

europæana
think culture

Integration of Heterogeneous Metadata in Europeana

Cesare Concordia

cesare.concordia@isti.cnr.it

Institute of Information Science and Technology-CNR

Outline

- What is Europeana
- The Europeana data model
- The Europeana Semantic Elements (ESE)
- Case study: the data ingestion in the Europeana prototype
- Conclusion and next steps

What is Europeana

- European Digital Library
- Open access to the digitized objects of European cultural institutions
- Cross multilingual search European cultural heritage at a single place
- Across cultural domains and across countries
- General public - User centered
- Digital Library technologies + Web 2.0

What is Europeana

- The Europeana Digital Library will be the result of a number of projects run by different cultural heritage institutions, among them there are:
 - **Athena** an aggregator that helps museums bringing their content to Europeana
 - **APENet** a BPN whose objective is to build an Internet Gateway for Documents and Archives in Europe
 - **EUROPEANAlocal** that aims to improve the interoperability of the digital content held by regional and local institutions
 - **European Film Gateway**: find solutions for providing integrated access to the Europe's cinematographic heritage
- All are part-funded by the European Commission's eContentplus programme.



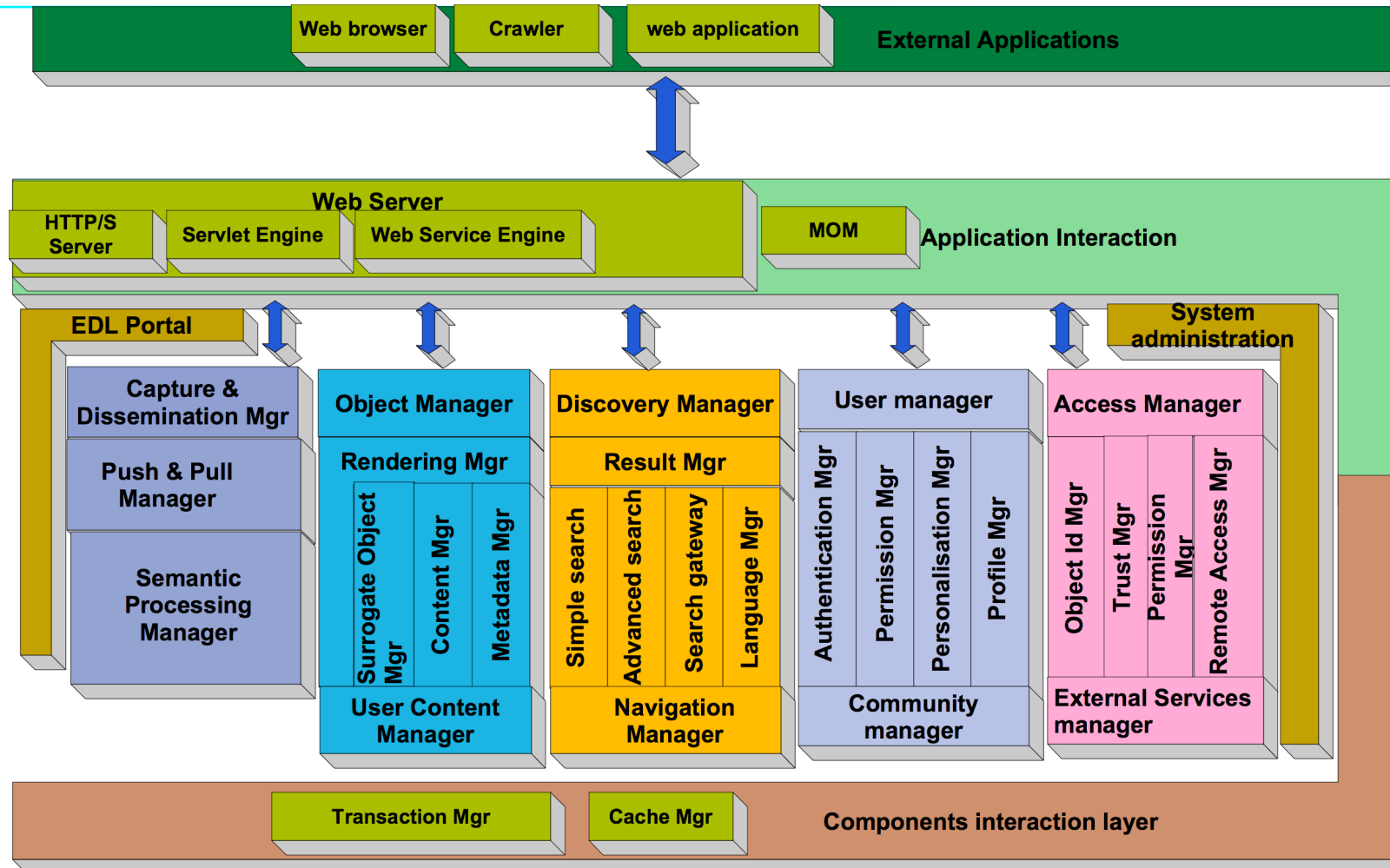
What is Europeana

- The full implementation of the Europeana is the goal of the two “core technology” projects:
 - **EuropeanaConnect** that will provide the technologies and resources to semantically enrich the digital content in Europeana.
 - **Europeana V1.0** that will implement the technology platform
- They are successors to the **EDLNet** thematic network which created the EDL Foundation and the Europeana prototype (www.europeana.eu)
 - in short Europeana v1.0 and EuropeanaConnect will turn the prototype into an operational service

