DELIVERABLE

Project Acronym:	**LoCloud**
Grant Agreement number:	**325099**
Project Title:	**Local content in a Europeana cloud**

D1.3: Content and metadata analysis

Revision: final

Authors:

Costis Dallas, Dimitris Gavrilis, Stavros Angelis, Dimitra Nefeli Makri and Eleni Afiontzi, Digital Curation Unit, Athena Research Centre

Contributors:

All partners

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

Revision History

Revision	Date	Author	Organisation	Description
V0.1	2013/09/15	Stavros Angelis, Nefeli Makri, Costis Dallas, Dimitris Gavrilis, Eleni Afiontzi	DCU	First Draft
V0.2	2013/09/30	Stavros Angelis, Nefeli Makri, Costis Dallas, Dimitris Gavrilis, Eleni Afiontzi	DCU	Additional partner contributions and final edit
V0.3	2013/10/04	Stavros Angelis, Nefeli Makri, Costis Dallas, Dimitris Gavrilis, Eleni Afiontzi	DCU	Kate Fernie and Costis Dallas comments

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

1. Executive Summary

This is the third deliverable within the LoCloud WP1: Planning, preparation and requirements and a part of Task 1.4: Content and metadata. During the planning stage of LoCloud, in association with content partners, the Athena Research Centre (DCU) evaluated and appraised content and metadata among collections participating in LoCloud. This includes new content contributed by content partners, content contributed by local institutions from the partner's regional networks and some content already ingested in the CARARE repository. The content was appraised and evaluated with regard to fitness-for-purpose, completeness and quality.

Section 2 describes the methodology followed to identify the content and metadata to be aggregated. In LoCloud content partners have two roles both as providers of content from their institution's collections and as regional or national aggregators of content from small institutions within their network. Content partners were asked to provide information about both their native collections and these third party collections by completing an online questionnaire survey. This survey aimed to identify information about existing collection management systems, native and third party collections, the objects contained in the collections, metadata schemas, vocabularies and thesauri, geographical information, metadata completeness, interoperability and rights related issues. The questionnaire was followed up by direct contact with partners by email. Three workshops were also organized provided a further opportunity to update and verify the information received from the questionnaire survey.

The results of the questionnaire survey and the content providers' workshops about the incoming content in LoCloud are presented in section 3. Content partners have identified a number of collections they want to contribute from their institution and are in the process of identifying additional third party collections from smaller institutions. These collections contain a diverse set of cultural objects that fall into one of the following categories: a) moveable objects (museum collections), b) immoveable objects (archaeological sites and historic buildings), c) library material and d) archival sources. Most partners will provide content from more than one collection. Several partners act as national aggregators and a few of these have already delivered content to Europeana directly. In other cases partners have identified content that they wish to collect and are in the process of creating collections or of contacting local institutions to invite them to participate. Less information was available about the third party collections at this stage in the project. The cultural objects contained in collections are mostly digitized images and text although there are a number of sound recordings, video and 3D objects. The metadata that will be submitted to LoCloud is mostly licensed under the CC0 license and under the Europeana Data Exchange Agreement.

Section 4 summarises findings on the metadata that content providers will be submitting for delivery to Europeana through LoCloud. The level of description and metadata vary amongst LoCloud partners from rich detailed descriptions to some content with no metadata descriptions yet available. An issue raised during the content providers' workshops is the need for a uniform way of describing items and a common metadata schema at national level. However, the survey revealed several different metadata schemas in use ranging from known standards (with extensions and local customisation) to local proprietory schemes. The completeness of metadata records and use of controlled vocabularies and thesauri were also found to vary widely. Geographic information is an important part of LoCloud content and about half the content partners support use of a standard geographic reference system.

One of LoCloud's objectives is to ensure interoperability between native content parters' metadata, the metadata stored in the aggregator repository and the metadata delivered to Europeana. In order to accommodate the diverse set of digitized cultural objects and metadata amongst partner collections an approach of mapping to a number of intermediate schemas has been proposed by LoCloud. The implications of the content and analysis, reported in Section 5, on the selection of intermediary schemas suggest that most providers can more easily deliver their metadata in CARARE, LIDO, EAD or a form of extended Dublin Core.

The aim of this content survey and metadata analysis has been to guide and inform planning of the aggregation strategy, to provide feedback for the selection of appropriate intermediary schemas to be used in metadata mapping in LoCloud, and provide input for the technical partners to the design and development of appropriate micro-services for LoCloud.

2. Methodology

In this section we describe the methodology followed to identify the content and metadata schemas of content providers, specific details about their content as well as key issues with their data. LoCloud aims to aggregate content from institutions with cultural content from across Europe and deliver this content to Europeana. In LoCloud the content partners will have two roles, i) to directly submit content from their institution's collections and ii) to act as national aggregators and collect content contributed from small institutions within their country. It became clear at an early point that we had to identify what both the directly submitted and contributed content consists of (digital object types, formats, quantity etc) in order to accommodate the specific characteristics of the content and pass on this knowledge to the technical partners in the form of requirements for the design and development of the technical infrastructure.

To this end the work has been divided in two complimentary steps. The first step was to conduct an online questionnaire survey. This survey was available for the content providers to complete from 06/08/2013 to 29/08/2013. In some cases further contact with content partners by email was needed in order to have a better description and more details about the provided information. The aim of the questionnaire survey was to capture information about the collections hosted by the content partners, as well as the content likely to be contributed by smaller providers. This questionnaire survey consisted of 28 questions with the purpose of identifying collections, metadata schemas, contributed collections, information about the quantity of digital resources and cultural objects, object types, language etc. A follow up of this questionnaire was a direct contact by email with providers asking for sample records from their collections.

