oversætte DK5 til engelsk og lægge parallel emne tekst på engelsk ind. Om dette vil være en relevant investering afhænger af en række forskellige forhold, herunder omfanget af DK-5 på digitaliseret materiale og den ikke afgjorte håndtering af hvordan flersprogethed i Europeana føres ud i livet.

Med venlig hilsen

Leif Andresen
Styrelsen for Bibliotek og Medier
Ekspert tilknyttet EDLnet Workpackage 2: Technical & Semantic Interoperability

Annex A
EDLnet D2.2 Initial Semantic and Technical Interoperability Requirements 17 December 2007

Metadata

4.1 Open Archives Harvesting Approach

This specification starts from the assumption that EDL will use the OAI-PMH [37] harvesting approach which implies that all EDL content providers must act as an OAI repository and meet the requirements as set out by the specifications of version 2.0 of the protocol.
In this approach, harvested information contains three parts:

- *Header*. This contains the unique identifier of the OAI item (defined as a constituent of a repository from which metadata about a resource can be disseminated) and properties necessary for selective harvesting, i.e. the datestamp, zero or more specifications of sets that the item belongs to, and an optional status attribute that can indicate that the metadata for an item has been withdrawn
- *Metadata*. This part contains a metadata record in a format requested by the harvester. Any OAI repository must be able to return records with metadata expressed in the Dublin Core format, without any qualification. Optionally, a repository may also disseminate other formats of metadata which can be requested by the harvester by means of an argument – the metadataPrefix – in the GetRecord or ListRecords request that produces the record.
- *About.* This part is an optional and repeatable container to hold data about the metadata part of the record, for example with a rights statement of provenance information.

In EDL, all content aggregators and contributors are required to provide metadata about their resources in unqualified Dublin Core [56]. This basic metadata will be used to build a basic index to be used for simple search.
To be able to provide more elaborate services, all content aggregators and providers are encouraged to provide more elaborate metadata. Investigations are ongoing to determine which formats will be recommended.

4.2 Dublin Core metadata

The list below is based on the common metadata set that is required for OAI-PMH, unqualified Dublin Core[1], and describes the usage of 14 of the 15 Dublin Core elements in EDL. The mandatory elements are presented first followed by the optional ones and the one that is not used.

Name	Obligation	Occurrence	Type	Encoding	Vocabulary	Comments
Format	Mandatory	Once	Literal	IANA MIME types [57]		MIME type that is appropriate. If none exists, use application/octet-stream
Identifier	Mandatory	Repeatable	Reference	URL [58]		Resolvable link to the object
Rights	Mandatory	Once	Reference	URL	EDL terms-of-use vocabulary or Creative Commons	Pointer to term in EDL terms-of-use vocabulary or Creative Commons license
Source	Mandatory	Once	Literal	UNICODE string	EDL content provider list	Name of the organization holding the digital object in the form given in the EDL content provider list
Subject	Mandatory	Repeatable	Literal	URL	EDL subject vocabulary and terms from acknowledged subject vocabularies	Term in EDL subject vocabulary
Title	Mandatory	Repeatable	Literal	UNICODE string		Name of the object and for any variants or abbreviations that helps with discovery of the object
Type	Mandatory	Repeatable	Literal	UNICODE string	DCMIType [59] and EDL Type vocabulary	Term in the DCMI Type vocabulary and, if necessary, term in the EDL type vocabulary
Contributor	Optional	Repeatable	Literal	UNICODE string		Names of any persons or organizations that have contributed to the content of the digital object, including the creator of the physical object that a digital object depicts
Coverage	Optional	Repeatable	Literal	W3CDTF [60] if date or date range	Named periods or Geonames [61] for geographic names	(a) Temporal coverage of the object, i.e. a (period of) time that the content is related to, and (b) Geographic coverage, i.e. a location that the content is related to

[1] http://www.openarchives.org/OAI/openarchivesprotocol.html#dublincore

Name	Obligation	Occurrence	Type	Encoding	Vocabulary	Comments
Date	Optional	Once	Literal	W3DTF		Most relevant date for discovery of the object
Description	Optional	Repeatable	Literal	UNICODE string		Any descriptive text that is associated with the object and is relevant for discovery
Language	Optional	Repeatable	Literal	ISO 639-3 [62]		Major language(s) in resources with text or spoken language
Publisher	Optional	Once	Literal	UNICODE string		Name of the formal publisher of the object
Relation	Optional	Repeatable	Reference	URL		Links to related resources, e.g. collection that the object is part of, other resources with the same content in different format or other related resources
Creator	Not used					