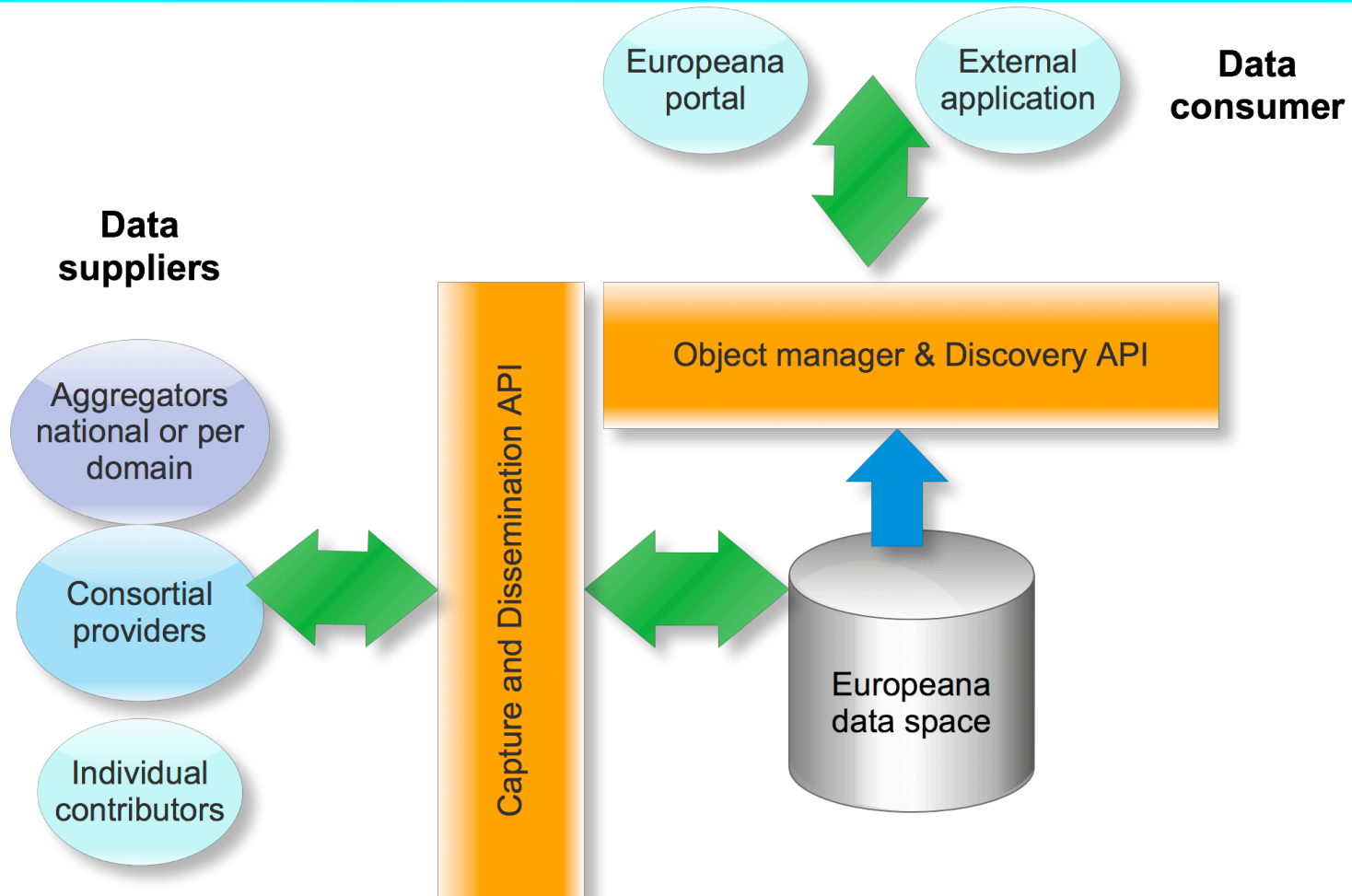
Europeana architecture

- Europeana is not a Web Portal
- Europeana is a **services platform** providing an Application Program Interface (API) enabling cultural institutions and users to
 - Access Europeana content
 - Provide content to Europeana
 - Build applications using Europeana functionalities for their own use.
- According to DELOS classification Europeana is a Digital Library System (DLS)
- The Europeana Portal is a web application using the Europeana API to access the Europeana Digital Library

Europeana DLS functional architecture



Europeana data flow



Europeana data space

- Europeana will create a data space that is a representation of the content providers data spaces
- The Europeana data space will contain Digital Surrogate Objects (DSO) that are defined as follows:
 - "the minimal significant documentary object unit 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 levels)" [EDLNet Deliverable 2.5]
- Each DSO will be a web resource, it will be identified by a URI owned by Europeana.

Digital Surrogate Objects

- There are several kinds of DSO, depending on the kinds of objects to be represented:
 - Real Physical Object (RPO): the physical object, for instance a painting, a building, a book
 - Digital Representation Object (DRO): a digital object obtained by digitizing an RPO, usually created by the data provider
 - Digital Primary Object (DPO), a "born digital" object, i.e. a digital object that is not a DRO