The second step was the organization of three content providers' workshops with the aim to discuss further the content providers' content and metadata, identify needs and extract requirements. During these three workshops the results from the questionnaire survey were presented to the content providers and further discussion was made regarding details about their collections and content to be contributed by institutions within their networks.

2.1. Online questionnaire survey

An online questionnaire survey was conducted as part of the LoCloud project. The main aim of this survey was to evaluate and appraise content and metadata among collections participating in LoCloud with regard to fitness-for-purpose, completeness and quality. It was also to take stock of the information systems, schemas, and standards used with metadata that will be aggregated by the project. The key challenge was to find out about content and metadata not only in partners' own collection, but also about content and metadata that they plan to source from other contributing institutions and provide it to LoCloud as part of their content plan.

This survey questionnaire consisted of 28 questions, divided into particular sections. These sections referred to:

1. general information about the existing digital libraries or collections management system or any used software from partners' aspect
2. collections owned by partners, as for the type of collections and the used metadata schemas
3. collections contributed by other institutions, regarding the material that would be gathered by third parties

4. objects, focusing on the quantity of digital resources and cultural objects as well as the object types
5. metadata, concerning about the used metadata schemas and issues such as the XML validation, the mandatory elements of the schema or the unique elements used
6. vocabularies/ thesauri, taking into consideration the controlled vocabularies and the use of SKOSified vocabularies
7. geographical information, as for the existing used geographic coordinates systems and the historical place names
8. metadata completeness, regarding statistics about the elements of the metadata schema
9. interoperability, as for the compliance of the metadata with protocols and standards (OAI-PMH)
10. rights, based on the used licences and archival policies that each partner used for their own material.

24 partners responded to the questionnaire, providing useful information about their collections. This information was further updated with the discussions in the three workshops.

2.2. Content provider workshops

Three workshops were organised in Copenhagen, York and Madrid respectively. The content providers' workshops shared a common programme; the providers were divided into three groups to give a better opportunity for discussion and to record the content providers' views, and as a means of verifying the survey results and identify possible issues more efficiently.

The first session of the workshops, "LoCloud source content and metadata" mainly focused on identifying the content and metadata of content providers. In that session we received important information that verified, expanded and updated the survey results. The second session "Intermediate metadata schemas in LoCloud" allowed the content partners to further understand possible intermediary metadata schemas and reflect on their schemas in comparison to the proposed intermediate ones. An important issue is that during the content providers' workshops it became apparent that some providers had an unclear view about the collections to be contributed by small institutions within their networks.

3. Content

In LoCloud content providers have a variety of content, as is clearly shown from the following analysis of the survey results and the workshops feedback. This content has many differences and particularities depending on various factors like the institution that holds it, the country it originates from etc. We present here information about the partners' collections, third party contributed collections and the metadata contained therein.

3.1. Collections

Collections that will be available through LoCloud vary in size and content. Partners have different kind of collections and systems varying from museum, digital library and archival content to archaeological and local history content. Partners will provide both content from their home collections and collections contributed from smaller providers. While most partners have a clear view of what their native collections hold, as yet in some cases they are unclear what the content of the external collections is. From the initial online questionnaire survey we extracted some important results, which we verified and updated with the feedback we got from the partners during the content providers workshops.

The first interesting result is that more than half of the content providers will submit items to LoCloud that belong to more than one collection held by their institution. These different collections vary in object types, level of description and the metadata schema used for their description. Table 1 shows if the content partner will submit content belonging to one or more collections and a brief description of the collections content. From the following providers Zavad Jara stated in the Madrid workshop that since they are already a national aggregator and deliver content to Europeana regularly, there is an issue with the workflow that they will follow with the content they plan to deliver through LoCloud. There are two possible options, i) deliver content from their repository directly which could make it difficult to identify and count the LoCloud content, ii) deliver content through the LoCloud repository in which case care will need to be taken to avoid duplication of content in Europeana (the same content being delivered twice). Future Library stated that they are now gathering their content in order to create their own collections therefore they cannot provide details about their content yet, only about third party collections.

Content provider	One collection	More than one collection	Brief description
PSNC (Poland)	✓		Collections contains movies, oral history, pictures and multimedia content
KUAS (Denmark)	✓		
BJC (Romania)		✓	Local photos and documents, newspapers and local history books
RCE (Netherlands)		✓	Historical Cultural landscapes, Archaeological reports, Controlled vocabulary of Dutch archaology. Dutch East India Company (VOC) RCE archive, Several other collections, probably including post-war built heritage, shipwreck archives