This proposal leaves the freedom to the content provider to decide exactly what to put in the Dublin Core elements. The last column in the document intends to give guidance as to the types of information that could be used. The actual decisions should be based on a view of which information would be useful for refining searches by a "general user" (not a domain specialist).
The following controlled vocabularies are proposed:

1. For Source: a list of EDL content providers to ensure consistent and persistent relationship between the objects and the holding institution.
2. For Rights: an EDL terms-of-use vocabulary to be used if an object is not governed by a Creative Commons [63] license.
3. For Subject: exchange to be based on formal subject vocabularies. Mandatory use of one or more terms from an EDL subject vocabulary and, additionally, use of terms from acknowledged subject vocabularies (e.g. vocabularies maintained by the national libraries and other domain-specific vocabularies such as the Getty Art and Architecture Thesaurus [64])
4. For coverage: vocabularies for (a) periods and (b) places to normalise temporal and geographic information.

The aggregators and content providers need to do the expansion of terms used in Source, Subject and Coverage from the controlled vocabularies at their end and deliver only textual information in these fields of the metadata to be harvested.

4.3 Additional metadata

In addition to the common metadata set, OAI-PMH allows the exchange of other metadata formats. EDL will allow content providers to make more elaborate metadata available for harvesting.
However, to limit the amount of mapping options, there is a need to limit the number of supported schemas. For the library world, the TEL application profile [65] (an example of a qualified Dublin Core profile) could be the basis for this. For the other domains, similar intra-domain solutions need to be found: for the AV

archives (e.g. EBU Core [66], Immix [67], Dismarc [68]), archives (e.g. EAD [69], Moreq2 [70]), and museums (e.g. CIDOC CRM [44]). The XML schemas for these metadata sets need to be provided by the content provider.
The central EDL index will use all textual information in these more elaborate sets of metadata for full-text searching.

Annex B
EDLnet D2.5 Europeana Outline Functional Specification For development of an operational European Digital Library

Public Draft Version 12 August 2008

Europeana Metadata Requirements

4.1.2.1 Object_metadata

The basic OAI-PMH mechanism may be used to harvest simple Dublin Core metadata from the content providers[2]. It is foreseen that Europeana will also receive additional object-specific metadata, either through the OAI-PMH getRecord request with appropriate metadataPrefix or through other means. This more detailed metadata should be delivered according to an XML format that is agreed between the content provider and Europeana management. Possible formats include: qualified Dublin Core conforming to an Application Profile such as the one defined for TEL[3], METS[4], EAD[5], EBU Core[6], Immix[7], CIDOC CRM, MODS[8], MARCXML[9],

---

[2] Dublin Core Metadata Element Set, version 1.1: http://dublincore.org/documents/dces/

[3] TEL Application Profile for Object: http://www.theeuropeanlibrary.org/handbook/Metadata/tel_ap.html

[4] Metadata Encoding & Transmission Standard (METS) - http://www.loc.gov/standards/mets/

[5] EAD – Encoded Archival Description: http://www.loc.gov/ead/

[6] EBU Core Metadata Set: http://www.ebu.ch/metadata/documentation/EBUCore/tec_doc_t3293_2008_FinalDraft.pdf

[7] iMMix, Nederlands Instituut voor Beeld en Geluid, contact Annemieke de Jong adjong@beeldengeluid.nl

[8] Metadata Object Description Schema (MODS) - http://www.loc.gov/standards/mods/