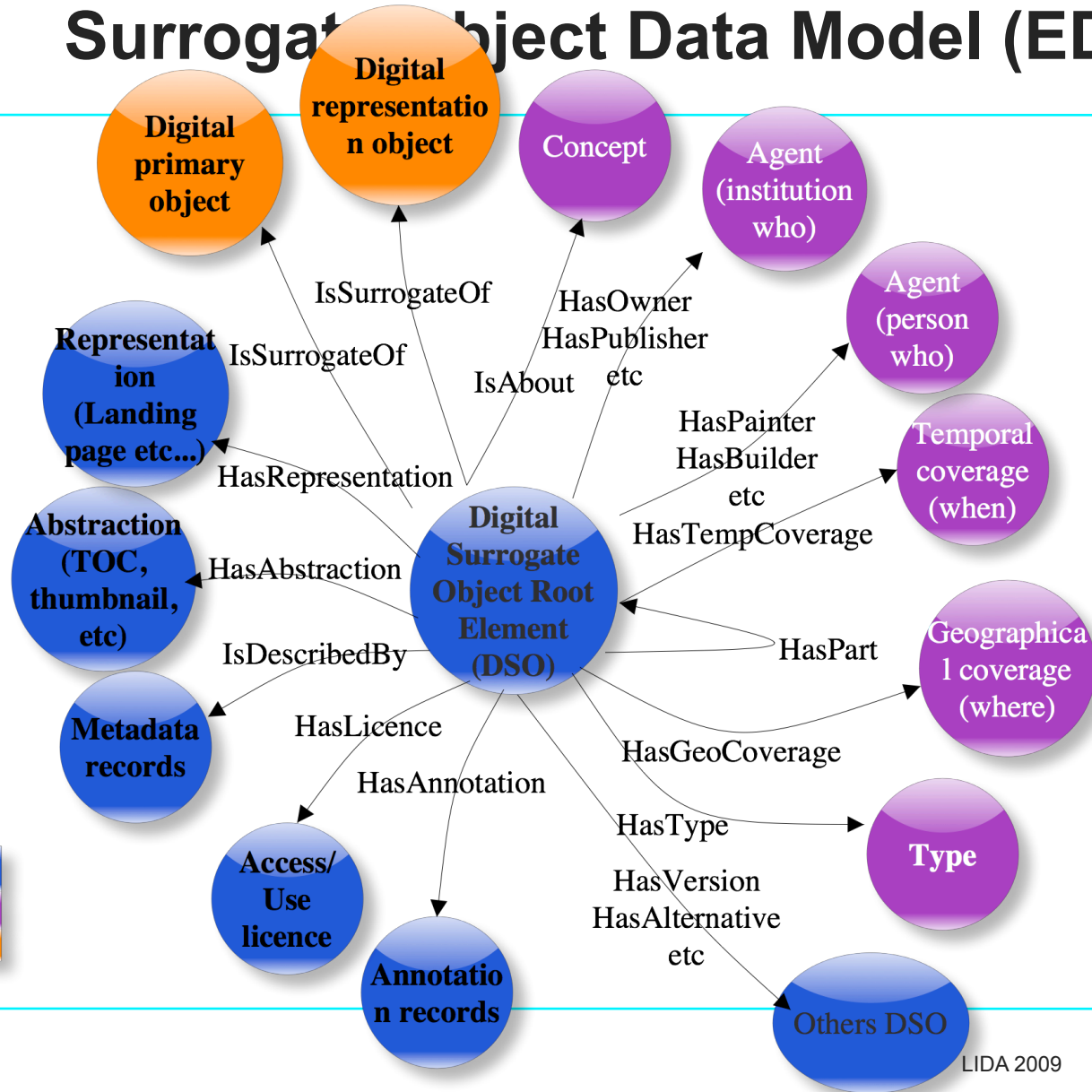
Digital Surrogate Objects

- Each DSO contains at least an identifier, a link to the Object in the content provider data space, the metadata record describing the object and some elementary technical and licensing information
- There should be a one-to-one correspondence between remote object entities and DSOs
- Surrogates can be linked each others

Digital Surrogate Objects

- On a very abstract level Europeana can be seen as a large collection of DSOs representing born digital or digitised cultural heritage objects
- Surrogates will be linked to semantic resources representing concepts as well as to reference entities such as persons, places and periods in time (contextualization)

Surrogate Object Data Model (EDLNet)



LIDA 2009

Europeana Semantic Elements

- In EDLNet the Surrogate Data Model has been implemented using the Europeana Semantic Elements (ESE) metadata format
- The ESE, consists of the Dublin Core (DC) metadata elements, a subset of the DC terms and a set of twelve elements which were created to meet Europeana's needs.

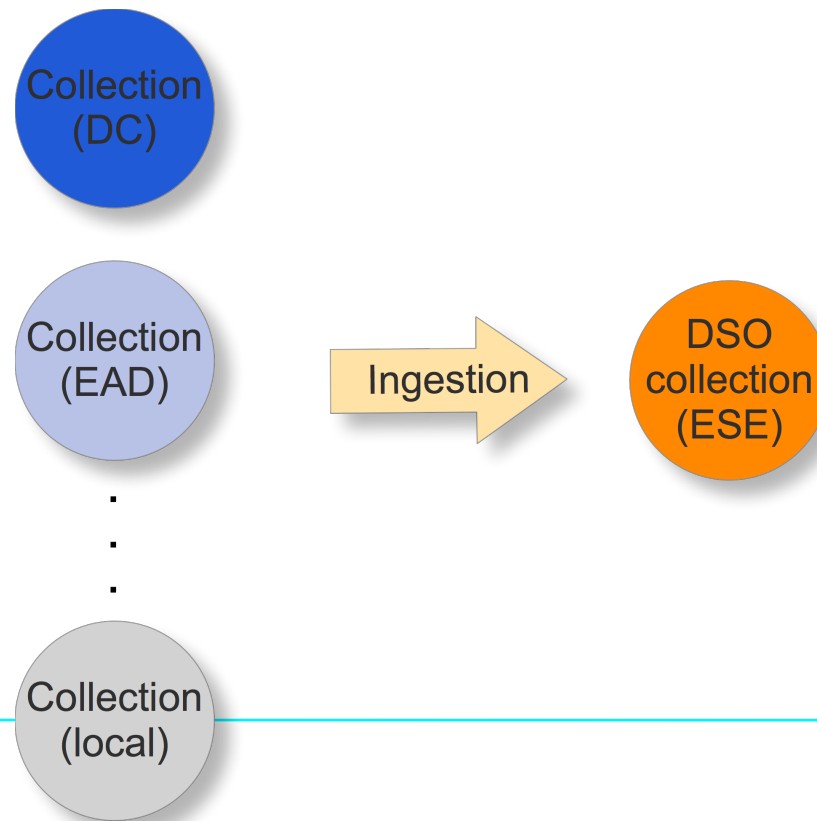
Europeana Semantic Elements

Source	Element	Refinement(s)
DC	title	alternative
DC	creator	
DC	subject	
DC	description	tableOfContents
DC	publisher	
DC	contributor	
DC	date	created; issued
DC	type	
DC	format	extent; medium
DC	identifier	
DC	source	
DC	relation	isVersionOf; hasVersion; isReplacedBy; replaces; isRequiredBy; ...

Source	Element	Refinement(s)
DC	coverage	spatial; temporal
DC	rights	
DC terms	provenance	
Europeana	relation	isShownBy; isShownAt
Europeana	userTag	
Europeana	unstored	
Europeana	object	
Europeana	language	
Europeana	provider	
Europeana	type	
Europeana	uri	
Europeana	year	
Europeana	hasObject	
Europeana	country	LIDA 2009

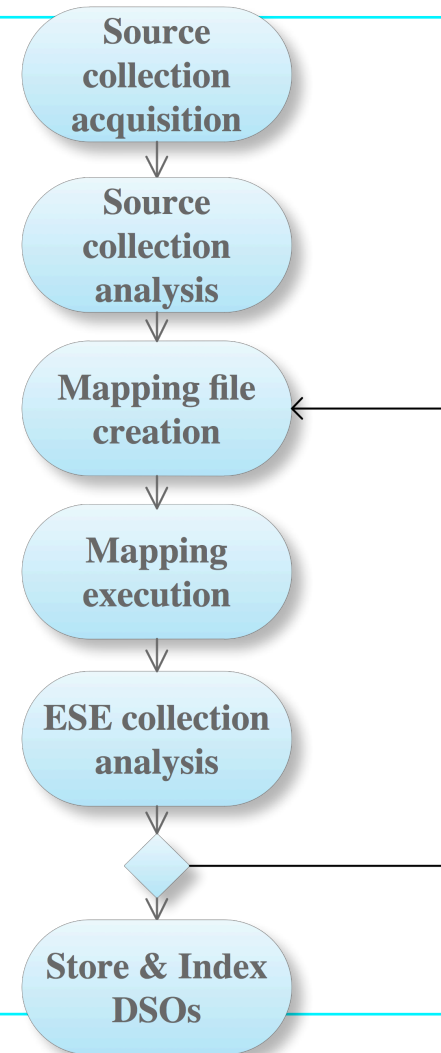
Europeana Semantic Elements (ESE)

- DSOs are created during the Europeana *data ingestion* process using the information provided by the content suppliers