NPU (Czech Republic)		✓	The State Archaeological List (SAL) of the Czech Republic Open and regularly updated information system of the State Archeological List of the Czech Republic, based on a digital map of archeological sites in the territory of the Czech Republic, interconnected with archeological sites database, including information on finds - both immovable and movable assets. Significant Archaeological Sites Database and map application contains the most significant archaeological sites in the Czech Republic from the point of view of their historical significance for archaeological heritage protection and conservation.
VUKF (Lithuania)		✓	
UoY ADS (United Kingdom)		✓	**Grey Literature Library**: The GLL, which consists primarily of PDF/A files of unpublished archaeological field reports, has already been published successfully in Europeana through the CARARE project. For LoCloud, we will provide an updated set of metadata, as the GLL has now grown by a further 3,000 reports. **ADS archived collections resource discovery metadata**: ADS will provide resource discovery metadata for all 450+ of its existing collections. **Proceedings of the Society of Antiquaries of Scotland (PSAS)**: ADS will provide metadata for the c. 4000 PSAS reports, dating from 1851 to the present. **Star Carr Archive**: ADS will provide metadata for around 2,500 artefacts (most with images, but not all) held in the following museums: British Museum Hull and East Riding Museum Museum of Archaeology and Anthropology, Cambridge Natural History Museum National Museum of Ireland Scarborough Museum Whitby Museum, Yorkshire Museum, York Wessex **Archaeology Image Archive**: ADS will provide metadata for this collection totaling about 300 images from the following small museums/county archives: Salisbury and South Wiltshire Museum Wiltshire Heritage Hampshire County Council Wiltshire Council **Southampton Museum's Archive**: The ADS will provide metadata for these collections, totaling about 424 images (and reports in PDF, CAD plans in DXF and a variety of other file types) from the Southampton City Council. ADS has convened its planning group to locate other small to medium sized organizations who might be interested in participating LoCloud. It is likely that these organizations will provide data that can be easily aligned to the ADS schema, and therefore to the CARARE schema.
IPCHS (Slovenia)	✓		
Provincie Limburg		✓	

(Belgium)			
CG33 (France)			the content is mainly archival, documents, textual, postcards, maps and cards
Zavad Jara (Slovenia)	✓		KAMRA is a digital library available to all Slovenian local CH institutions. It contains over 170 collections, related to local history, contributed by various organisations.
Future Library (Greece)			
FMNF (Portugal)		✓	Archive collection - Textual digitalized documents and photographs about the beginning of railway in Portugal; Museum collection - Photographs of artifacts, trains, locomotives, buildings etc. about the beginning of railway in Portugal.
AIT (Austria)	✓		Heterogeneous (archives, images, library materials, numismatic, archaeological images, theatre texts, performance
ABMR (Sweden)		✓	**Birgittamuseet** - Collections from a local museum of medicine history **Landsarkivet** - Parchment and paper collection of letters **Ånges fotosamling** - Collection of photographs from Ånge municipality **Kubikenborgs skolas intresseförening** - Photo Collection related to the school of Kubikenborg
PSRL (Bulgaria)	✓		
BGB (Serbia)	✓		old photos and articles (newspapers and magazines)
HU (Turkey)		✓	Vekam Archival Collection separated from Library Collection
CUT (Cyprus)	✓		
AHAI (Iceland)		✓	1. Fornleifarannsóknir á Íslandi /Excavations in Iceland 1870-present Other possible collections are: 2. Designated grave-marks in Iceland (gravestones, crosses 100 years and older Part of CARARE project but could be submitted again: 3. Listed houses in Iceland 4. Listed archaeological sites in Iceland
PrifUK KAEG (Slovakia)	✓		Geophysical images of buried archaeological structures (buildings, chapels, etc) are included.
DP (Ireland)		✓	**Leo Swan Aerial photo collection**: aerial archaeology images taken over 20 years by the archaeologists **Leo Swan Discovery programme image collection**: range of graphic content including surveys, pans, illustrations and photographs generated over the last 21 years. **Monastic Ireland**: Images and photographs from Irish Monasteries
FRS (Italy)		✓	The collections, part of which, at any one time, are on display in the house museum, are kept in a state-of-the-art storage area. They include paintings from 16th to 20th c., some of considerable artistic importance; over 600 pieces of porcelain by Italian and European

			manufacturers, including a precious 18th century table service from the Florentine manufacturer Ginori; about 2,800 engravings of various subjects and 180 drawing dating to 16th and 20th c.; over 600 embroidered textiles produced by the School of Embroidery founded in 1904 by Romeyne Robert Ranieri di Sorbello. There are also about 3000 historic photos, including photos of local monuments and historical events, as well as 130 maps, ranging from the 17° to the 20° c. The Palazzo also hosts an important old family library, initiated in the late 18th century, which, through continuous acquisitions, now includes about 30.000 items, including 1500 ancient volumes, e.g. the Spaera Mundi manuscript from the 15th century and a 1770 edition of the French Encyclopaedia. It also has an original library catalogue from 1802, accessible now in digital form (complete of images and metadata MAG, with Dublin Core set of metadata elements) from its own website http://catalogo1802.wordpress.com/.

Table 1. Collections

3.2. Contributed collections

More than half of the providers plan to submit items that belong to collections contributed by other institutions/sources. At this moment the actual size of this contributed content is unclear, as most partners are in the point where they make contact with smaller providers. Most partners have an idea about the object types of the contributed collections (mostly text and photographs), and about the level of description (low to none). Table 2 show if the contributed content that will be submitted belongs to one or more collections and a brief description of the collections content.

Content provider	**One collection**	**More than one collection**	**Brief description**
Norsk Kulturrad (Norway)		✓	
PSNC (Poland)	✓		collections from very small institutions with city structures and oral history
MECD (Spain)		✓	a digital network in Spain with 20-25 collections
KUAS (Denmark)		✓	100 museums use Regin. There are two main categories: art museums and local history museums.
BJC (Romania)		✓	Library documents from County Public Libraries, Archive documents from memorial house
RCE (Netherlands)		✓	Several local/regional museums and heritage organizations We are discussing content to be delivered by several CH organisations
NPU (Czech	✓		