[9] MARC 21 XML Schema - http://www.loc.gov/standards/marcxml/

MPEG-21 . CDWA[10], Dismarc[11], museumdat[12] and Moreq2[13]. The XML schemas for these metadata sets need to be provided by the content provider.
All incoming metadata in one of the agreed formats need to be converted to a common internal format, the semantics of which are described in the table below. The table is intended to include the full list of metadata elements that will be understood by Europeana. In addition, other metadata, if supplied by the data providers, can possibly be used for full-text indexing.
The link with the metadata format listed in section 4.2 of D2.2 is that some of the metadata elements below may be derived from Dublin Core metadata that is delivered as part of the initial OAI-PMH harvesting mechanism. As the default format for OAI-PMH is just simple Dublin Core, the metadata received through that mechanism will not meet all of the more detailed requirements outlined below. Therefore, additional object-specific metadata will also be harvested from providers if available.
A particular case where the object-specific metadata is necessary is the case of what we refer to as 'complex' objects. For these objects, structured metadata needs to be available as well that contains a description of a coherent collection of objects that need to be seen in context of the collection.
While for 'simple', 'atomic' objects, the processing of incoming metadata can be a more or less straight-forward conversion from the XML format provided to the internal metadata format, in the case of 'complex' objects, received metadata will need to go through a more elaborate process. Surrogates need to be generated for each of the components of the object and the metadata need to be decomposed into metadata records for the individual components. Appropriate linking between the surrogate for the 'root' object and the component surrogates and between the component surrogates, is necessary to precisely reflect the internal structure of the object.
The metadata elements described in the table need to be seen from a purely functional and semantic perspective. There is no pre-defined mapping to any particular implementation or metadata standard, although the Dublin Core properties are used as examples. The actual encoding of this internal format is left to the implementers.
Of the elements listed in the table only the first four are mandatory (location, owner, format and rights). For the other elements, as many as possible and relevant should be made available. These metadata can be derived from OAI-PMH metadata (if that is the mechanism used by a particular provider), from specific metadata that the provider is able to supply or from manual intervention by either experts or end-users. The metadata can be enhanced in various ways, harmonizing and/or linking to controlled vocabularies or authority files.

Semantic description	*Source – comment*	*Example DC property*

10 http://www.getty.edu/research/conducting_research/standards/cdwa/

11 DISMARC – Discovering Music Archives: http://www.dismarc.org/

12 http://museum.zib.de/museumdat/

13 Moreq2 – Model Requirements for the Management of Electronic Records: http://www.moreq2.eu/

Semantic description	*Source – comment*	*Example DC property*
Location of object (mandatory)	Expressed as a URI. For simple objects, from dc:identifier in OAI-PMH metadata and/or from specific metadata; for complex objects, pointers to the individual components are derived from further processing of detailed metadata; to be used for linking to the original object	dc:identifier
Institution holding the object (mandatory)	From dc:source in OAI-PMH metadata and/or from specific metadata; incoming data needs to be in standardised form so it can be converted to link to the record in the Europeana provider name authority file; to be used for institution-based searching and for grouping results by institution	dc:source
Object format (mandatory)	From dc:format in OAI-PMH metadata or from specific metadata needs to be one of the file types supported by Europeana; to be used for format-based searching and narrowing of results	dc:format
Rights	From dc:rights in OAI-PMH metadata and/or from specific metadata should be either a term of Europeana terms-of-use vocabulary or a Creative Commons[14] license; to be used in presenting usage restictions to the user	dc:rights
Contributor	From specific metadata; possibly enhanced through automatic processing, linking to name authority file, manual enhancement by experts; to be used for simple search and name-based searching	dc:creator and dc:contributor, including refinements
Creation date	From specific metadata; to be used for narrowing results	dcterms:created
Description	From specific metadata; possibly in multiple languages; to be used for simple search	dc:description
Geographic coverage	From specific metadata; automatic enhancement; possible further manual enhancement by experts; to be used in map-based searching and	dcterms:spatial

[14] Creative Commons: http://creativecommons.org/

Semantic description	*Source – comment*	*Example DC property*
	presentation	
Language	From dc:language in OAI-PMH metadata; to be further processed and harmonized using ISO 639[15]; to be used to restrict simple searches to specific language or narrowing results	dc:language
Modification date	From specific metadata; date of last update; to be used for narrowing results	dcterms:modified
Object type	From dc:type in OAI-PMH metadata and/or from specific metadata; possibly enhanced by particular genre (e.g. painting, book, video) or domain (library, museum etc.) or theme (an initial list as defined for the prototype) that could be derived from detailed metadata or added by further processing or manual enhancement by experts; to be used for relevance ranking and for narrowing of results	dc:type
Publication date	From specific metadata; to be used for narrowing results	dcterms:issued
Publisher	From dc:publisher in OAI-PMH metadata and/or from specific metadata; possibly enhanced and harmonized; to be used for name-based searches	dc:publisher
Relation	From specific metadata and from processing of 'çomplex' object metadata to create a network of surrogates reflecting the structure of the 'complex' object; needs to be further analysed to define the appropriate set of relation types	dc:relation and various standard and local refinements
Subject	From dc:subject in OAI-PMH metadata and/or from specific metadata; processing to link to 'semantic nodes'; possible enhancement through manual enhancement by experts and end-users; to be used in simple search and subject-specific search	dc:subject
Temporal coverage	From specific metadata; automatic	dcterms:temporal or