EDLNet Data Ingestion Workflow

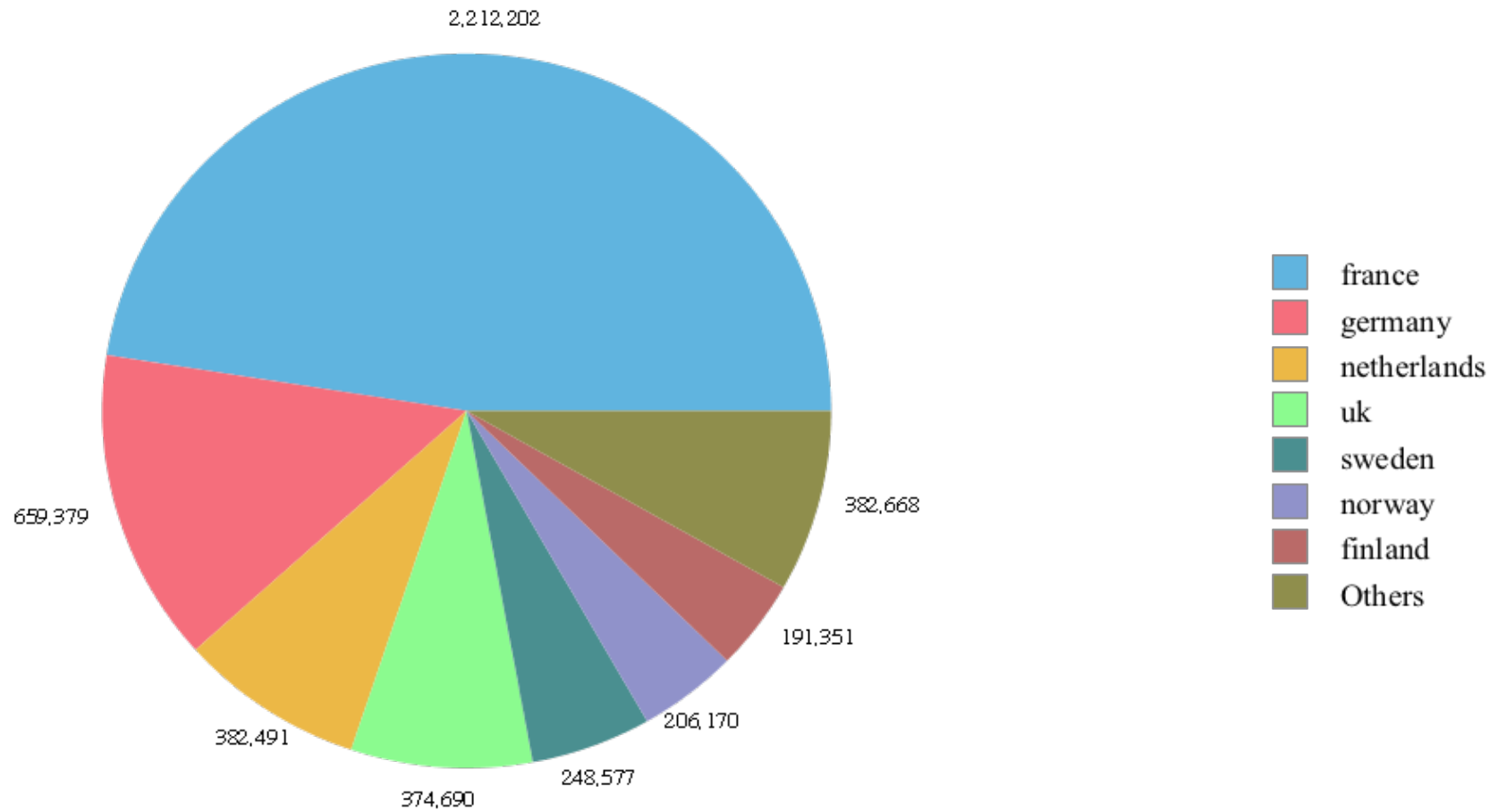
- Source collections are acquired via harvesting or received by content providers as XML files
- The *mapper* checks the source collections and creates the mapping rules
- XSLT is used to implement the mapping
- ESE collections are stored and indexed