Republic)			
VUKF (Lithuania)		✓	The Atlas of Lithuanian hill forts and castles consist of two digital collections: 1. Collection of hill forts – database of Lithuanian hill forts (texts, geodetic data, digitised and digital photos, aero photos, etc.). 2. Collection of castles and fortified sites – database of Lithuanian castles and fortified sites (texts, geodetic data, digitised and digital photos, aero photos, etc.) Both are owned by Society of Lithuanian Archaeology.
UoY ADS (United Kingdom)		✓	In addition to the different collections we hold outlined on the previous page, we have convened our planning group and contacted several other potential contributors as a result. While nothing has yet been decided, we assume that any outside collections will be easily mapped to the CARARE schema.
IPCHS (Slovenia)		✓	
Provincie Limburg (Belgium)			Needs to be determined on the basis of what can be offered.
CG33 (France)		✓	Local history society archive, museum, local environmental and historical preservation association
Zavad Jara (Slovenia)		✓	There are 170 digital collections contributed by over 50 partners, local cultural heritage organisations. The numbers are constantly growing. Partners are mostly public libraries, but also museums, archives, local associations, schools. In September we will add new module to create user generated content.
Future Library (Greece)		✓	Our organisation (Future Library, Greece) is coordinating a group of 117 Greek public and municipal libraries. Out of these libraries, at least 9 will have soon a collection of digital material of local content, coming from their media laboratories: Drama, Kozani, Livadia, Korinthos, Nafpaktos, Keratsini-Drapetsona, Athens, Ilioupoli, Athens, Chania. The collections will include digital stories (video and audio), local pictures of cultural and historical value, local texts of cultural and historical value.
FMNF (Portugal)		✓	Archive collection – Textual digitalized documents and photographs about the beginning of railway in Portugal; Museum collection – Photographs of artefacts, trains, locomotives, buildings etc. about

			the beginning of railway in Portugal.
AIT (Austria)		✓	University of Graz: Archeological Collections at the University of Graz Hugo Montfort Digital Edition Numismatic Collection at the University of Graz Visual Art of South-Eastern Europe Don Juan Archive Vienna: theatre related texts
ABMR (Sweden)		✓	Birgittamuseet – Collections from a local museum of medicine history. (Local history society) Landsarkivet – Parchment and paper collection of letters. (Local archive) Ånges fotosamling – Collection of photographs from Ånge municipality. (Local history society) Kubikenborgs skolas intresseförening – Photo Collection related to the school of Kubikenborg. (Local history society)
PSRL (Bulgaria)	✓		
BGB (Serbia)		✓	
HU (Turkey)	✓		
CUT (Cyprus)		✓	Images and Books which belongs to the local archive of the Limassol Municipality – 3D Icons which belongs to the Church of Cyprus – Audio-visual aerial which belongs to the CyBC (the only state CY Radio-TV Station)
AHAI (Iceland)		✓	Excavation Field Data. Collection that belongs to the Museum of Skagafjörður. Field data from 50 different places in the municipality of Skagafjörður
DP (Ireland)		✓	George Victor Du Noyer Collection: provided by the Royal Society of Antiquities Ireland (RSAI) this is the digitised collection of antiquarian drawings and paintings from 1834-1868. (Antiquarian Society) RSAI lantern slide collection: scanned images from lantern slides captured between 1891 and 1926 (Antiquarian Society) Dublin Institute of Advanced Studies (DIAS) Ogham collection.: textural, images and 3d models of organ stones around Ireland (Institute)
FRS (Italy)	✓		

Table 2. Contributed Collections

3.3 Cultural objects

The items (digital resources, cultural objects) that will be submitted to LoCloud are possible to belong to various categories. Table 3 presents the estimated number of items a content provider will submit per category. These categories include:

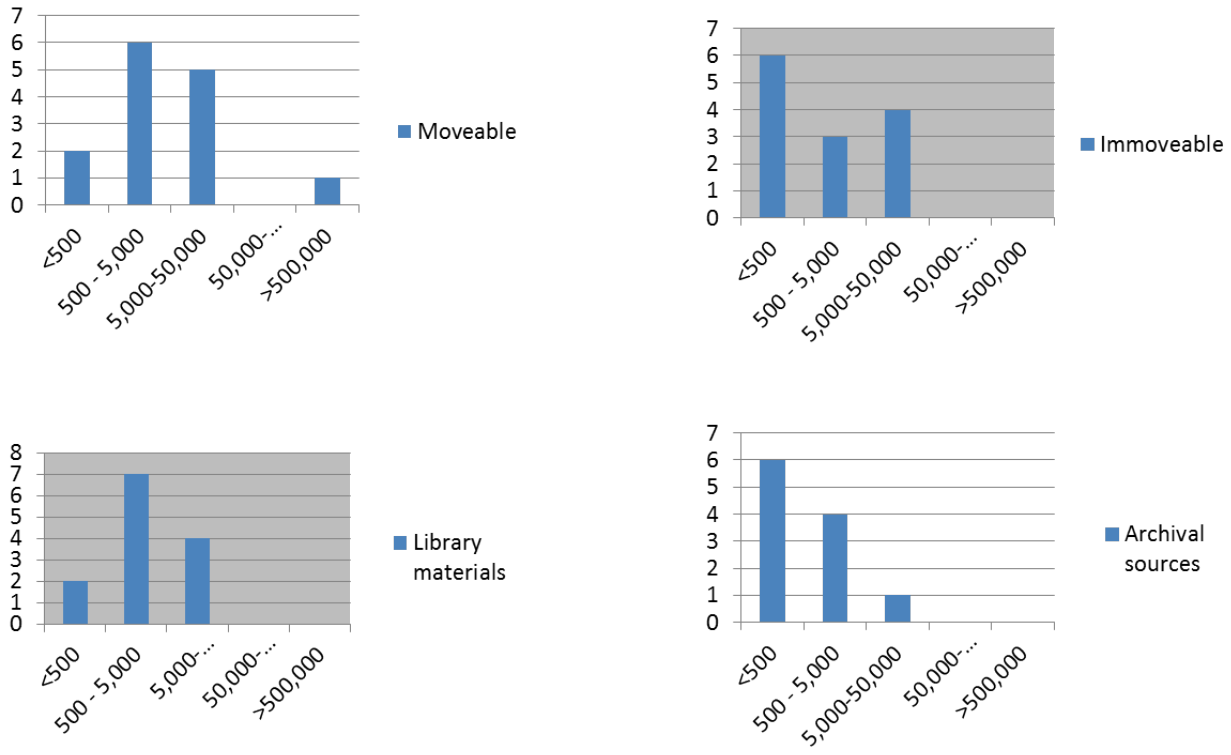
i) moveable cultural objects (artefacts, museum objects, artworks etc.)
ii) immoveable cultural objects (monuments, buildings etc.)