[15] ISO 639-3 – Codes for the representation of names of languages: http://www.sil.org/iso639-3/

Semantic description	*Source – comment*	*Example DC property*
	enhancement; possible further manual enhancement by experts; to be used for time-line presentation and navigation	dc:coverage.temporal
Title	From dc:title in OAI-PMH metadata and/or from specific metadata; any formal, informal, abbreviated or parallel title should be included; to be used in simple search and subject-specific search	dc:title

For every metadata statement, it needs to be recorded, as a minimum, (a) when it was last modified and (b) who made the modification (system, expert, user) in order for the system to be able to assign weights to the values depending on the trustworthiness of the metadata and the particular use that is made of the metadata (e.g. to allow for the functional requirement that user-provided metadata has a higher value than expert-provided metadata).

4.1.2.2 User metadata

Property	Comment
User ID	Internal, unique identifier
User status	Anonymous/registered
Last access	Date and time that user was last in the system (from cookie if user is anonymous, from login for registered users). May also be used to delete user records after certain period of inactivity.
Cookie info	
Login name/password	As provided by user when registering; password encrypted
User type	General/expert (to be specified by user when registering)
User domain	Archive/Museum/Library/Audiovisual Archive/Other (to be specified by user when registering)
User interest profile	Any type of information that the registered user wants to provide and that can help better ranking and suggestions. Possibly also generated by the system on the basis of past behaviour (also for returning anonymous users)
User subscriptions	Indicating which Europeana Communities the user is a member of (for registered users only)
User space	Pointer to private space assigned to user (for registered users only)
User e-mail address	If provided by user (for registered users only)

4.1.2.3 Provider metadata

Property	Comment
Provider ID	Internal, unique identifier
Provider status	Content holder/Aggregator
Provider domain	Archive/Museum/Library/Audiovisual archive/ Archaeology/Monuments/Other
Last harvest	Date and time of last successful harvesting
Harvesting mechanism	OAI-PMH/FTP/etc. (code for any supported mechanism)

Property	Comment
Harvesting format	Available metadata format for harvesting
Harvest address	URL where file will be available for harvesting
Provider Name	As supplied by provider
Address information	
Web site address	
Contact person name	
Contact e-mail	
Aggregator link	Provider ID of Aggregator if content from this provider is received from an aggregator
Content holder link	(Repeatable) Provider ID of Content holder that this Aggregator aggregates

Annex B
EDLnet D2.5 Europeana Outline Functional Specification For development of an operational European Digital Library

Public Draft Version 12 August 2008

Logical data model: Objects and Surrogates

A central principle for building Europeana is that a network of semantic resources will be used as the primary level of user interaction. In a classical librarian catalogue model all user access to information objects is mediated by descriptive metadata as illustrated in Figure 1 below:

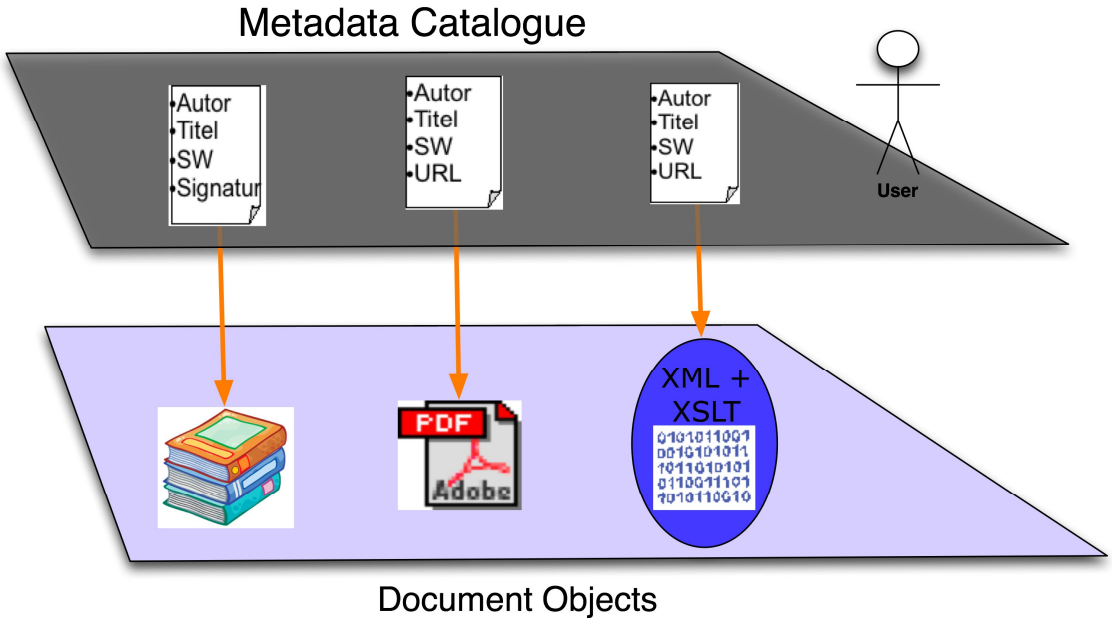

Figure 1: Catalogues and Information Objects in Digital Libraries

Unlike in such librarian functional models users are expected to explore the Europeana data space using semantic nodes as primary elements for searching and browsing along paradigms indicated by the questions as to "Who?", "Where?", "When?" and "What?" The intended relation between the

semantic and the object representation layers with respect to the Europeana user interface is illustrated in Figure below:

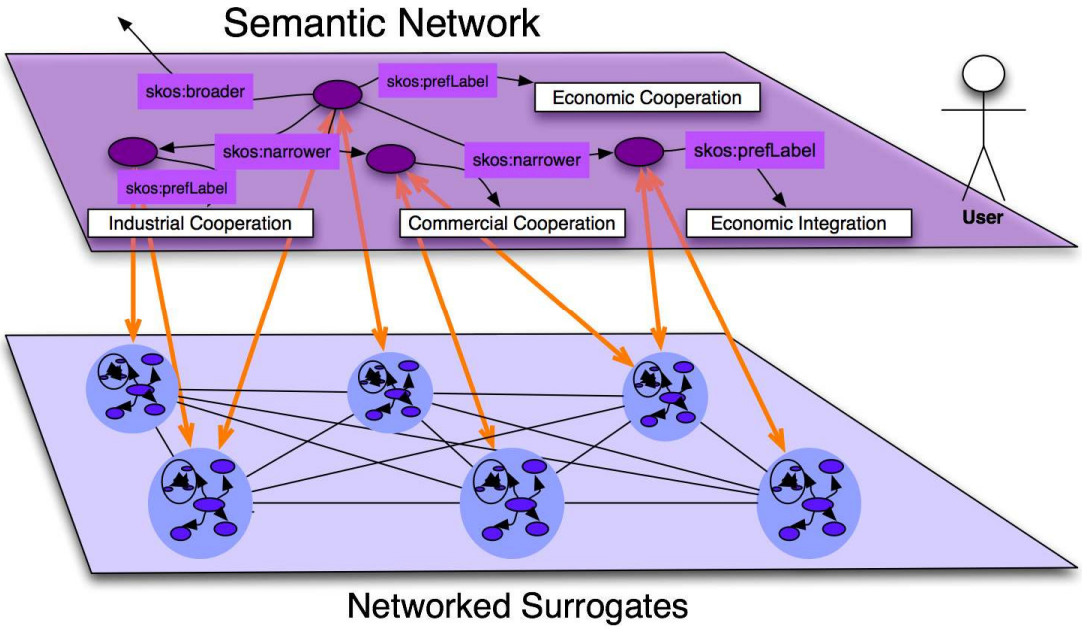

Figure 2: Semantic and Surrogate Layer in Europeana

The user now primarily interacts with the semantic network to explore the Europeana surrogate space which now has the metadata as parts of the surrogates and surrogate aggregations.

In the perspective of this approach, Europeana can be thought of as a ***network of inter-operating object surrogates enabling semantics based object discovery and use***. This network in turn is an integral part of the overall information architecture of the WWW.