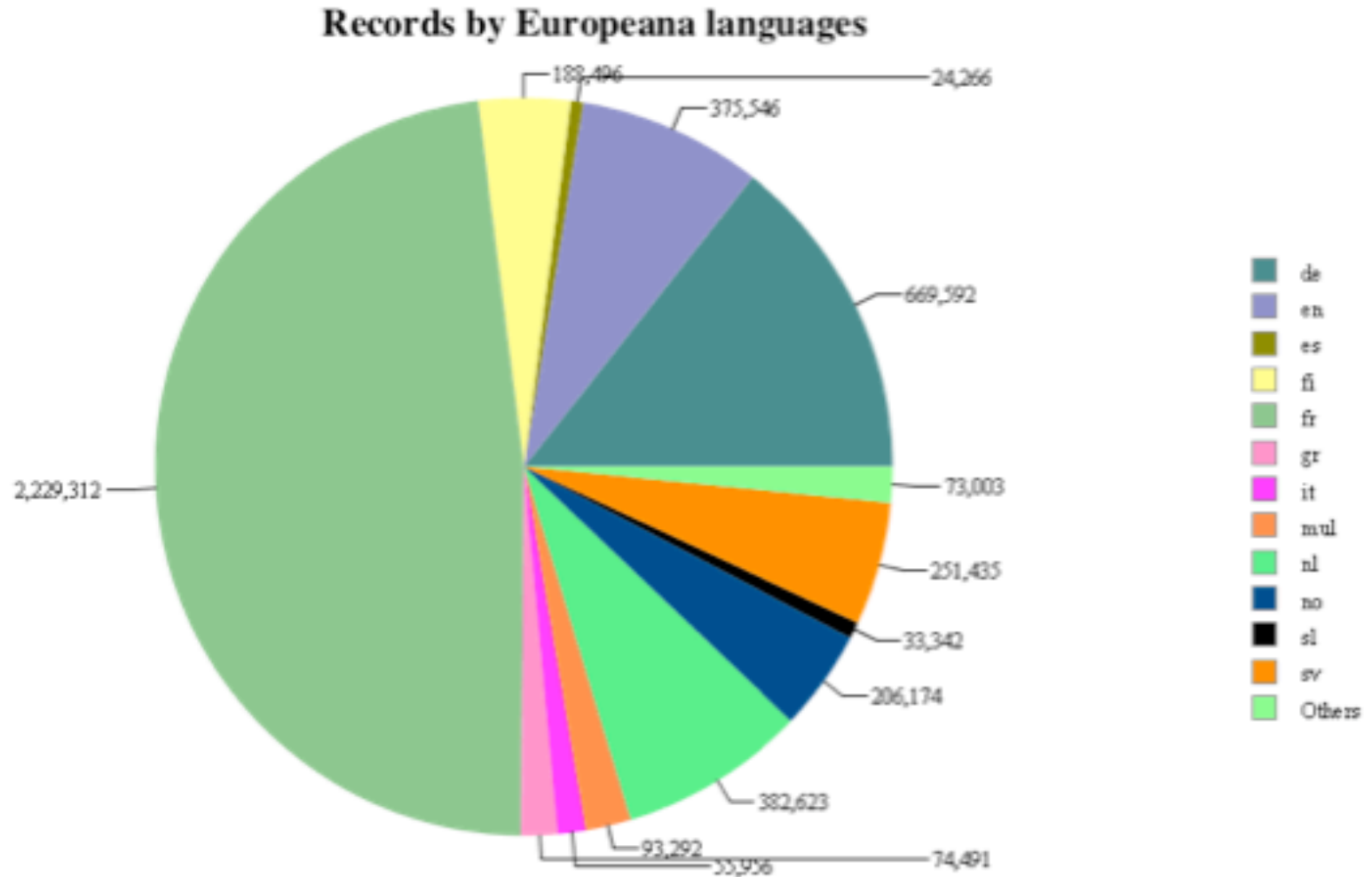
Europeana prototype data space

- The first Europeana prototype has been presented in November 08 as result of the EDLNet project
 - Next prototype will be released in September 09
- As of April 09, it contains DSOs referring information objects provided by 54 cultural institutions from 24 European countries
 - 4.5 milion of surrogate digital objects stored in the data space
- 15 different metadata formats