iii) library materials (digital library assets, photographs etc.)
iv) archival sources

Content provider	Moveable	Immoveable	Library materials	Archival sources
PSNC (Poland)			5.000-50.000	
KUAS (Denmark)	>500.000			<500
BJC (Romania)			500-5.000	
RCE (Netherlands)	5.000-50.000	5.000-50.000	5.000-50.000	500-5.000
NPU (Czech Republic)		500-5.000		
VUKF (Lithuania)	5.000-50.000	5.000-50.000	500-5.000	
UoY ADS (United Kingdom)	500-5.000	<500		5.000-50.000
IPCHS (Slovenia)	5.000-50.000			
Zavad Jara (Slovenia)	<500	<500	500-5.000	<500
Future Library (Greece)	500-5.000	500-5.000	500-5.000	
FMNF (Portugal)	500-5.000	<500	<500	500-5.000
AIT (Austria)	5.000-50.000		5.000-50.000	<500
ABMR (Sweden)	5.000-50.000			<500
PSRL (Bulgaria)	500-5.000	<500	500-5.000	<500
BGB (Serbia)			500-5.000	
HU (Turkey)				500-5.000
CUT (Cyprus)	<500	<500	<500	<500
AHAI (Iceland)		<500		
PrifUK KAEG (Slovakia)		500-5.000		
DP (Ireland)	500-5.000	5.000-50.000		
FRS (Italy)	500-5.000		500-5.000	

Table 3. Categories of items

The following graph depicts an approximate estimation of the number of items that will be submitted through LoCloud per item category. The x-axis shows the number of items and the y-axis the number of providers.



Graph 1. Approximate number of items per category

With regard to resource type, the items (digital resources, cultural objects) submitted to LoCloud can be categorized as:

i) sound
ii) image
iii) text
iv) video
v) digital 3D representation / model

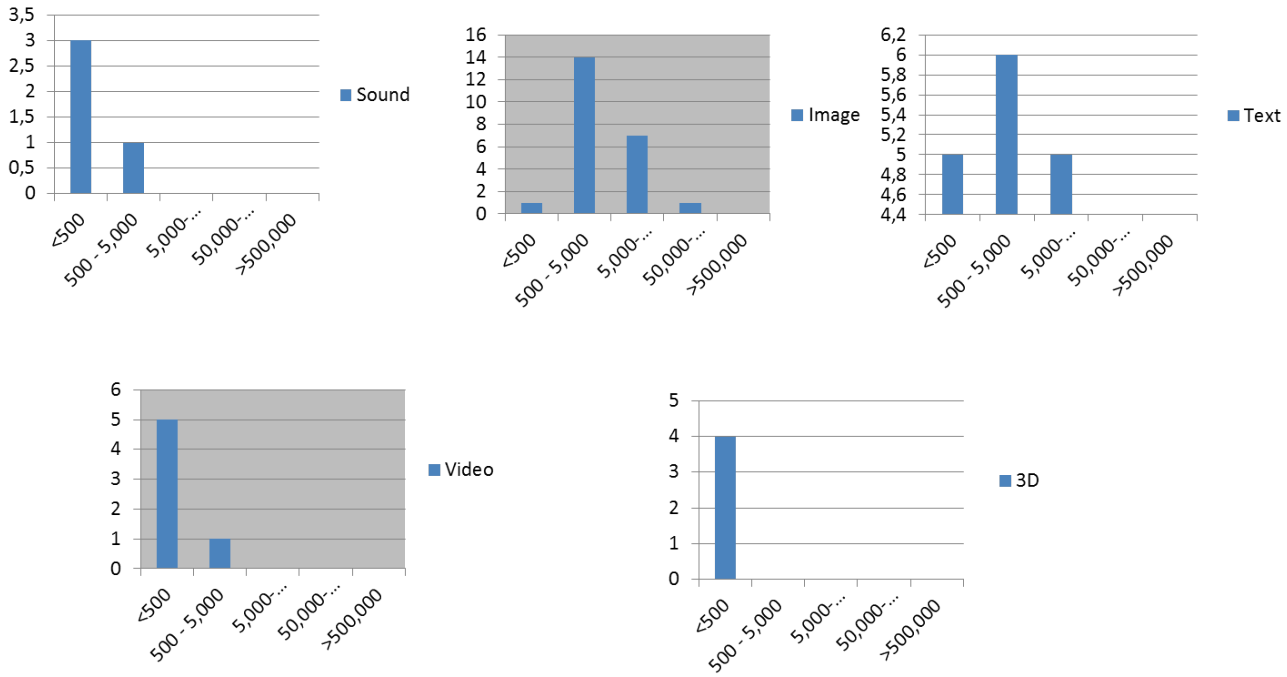
The following table presents the estimated number of item file types for each category per content partner (Table 4).