Furthermore, the Europeana object model is based on the assumption that the central Europeana data store will only contain object surrogates and index files, whereas original objects are located at the content provider sites. Europeana thus will create a parallel data space inside the system that is a representation of the real world object space. As a consequence, we distinguish 'object entities' (to indicate an external object plus any associated metadata about that object) and 'surrogate entities' (to indicate the internal object with associated metadata and other composite elements). Likewise, two separate data spaces need to be distinguished: an external space of objects entities and an internal space of surrogate entities.

(i) Surrogates

All surrogates in the Europeana data space are web resources in the sense defined by the W3C and thus have a URI[16] identifier. They also contain a link to the object entity in case this object can be identified as a web resource.

[16] Uniform Resourc Identifier (URI) - http://tools.ietf.org/html/rfc3986

Otherwise the link will be to an external application permitting access to this object. In some specific settings requiring exclusive control by the content provider of all access methods and functionality a surrogate thus can be limited to being an entry to a content access point under control of the content provider.

In such an approach the Europeana surrogate model can be completely agnostic about where the original objects are stored: the URI link to the object syntactically remains the same. The surrogate model thus isn't affected, in case the option of also keeping original objects within Europeana is needed (for instance for content providers that do not have a content store of their own or for some other reason prefer to store their objects within the Europeana environment): these objects would still be kept in a separate Europeana object store and be referenced from their surrogates just as external objects would be.

The model is conceived from an 'atomic', bottom-up perspective: the basic building blocks are surrogates representing the minimal significant documentary object units a given content provider is able / willing to identify (in the case of textual object there thus can be surrogates on the level of the entire document, on chapter level or on page, paragraph, sentence or even word level).

Each of these surrogates contains at least a URI, a link to the original object, metadata as well as different kinds of abstractions, aggregations or derivatives depending on object characteristics. Examples of such abstractions/aggregations/derivatives are tables of contents and indexes, full text index items, thumbnails, music and video abstractions (e.g. colour histograms or shape abstractions) and signatures. Surrogate metadata records as part of these surrogates are sets of RDF triples.

These atomistic surrogates can be linked to each other to form complex aggregations, which in turn can be organised as Description Sets based on the DCMI Abstract Model[17] or OAI-ORE Resource Maps[18]. These surrogate aggregations correspond to compound logical objects on the content provider's side such as scanned books or multipart multimedia objects, to give just two examples. The central (and mandatory) element of each surrogate aggregation (everything within the light blue circle in the diagram below) is an aggregation root element with a URI of its own and containing some elementary technical and (mandatory!) licensing information. A 'landing page' rendition of this root element is used to expose Europeana DL content to external software agents such as search engines.

There should be a one-to-one correspondence between remote object entities and internal surrogate entities as well as between remote compound logical object entities and complex aggregations of Europeana surrogates. In such a perspective, the decision regarding the actual boundaries of a complex

---

[17] DCMI Abstract Model: http://dublincore.org/documents/abstract-model/

[18] OAI-ORE – Open Archives Initiative Object Exchange and Reuse: http://www.openarchives.org/ore/

surrogate aggregation largely depends on the way the object providers conceive the entities they want to make accessible via the Europeana surrogate space.

Furthermore, Europeana surrogates as well as surrogate aggregations will systematically be linked to semantic resources representing concepts as well as to external reference resources representing reference entities such as persons, places and periods in time. Links to these reference resources are used in order to create context for the Europeana surrogates. The reference resources may be part of the Europeana data space or external to it: they are referred to as web resources using a URI in either case.

In either case, the semantic resources surrogates are linked to will be organised as semantic web ontologies (hitherto referenced as 'ontologies'), containing the vocabularies for describing the meaning of surrogate aggregations. Semantic ontologies include thesauri, classification schemes, subject heading systems, taxonomies, and the like. Semantic ontologies will be used to define entities (which mostly have a lexical counterpart) both at the concept/class level (by defining the entity classes and the relations between them), and at the word/object level (by defining the allowed instances of semantic ontology classes). The latter mechanism will allow to model authority files in the sense of collections of valid instances. Moreover, domain knowledge such as historical events or biographical information is also modelled via semantic ontologies; a notable example of this is the CIDOC CRM[19].