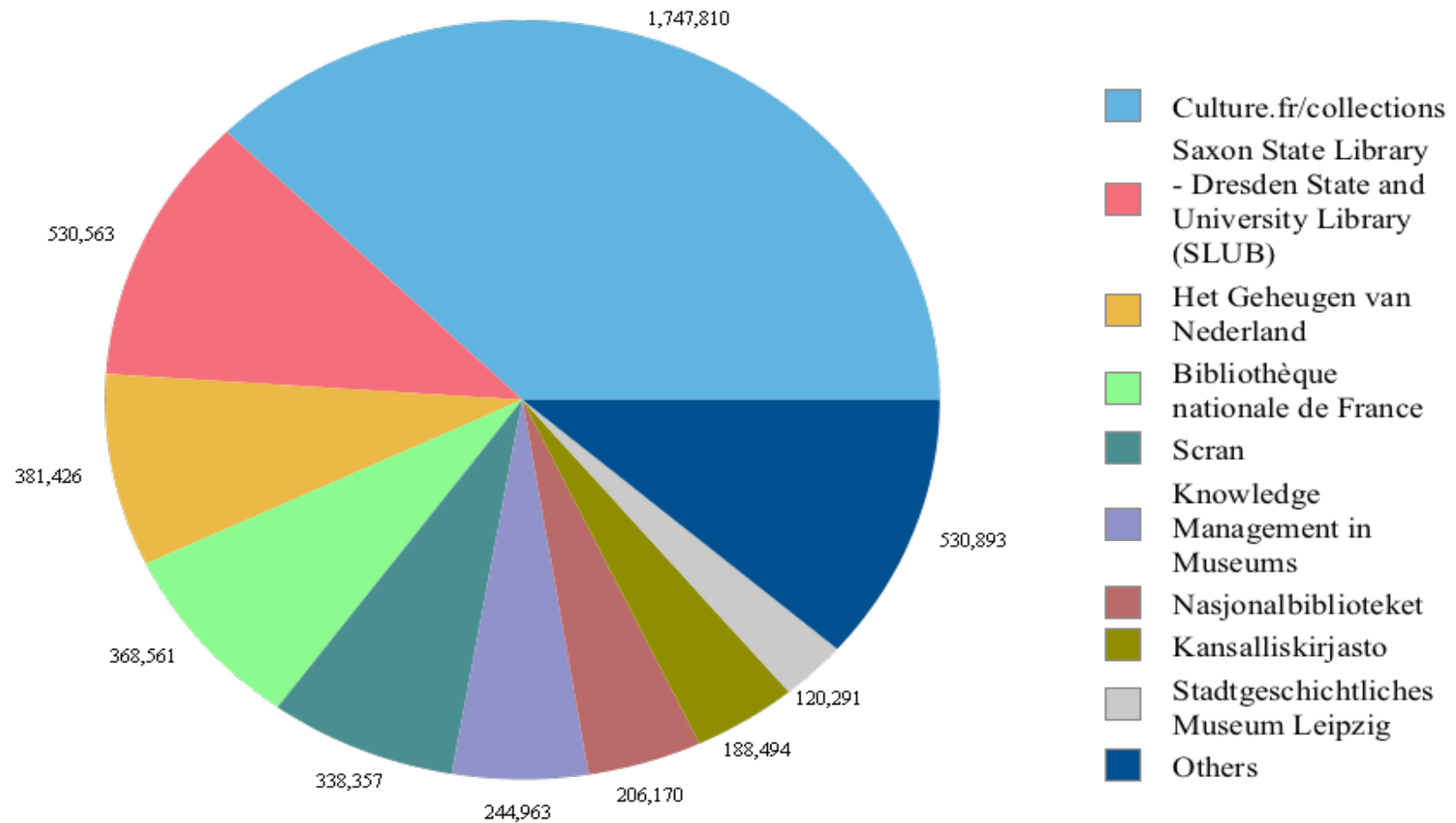
Heterogeneity: records by countries



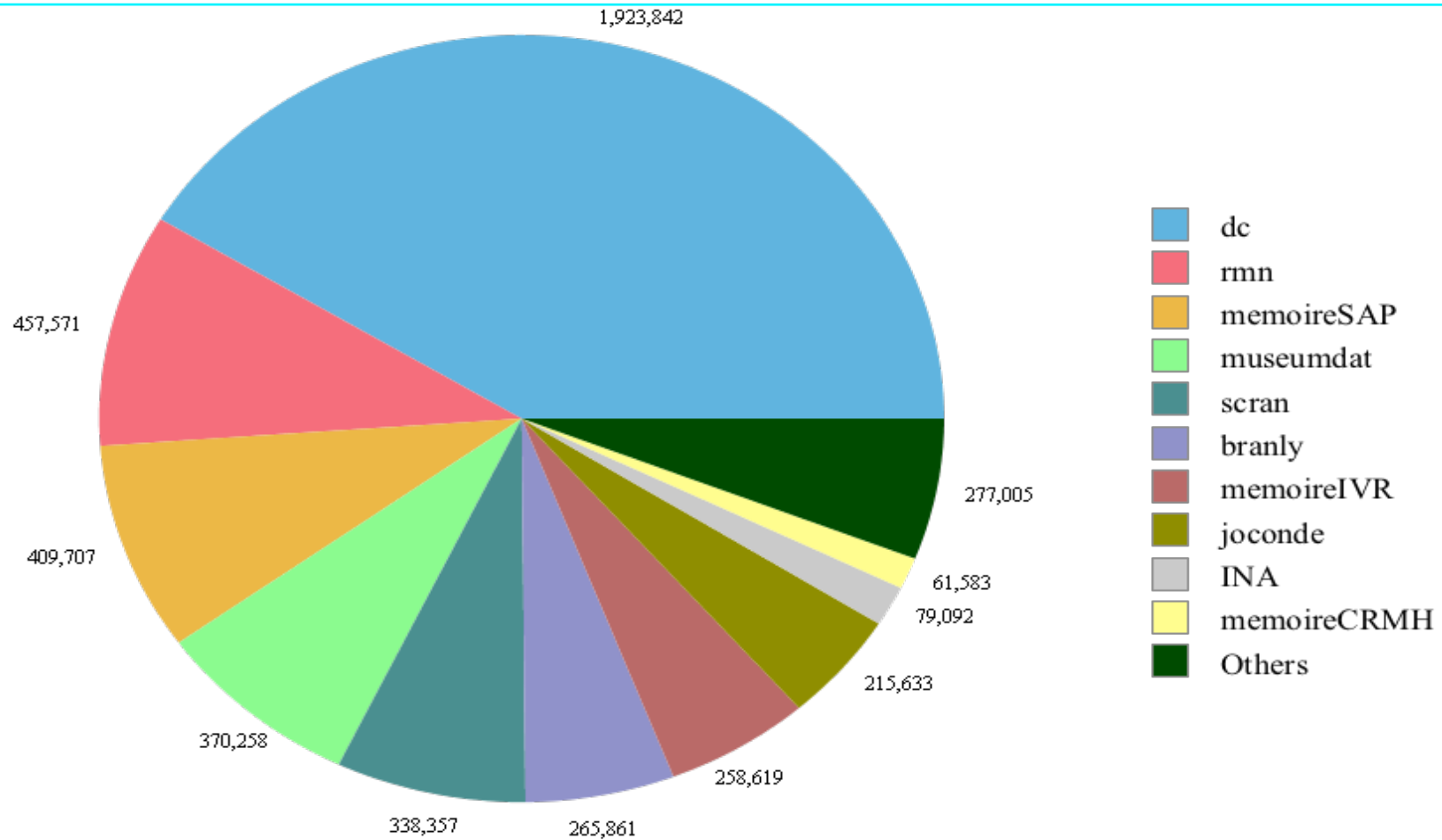
Heterogeneity: records by languages



Heterogeneity: records by data providers



Heterogeneity: records grouped by metadata format



Source collection snippet: DC

<ListRecords>

<record>

<dc:title>Από τα γλυκοχαράματα της ζωής μου: Σαλαμίσ**</dc:title>**

<dc:creator>Κ. Ν. Κωνσταντινίδης**</dc:creator>**

<dc:subject/>

<dc:description/>

<dc:publisher>Νέα Ζωή**</dc:publisher>**

<dc:contributor/>

<dc:date>1970-01-01**</dc:date>**

<dc:type>Articles**</dc:type>**

<dc:format>image/jpeg**</dc:format>**

<dc:identifier>http://xantho.lis.upatras.gr/kosmopolis/index.php/nea_zoi/article/view/313 **<dc:identifier>**

<dc:source>Νέα Ζωή; Vol 1, No 1 (1904); σελ. 07-08**</dc:source>**

<dc:language>gr**</dc:language>**

<dc:coverage/>

<dc:rights/>

</record>

Example: DC mapping

Source element	ESE element
dc:creator	dc:creator
dc:date	dc:date
dc:format	dc:format
dc:identifier	europaena:isShownAt
dc:language	dc:language, europaena:language
dc:publisher	dc:publisher
dc:source	dc:relation
dc:title	dc:title
dc:type	dc:type

Source collection snippet: MemoireSDAP

<BASE>

<NAME>Mémoire</NAME>

<DOMAINE>SDAP</DOMAINE>

<NOTICES>

<NOTICE ID="AP080050805330033">

<REF> AP080050805330033 </REF>

<ADRESSE> rue Petit ; villa "Rosario", ilot A</ADRESSE>

<AUTP>Richard, Fran√boise</AUTP>

<AUTOR>Delamotte, Patrick (architecte)</AUTOR>

<COM> Mers-les-Bains </COM>

<MCL>Secteur sauvegardé</MCL>

<LEG>cartouche céramique en relief façon "cuir" ; décor briques vernissées.</

LEG>

<COULEUR>OUI</COULEUR>

<LIB>Epoque 19ème</LIB>

<PAYS>France</PAYS>

<INSEE>80533</INSEE>

<TYPEIMG>JPG ; oui</TYPEIMG>

<TYPESUPP>DS1</TYPESUPP>

<REFIM>AP080_050805330033NUCA_P.JPG,DS1,,</REFIM>

<VIDEO>/Wave/image/memoire/1047/ap080_050805330033nuca_p.jpg;/Wave/

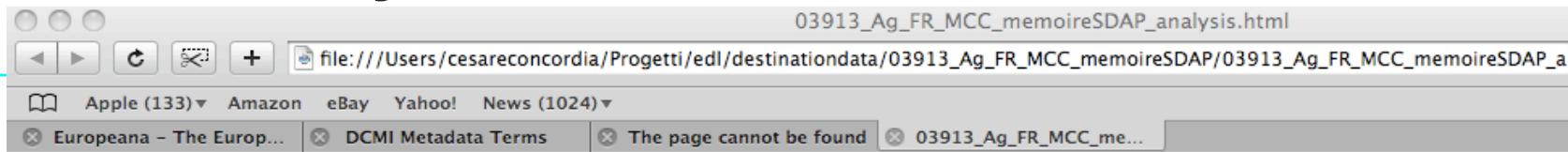
image/memoire/1047/ap080_050805330033nuca_v.jpg</VIDEO>



</NOTICE>

LIDA 2009

Analysis file of the MemoireSDAP collection



Europeana.eu analysis file for 03913_Ag_FR_MCC_memoireSDAP.xml

This is the list of properties. Click on it to see the top 100 values of each property. Click here to sh