Content provider	Sound	Image	Text	Video	3D
PSNC (Poland)	500-5.000	5.000-50.000		500-5.000	
KUAS (Denmark)		50.000-500.000	<500	<500	
BJC (Romania)		<500	<500		
RCE (Netherlands)		5.000-50.000	5.000-50.000	<500	<500
NPU (Czech Republic)		500-5.000	<500		
VUKF (Lithuania)		5.000-50.000	5.000-50.000		
UoY ADS (United Kingdom)		5.000-50.000	5.000-50.000		<500
IPCHS (Slovenia)		500-5.000	500-5.000		

Provincie Limburg (Belgium)		500-5.000			
Zavad Zara (Slovenia)	<500	500-5.000	<500	<500	
Future Library (Greece)	<500	500-5.000	500-5.000	<500	
CG33 (France)					
FMNF (Portugal)		500-5.000	500-5.000		
AIT (Austria)		5.000-50.000	5.000-50.000		
ABMR (Sweden)		5.000-50.000	<500		
PSRL (Bulgaria)		500-5.000	500-5.000		
BGB (Serbia)	<500	500-5.000	500-5.000		
HU (Turkey)		500-5.000			
CUT (Cyprus)		5.000-50.000	500-5.000	<500	<500
AHAI (Iceland)		500-5.000	<500		
PrifUK KAEG (Slovakia)		500-5.000			
DP (Ireland)		500-5.000	<500		<500
FRS (Italy)		500-5.000			

Table 4. Item file types

The following graph depicts the approximate estimation of the number of item file types to be submitted through LoCloud (Graph 2). The x-axis indicates the number of items while the y-axis the number of providers.



Graph 2. Approximate number of item files types

Digital objects that are going to be submitted to LoCloud are complex objects and consist of one or more datastreams.

Table 5 shows the different datastreams an object may consist of. These datastreams include:

i) XML Metadata
ii) thumbnail images (JPG, PNG)
iii) full images (JPG, PNG, TIFF)
iv) text (PDF, DOC, PDF/A)
v) sound (MP3)
vi) videos (FLV, MPG4, WAV, AVI)
vii) database exports (MDB)
viii) geospatial vector files (SHAPEFILE)
ix) 3D items (3DPDF, CAD)

Content provider	Object datastreams
PSNC (Poland)	Metadata (XML, RDF, bibtex), thumbnais & images (JPG), sound (MP3), video (FLV)
KUAS (Denmark)	XML metadata, images
BJC (Romania)	text (DOC, PDF) thumbnails, images
RCE (Netherlands)	XML metadata, text (PDF), databases exports (MDB), images (JPG, PNG), videos (MPG4)
NPU (Czech Republic)	XML metadata, thumbnails, images
VUKF (Lithuania)	XML metadata, images
UoY ADS (United Kingdom)	XML metadata
Provincie Limburg (Belgium)	metadata, thumbnails, images, text (PDF)
CG33 (France)	XML metadata for the AD33 collections and digital objects for the partners. Quantities have to be defined.
Zavad Zara (Slovenia)	XML Metadata
Future Library (Greece)	sound (WAV, MP3) video (AVI), images (TIFF, JPEG), text (PDF)
FMNF (Portugal)	XML metadata
AIT (Austria)	XML metadata

ABMR (Sweden)	XML metadata, text (PDF), thumbnails, images
PSRL (Bulgaria)	XML Metadata, thumbnails, images, text (PDF)
BGB (Serbia)	Sound (mp3 file, XML metadata), image (pdf file, XML metadata), text (pdf file, XML metadata)
HU (Turkey)	images
CUT (Cyprus)	images, XML metadata
AHAI (Iceland)	XML metadata, text (DOC, PDF), geospatial vector files (SHAPEFILE), images, thumbnails
PrifUK KAEG (Slovakia)	images
DP (Ireland)	thumbnails, images, XML metadata, 3D (3DPDF)
FRS (Italy)	XML metadata, image

Table 5. Object datastreams

3.4 Rights

Most providers have no licence issues about the metadata that will be submitted in LoCLoud. In fact 21 out of 23 partners have their metadata openly accessible through one of the Creative Commons licenses (Table 9).

Content Provider	**Creative Commons License**
PSNC (Poland)	CC0 should be possible
KUAS (Denmark)	Not yet but we plan to make them available under CC0 license
BJC (Romania)	Public Domain
RCE (Netherlands)	CC0
NPU (Czech Republic)	Europeana Exchange Agreement
VUKF (Lithuania)	Available through Europeana Exchange Agreement
UoY ADS (United Kingdom)	CC0
IPCHS (Slovenia)	Free for noncommercial use
Provincie Limburg (Belgium)	Europeana DEA

CG33 (France)	CC0
Zavad Zara (Slovenia)	CC0
Future Library (Greece)	It would be possible to make the metadata available using CC0 license
AIT (Austria)	DEA
ABMR (Sweden)	CC0 and Europeana DEA
PSRL (Bulgaria)	Europeana DEA
BGB (Serbia)	CC0
HU (Turkey)	Planning to use Creative Commons licenses
CUT (Cyprus)	CC0
AHAI (Iceland)	The metadata is currently not openply accessible but we plan to have it open
DP (Ireland)	Discovery Programme & RSAI: CC0 through the European Exchange Agreement DIAS still to sign up to the EEA for CC0
FRS (Italy)	Having submitted the declaration of DEA, we will allow the publications of the metadata of our collections making them available through Creative Commons CC0.