Semantic ontology classes and objects will be associated to surrogate aggregations in two ways:

(1) implicitly, as string metadata attributes, such as the dc:subject attribute connecting an aggregation to a term from a classification scheme, or the dc:creator attribute connecting an aggregation to an entry in an authority file;

(2) explicitly, via *classification* association links to web resources (URIs). An example of a classification association is the "is about" association, relating an aggregation to a topic (i.e., class in a semantic ontology); another example is the "represents" association, relating an aggregation (or a single surrogate) to an object, instance of a person or an event class.

(ii) Associations

Both the links within aggregations (part-whole relations) and between aggregations as well as between surrogates/aggregations and reference resources are not yet given definitive types in this version of the specification document but the need of more 'specialisation' is evident. This 'specialisation' is likely to draw upon the Resource and Content domains of the DELOS reference model[20], standardisation attempts such as MPEG21 DIDL[21] or

---

19 CIDOC CRM – Conceptual Reference Model: http://cidoc.ics.forth.gr/

20 DELOS Reference Model: http://www.delos.info/ReferenceModel

PRISM[22], as well as on the ORE Abstract Model and the related vocabulary as part of the ORE specifications[23]. Another promising starting point for typing relations is the list of FRBR property declarations as part of FRBRoo [24] as well as the CIDOC CRM properties referred to in this document. The result of the work on typing relationships/links will be a framework for expressing complex multimedia object structures as well as structural relations between objects and reference entities. An ontology (or several ontologies) of such structural relations - also known as a content model – will be needed in this respect.

However, we can assume some initial guiding lines and distinguish the following association types:

- *content* associations, relating a surrogate to other surrogates, to reflect structural relationships between the corresponding objects. These associations can be further characterized as:
    - associations defining object structures; different content models will have different types, but they can be taxonomized as specializations of the IsPartOf relation;
    - associations capturing versioning; there are several models which can be used for inspiration, depending how sophisticate the underlying mechanisms need to be; versions can form a single line or a tree or a directed acyclic graph;
    - the FRBR associations.
- *description* association, relating surrogates to the metadata objects describing them in some description ontology. The singular here means that in principle one should not have more than one association; but if needed, these associations may form a taxonomy with one specific association being the root;
- *naming* associations, relating surrogates to their appellations, that is object-level elements of terminological ontologies. There is an implicit associative mechanism here, because if x is the name of a surrogate and y is a synonym of x, then also y can be used as a name for that surrogate;
- *classification* associations, relating surrogates to the concept-level elements of terminological ontologies. An example is the "is about" association, relating an surrogate to a topic; another example is the "represents" association, relating an surrogate (or a portion thereof) to a world entity representation; a third example is the "instance of" association, relating a surrogate to a class, like Monna Lisa being an instance of Renaissance Art. Inference plays a major role also here, in the sense that concepts in terminological ontologies may be connected by logical relations (sumbsumption, equivalence), which therefore apply to the relatum. A classical example is the inference that has as antecedents "object X is an instance of Renaissance Art", "Renaissance Art is Art" and as consequent "object X is an instance of Art". Different associations may have different logical properties, which should be stated by specifying the semantics of the associations.

---

[21] MPEG 21 DIDL: http://xml.coverpages.org/MPEG21-WG-11-N3971-200103.pdf

[22] PRISM, Publishing Requirements for Industry Standard Metadata, http://www.prismstandard.org/

[23] OAI-ORE Abstract Data Model: http://www.openarchives.org/ore/0.1/datamodel

[24] http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBR_oo_V0.9.pdf

- *similarity* associations, relating the surrogate of an object to the surrogates of the objects that are similar to it. Similarity can be defined along several axes:
    - *content-based similarity,* capturing resemblance between text (as established by information retrieval models), images, audio, video and audio-visual objects. These associations are typically computed on demand rather than stored. An important point is that this kind of associations are typically application-dependent, so their set should be extensible.

recommendation, capturing resemblance between objects as established by experts (possibly via annotations), usage (people who accessed one object often accessed the other one), or other criteria.

In this respect, it remains to be determined whether Europeana will fundamentally distinguish relations within an aggregation from those linking aggregations to each other or to reference resources. This issue as well as the potential internal recursive structure of Europeana surrogate aggregations has a counterpart in the OAI-ORE regarding the distinction of internal and external relations. While abstracting from the wealth of object modelling options and choosing a few, general ones that capture structural relations in the most popular existing standards Europeana should take good care to evolve in line with the ORE model in order to preserve the interoperability potential with the repository community.