/BASE/DOMAIN, total values: 1, All Unique

/BASE/NAME, total values: 1, All Unique

/BASE/NOTICES/NOTICE/ADRESSE, total values: 642

/BASE/NOTICES/NOTICE/AUTOR, total values: 148

Description	Value
Xpath:	/BASE/NOTICES/NOTICE/AUTOR
QName path:	BASE;NOTICES;NOTICE;AUTOR

Value	Occurrences of this value	Coverage records, %
Clémence	113	76
Patrick, Delamotte (Architecte)	30	20
Patrick, Delamotte (architecte)	4	2
Delamotte, Patrick (architecte)	1	0

LIDA 2009

/BASE/NOTICES/NOTICE/AUTP, total values: 645

Example: MemoireSDAP mapping

Source element	ESE Element	Comment
ADRESSE	europeana:unstored	
AUTOR	dc:creator	
AUTP	dc:contributor	
DENO	dc:subject	
DIFF	dc:rights	
DOM	europeana:unstored	
IDPROD	dc:source	
LEG	dc:description	
LIB	europeana:unstored	
LOCA	dc:description	
REF	europeana:isShownAt	Prefix the value with http://www.culture.gouv.fr/public/mistral/memoire_fr?ACTION=CHERCHER&
VIDEO	europeana:object_&& europeana:isShownBy	take the first value before the ";" and prefix it with 'http://www.culture.gouv.fr'

XSLT snippet

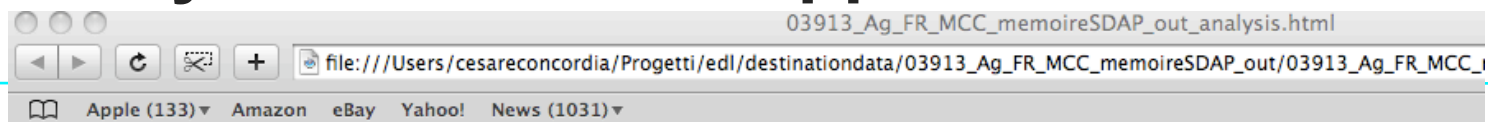
...

```
<xsl:template match="BASE">
  <metadata>
    <xsl:comment>europeana:type has 'image' value </xsl:comment>
    <xsl:apply-templates select="NOTICES/NOTICE"/>
  </metadata>
</xsl:template>
<xsl:template match="NOTICES/NOTICE">
  <record>
    <xsl:apply-templates select="DENO"/>
    <xsl:apply-templates select="IDPROD"/>
    <xsl:apply-templates select="ADRESSE"/>
    <xsl:apply-templates select="AUTOR"/>
    <xsl:apply-templates select="AUTTI"/>
    <xsl:apply-templates select="COM"/>
    <xsl:apply-templates select="VIDEO"/>
  </record>
</xsl:template>
</xsl:stylesheet>
```

XSLT Snippet

```
<xsl:template match="VIDEO">
  <europeana:isShownBy >
    <xsl:text>http://www.culture.gouv.fr</xsl:text><xsl:value-of select="substring-
before(.,';)"/>
  </europeana:isShownBy>
  <europeana:object >
    <xsl:text>http://www.culture.gouv.fr</xsl:text><xsl:value-of select="substring-
before(.,';)"/>
  </europeana:object>
</xsl:template>
```

Analysis file of the mapped ESE collection



Europeana.eu analysis file for 03913_Ag_FR_MCC_memoireSDAP_out.xml

This is the list of properties. Click on it to see the top 100 values of each property. Click he

/metadata/record/dc:contributor, total values: 645

/metadata/record/dc:creator, total values: 148

Description	Value
Xpath:	/metadata/record/dc:creator
QName path:	metadata;record;{http://purl.org/dc/elements/1.1/}creator

Value	Occurances of this value	Coverage records, %
Clémence	113	76
Patrick, Delamotte (Architecte)	30	20
Patrick, Delamotte (architecte)	4	2
Delamotte, Patrick (architecte)	1	0

/metadata/record/dc:description, total values: 645

/metadata/record/dc:rights, total values: 645

/metadata/record/dc:subject, total values: 1290

LIDA 2009

/metadata/record/dcterms:spatial, total values: 1935



Theory and Practice

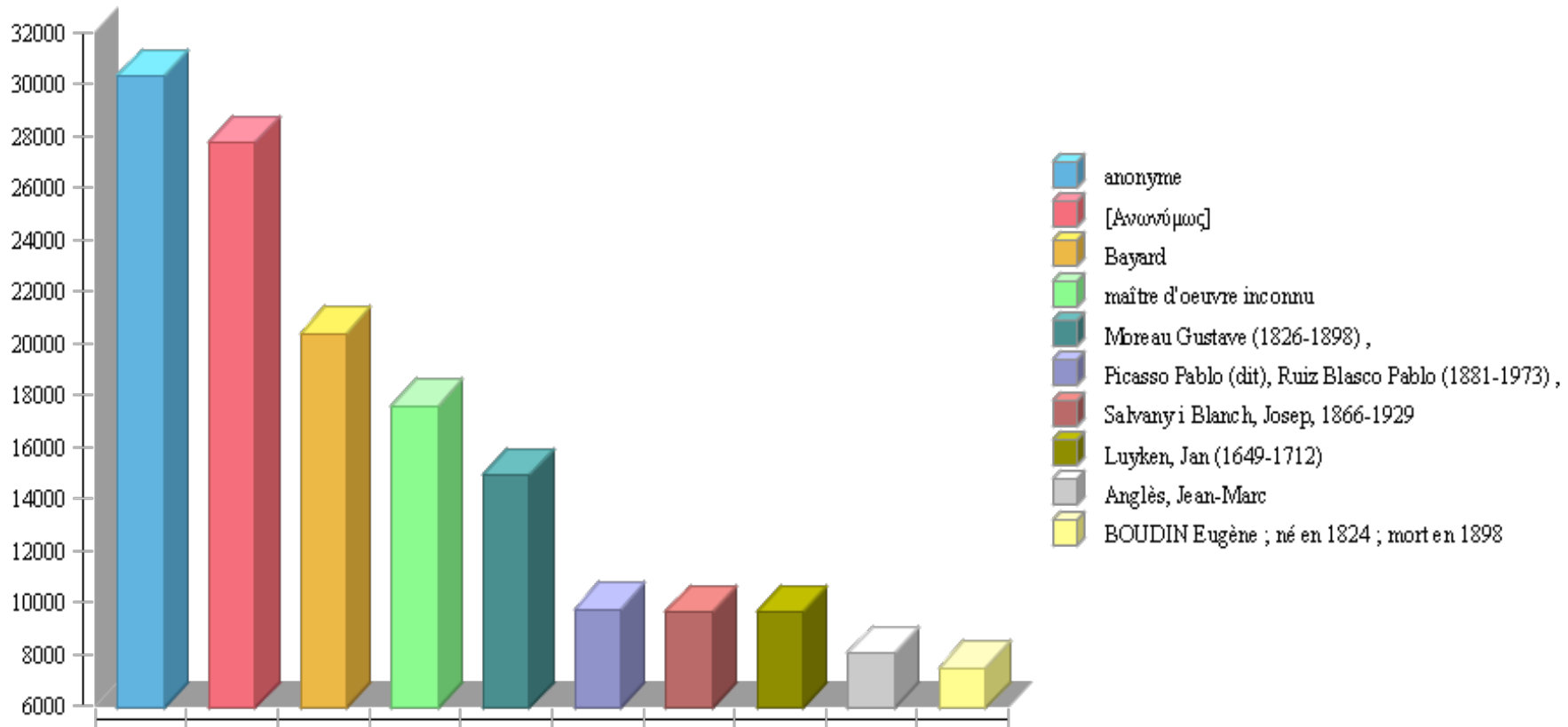
- Lot of *manual* work for writing the mapping rules and implement them
- Often mapping files cannot be reused
 - Same metadata element have different kind of values for different collections
- It is difficult distinguish metadata records describing DR, DP or RP objects
- Many content providers provide minimal metadata records, it is difficult to build significant DSOs
- Few relationships among digital objects in the metadata