Table 9. Creative Commons licenses

There are partners that provide access to only a part of their items and serve the end user with a thumbnail and a short description (e.g HU). These providers follow a pay-as-you-go model in order to provide full access to their content.

Only 3 partners do not allow open access to their cultural assets. Open access is given to some or all parts of objects and in a specific form (Table 10).

Content Provider	**Open Access**
PSNC (Poland)	Thumbnails and metadata are openly available for all objects planned for submission.
KUAS (Denmark)	Open access to documentary photo. Not images of art works.
BJC (Romania)	We allow open access to all our digital objects
RCE (Netherlands)	Archive, historic information, maps, Thumbnails of images, images (800x800), documents
VUKF (Lithuania)	Metadata – Full Access, Images – Free Access
CG33 (France)	Thumbnails of images
Zavad Zara (Slovenia)	Access is for majority of object allowed under CC BY-NC or Europeana Right Reserved-Free Access licenses.
FMNF (Portugal)	Non-commercial share-alike (from CC)
AIT (Austria)	Thumbnails and images
ABMR (Sweden)	Each collection has its own cc-license, mainly cc by-sa.
PSRL (Bulgaria)	Thumbnails
HU (Turkey)	Access to thumbnails of images and catalogue entries
CUT (Cyprus)	To all the content
AHAI (Iceland)	All files where author has given permission to published

PrifUK KAEG (Slovakia)	Access to full images
DP (Ireland)	All content will be available except for RSAI data. For RSAI donated data (Lantern Slides collection & Du Noyer watercolour collection) thumbnails will be provided under CC0. Full access to high resolution images will be available through additional license
FRS (Italy)	In some cases we will allow open access to full images, in other ones only the thumbnails of the images.

Table 10. Open access

4. Metadata

This section of the reports summarizes our findings on the medatata content providers will be submitting for delivery to Europeana through LoCloud.

4.1 Schemas, extensions and object identity

The level of description and metadata vary throughout the LoCloud partners. Collections have a different level of description, from rich detailed descriptions, medium descriptions in extended Dublin Core, to no descriptions (e.g. photographic collections from small providers). An issue raised during the content providers' workshops is that there is a need for a uniform way of describing items and a common metadata schema at a national level and there were content partners that viewed the LoCloud project as an opportunity to work towards that goal, e.g. Discovery Programme.

The online questionnaire survey indicated that approximately half content partners describe all their collections using a metadata schema. The workshops showed that most providers use a standard metadata description. Only a couple have native schemas and store their information in various databases. In these cases the structure of the database is unclear. 8 out of 24 providers are aware of specific metadata schemas in order to describe objects in these collections. We identified the following descriptive or metadata schemas used for metadata submitted to LoCloud that we identified:

i) Dublin Core
ii) Extended Dublin Core
iii) SPECTRUM
iv) EDM
v) ESE
vi) CARARE
vii) LIDO
viii) SKOS
ix) EAD
x) TEI
xi) other local schemas (MAG, ARUODAI)

The metadata schemas mentioned above apply to both native collections and third party collections. Most content partners have a clear view about the metadata schema used to describe items in their native collections but are unclear about the metadata available for third party collections. Only minimum information is available at this point of the project about metadata schemas used to describe contributed collections as content partners are still in the point of connecting with smaller providers that are interested to participate. This results to content partners having in some cases a sense about the level of metadata description in third party collections, but many content partners indicated that many third party collections will have minimum to none metadata description. The identified metadata schemas are presented in the following table (Table 6).

Content Provider	Metadata schema
Norsk Kulturrad	EAD
PSNC (Poland)	slightly customized Dublin Core, flat native schema
MECD (Spain)	MARC21 exports to EDM, Dublin Core mapped to ESE
BJC (Romania)	ESE, Dublin Core
RCE (Netherlands)	native schema exports to CARARE, Dublin Core, SKOS
NPU (Czech Republic)	native schema exports to CARARE
AVINET (Norway)	EAD
VUKF (Lithuania)	native schema (ARUODAI) exports to CARARE
UoY ADS (United Kingdom)	native schema exports to CARARE, extended Dublin Core
IPCHS (Slovenia)	native schema exports to CARARE, Dublin Core
Provincie Limburg (Belgium)	SPECTRUM, LIDO can be implemented
Zavad Zara (Slovenia)	extended Dublin Core
CG33 (France)	EAD
FMNF (Portugal)	LIDO, EAD, CARARE
AIT (Austria)	Dublin Core, EDM, ESE
ABMR (Sweden)	SPECTRUM
PSRL (Bulgaria)	Dublin Core, ESE
BGB (Serbia)	Dublin Core, native XML formats
HU (Turkey)	extended Dublin Core
CUT (Cyprus)	EDM
AHAI (Iceland)	no metadata
DP (Ireland)	Dublin Core

Table 6. Metadata schemas

Institutions that use a metadata schema have in several occasions extended a standard schema with customized elements in order to accommodate for specific information and to better describe items in their collections. The following table summarizes the elements that were customized by content providers in their schemas:

- status of object
- COBISS search

- group
- postal number
- type of display
- rights statement
- additional title
- subject
- additional description
- physical description (maps and photos)
- keywords
- receiving date
- thumbnail URL
- signature
- place of publication
- extended coverage (temporal – spatial)
- extended subject (archaeological subjects e.g interventions)
- C14 dates, Aerial photos
- artist/author/publisher
- artist group
- building phases of monuments
- monument id from the national monuments registry
- geographic place names
- vocabularies

Table 7 (below), on the other hand, summarizes schema extensions introduced by each content provider.