Europeana object surrogates can thus be simple entities or can be aggregated into potentially complex logical entities and related to other surrogates and reference resources. A logical overview of this is given in Figure 2 hereafter:

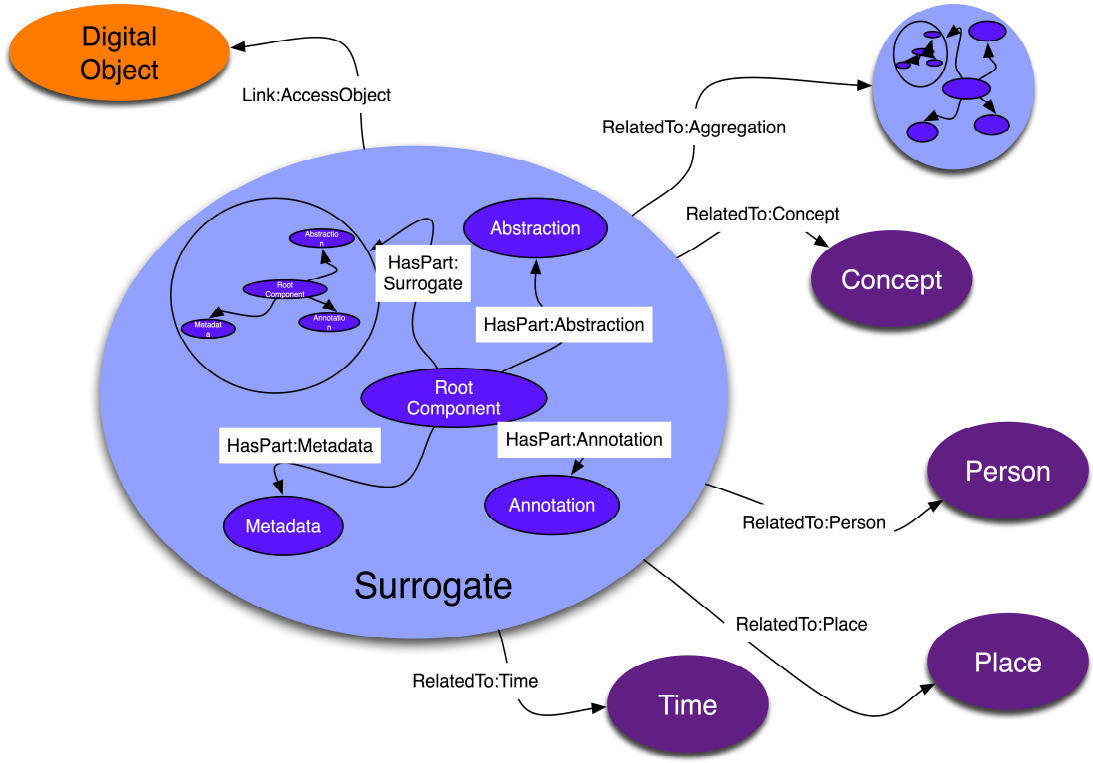

Figure 2: Surrogate Model Logical Overview

Surrogates will have, metadata, abstractions (such as tables of content or colour histograms) and annotations as parts, surrogate aggregations additionally have component surrogates which in turn may have a complex internal structure as described above. The 'HasPart' link is used to point from the surrogate to its components. The 'RelatedTo' link is used to point from the surrogate to external entities. As said before, both links evidently need to be specialised and 'typed', and one important task in further working out the functional specifications of Europeana is to produce a list of values specifying the type of relation (work to be done jointly with the ORE initiative).

Surrogates can be exposed via the Europeana API and/or Europeana portal services, but the API should be underlying the portal, too, and exposure via API should thus be considered the standard way of surrogate delivery.

An important consideration for the surrogate model is that it is necessary that surrogates can be referenced.

The degree of object granularity to be delivered is determined by the content provider who additionally should be given the possibility to indicate the object modelling schema he is referring to in conceiving object building blocks (e. g. TEI[25] or DocBook[26] or MPEG21[27]). This information will be used by Europeana when creating surrogate representations of these objects for translating the relations the content provider has conceived on object level to the structural relations known within the Europeana content model (cf. above).

---

25 Text Encoding initiative TEI) - http://www.tei-c.org/Guidelines/P5/

26 DocBook - http://www.docbook.org/

27 MPEG21 - http://mpeg-21.itec.uni-klu.ac.at/cocoon/mpeg21/