Theory and Practice

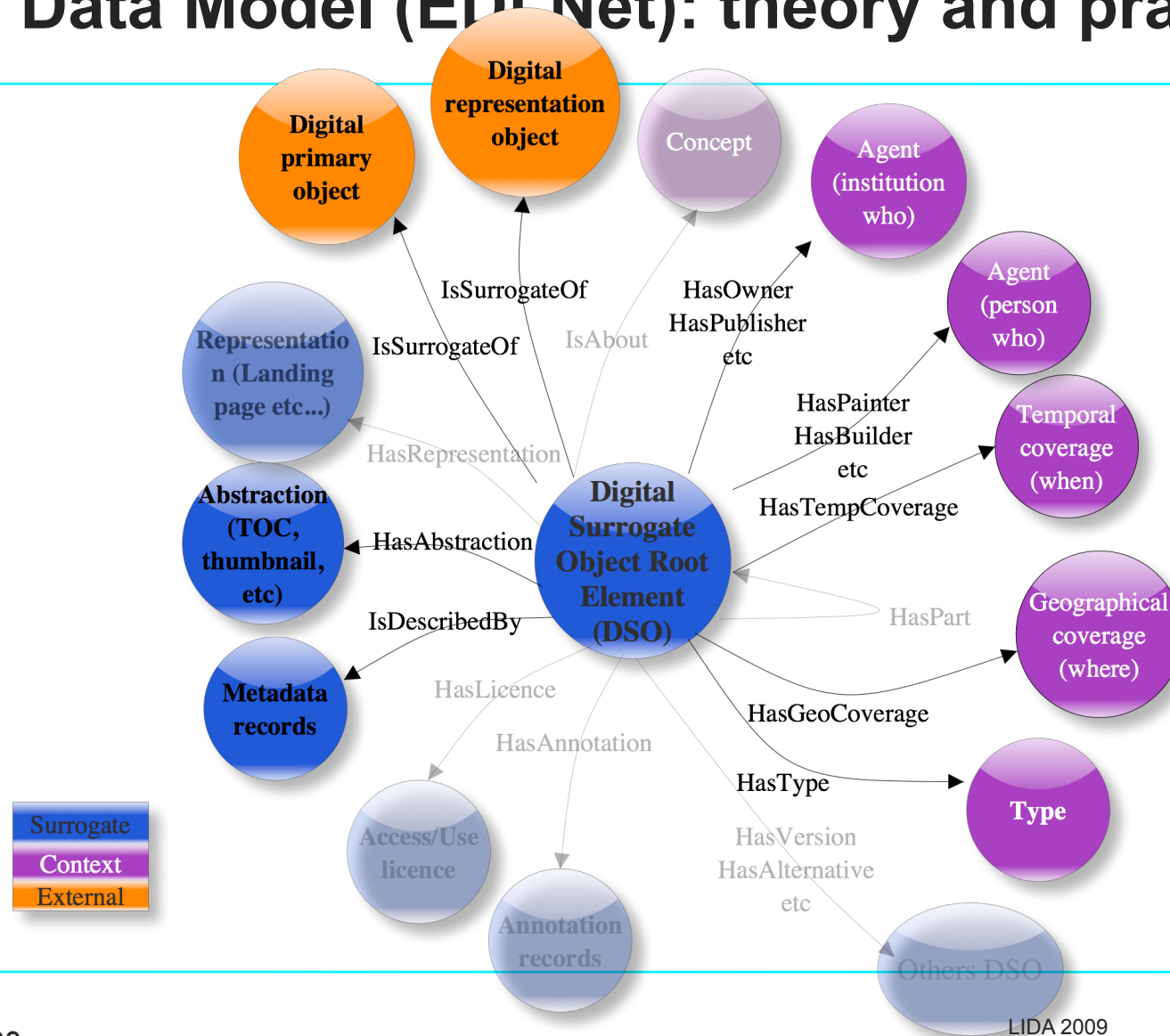
- Surrogates contextualization is a complex task
 - Implicit contextualization i.e. matching elements and attributes values with (few) classification schemes and/or authority files has been applied to several collections
 - Explicit contextualization i.e. using elements values directly linking to semantic resources (ex. IsAbout) has been in practice never applied
- Need to adopt authority files

Data normalization problem: example

'dc:creator' top ten



DSO Data Model (EDL Net): theory and practice



Ingestion: from EDLNet to Europeana

- The DSO data model is currently being reviewed, it should move from DC to CIDOC-CRM
 - Extracting and adding snippets (research)
 - Provenance
 - Events
 - Model adopted for Europeana will be released in September 09
- More involvement of content providers in the ingestion workflow (EuropeanaConnect)
- Extracting knowledge from the data space to contextualize and create relationships among DSOs (research)
- Try to make the mapping a semi-automatic process (research)
- Going open source



Acknowledgements

- The Digital Surrogate Object model has been defined by WP leaders of EDLNet Work Package 2: Makx Dekkers, Stefan Gradmann and Carlo Meghini and reviewed by EDLNet members
- The ESE model has been defined by Go Sugimoto (EDL Office) EDLNet Interoperability Manager and reviewed by EDLNet members.
- The xml-analyzer program has been developed by Gerald de Jong and Sjoerd Siebinga (EDL Office).

• Thanks