Content Provider	Metadata schema
PSNC (Poland)	Signature and Place of publication were added
BJC (Romania)	Greenstone software allows the change of the schemas through Metadata Set Editor
UoY ADS (United Kingdom)	Coverage was extended to support more precise definitions of spatial and temporal elements - Subject was extended to support archaeological subjects such as interventions, C14 dates, Aerial Photos, etc
Zavad Jara (Slovenia)	Status of the object (the purpose is to support editing) COBISS search (it is possible to transfer the metadata from the bibliographical system COBISS) Group (internal element) Postal number (the smallest controlled geographical unit) Type of display (photo gallery or/and as an object with metadata) Rights statement (CC BY-NC , free access, unknown) Additional title Subject Description
HU (Turkey)	keywords, description, physical description (for maps and photos) and receiving date, thumbnail link
DP (Ireland)	Extended to match CARARE v 1.1: Metadata for Heritage Assets and Digital Assets completed with dublin core terms e.g. Heritage asset type dc.subject. No metadata

	created for Heritage Asset Identification/Designation fields within CARARE. In process of extending to new requirements of CARARE V2.0 once full documentation and tutorials are provided as part of 3D-ICONS project so schema can be applied to documentary files and images.

Table 7. Metadata schema extensions

Most providers use Dublin Core as a base schema and build their native schema on top. They find really important to be able to store spatial and temporal information. Many don't have this information but think it is really important to be able to enrich their content with such information. Many providers use gazetteers, vocabularies and thesauri.

Only 3 out of 24 partners have a XSD describing the schema for the metadata which will be submitted to LoCloud. 9 partners validate their XML metadata schemas while 7 do not. 15 partners use UTF-8 encoding for their metadata while 2 do not. Most providers said that there is no character encoding issue as their systems export content in UTF-8.

Two significant elements in terms of metadata quality are the Title (appellation) element and the description element. In the following table (Table 8), the number of records (percentage) that provide this information is depicted.

Percentage of records	Title (appellation) element	Description element
None	4	6
1%-25%		1
25%-50%		1
50%-75%	2	2
75%-99%	2	6
ALL	**16**	**8**

Table 8. Records with Title and Description elements

A conclusion that came out of the workshops is that there are third party collections with minimum to no description. On the other hand a few third party collections have a description element that contains rich information about the object (e.g. photograph collections presented by PSNC) and metadata could probably be extracted from that description.

4.2 Controlled vocabularies and thesauri

Almost half the partners use controlled vocabularies, in the form of term lists or thesauri. There is a variety in the elements a partner uses a controlled vocabulary or thesaurus with (Table 11). Most providers identify the importance of vocabularies and vocabulary services; they however believe

that most contributed collections won't include vocabulary elements and the extra effort involved in enriching content with vocabularies will most probably discourage small providers from using them.

Content Provider	Elements populated with controlled vocabularies
PSNC (Poland)	For Language, Resource Type, Format
KUAS (Denmark)	For works of art: Type Material For museum objects: Period
RCE (Netherlands)	System Item types Predicates Overig AAT Facet Stijlen en perioden ABR ABR Complextypen Archeologische verwervingswijzen Archeologische verzamelwijzen ABRN ABRN Vondst Eigenschappen artefact baksels artefact categorieën artefact functies artefact materialen artefact onderdelen artefact technieken artefact types artefact versieringen artefact vormen Deventercodes baksels Deventercode vormen ABRN Artefact Concepten ABRN Artefacten ABRN Complextypen ABRN Culturen ABRN Perioden ABRN Sporen ABRN Structuren
NPU (Czech Republic)	Locality
VUKF (Lithuania)	IS "Aruodai" - YES AKMENS AMŽIAUS STOVYKLAVIETĖ
UoY ADS (United Kingdom)	Subject Period Location
Provincie Limburg (Belgium)	All relevant elements (compatible with Spectrum). Object types, materials, places (location, collection, creation, represented), time periods, creators, represented objects/persons/places/..., etc.
Zavad Zara (Slovenia)	Partly. Beside controlled vocabulary contributors can add their own tags.
AIT (Austria)	DISMARC vocabularies: ERAs, Languages, Geography, GeoHistorical Other Vocabularies for collection description: AccuralMethod, AccrualPeriodicity, AccrualPolicy
PSRL (Bulgaria)	yes dcterms:spatial - geografical names dc:subjects
AHAI (Iceland)	Site type Type of research Type of method
DP (Ireland)	dcterms:spatial, Getty Thesaurus of Geographic Names, Geonames, Heritage Asset/Actors metadata Art & Architecture Thesaurus, Agent Facet Digital Resource, dc:format - MIME Media Types list, dc:type - DCMI Type Vocabulary
FRS (Italy)	At the moment we are starting to use the software "Samira" which provides controlled vocabularies.

Table 11. Elements populated with controlled vocabularies – thesauri terms

The following table presents standard controlled vocabularies or thesauri used by content partners (Table 12). These are:

i) ABR (Archeologisch Basis Register)
ii) CZ_RETRO
iii) MEDIN
iv) NMR
v) MDA
vi) RCAHMS
vii) Library of Congress Subject Headings
viii) Getty Thesaurus
ix) MIDAS
x) AAT-Ned (Getty AAT, with Dutch translation)
xi) ÖFOS
xii) BIC Standard Subject Categories
xiii) local and native vocabularies

Content Provider	Standard controlled vocabularies - thesauri
PSNC (Poland)	Library of Congress Subject Headings. Some libraries are using central catalogue KABA as a source of subject headings. It is maintained by http://centrum.nukat.edu.pl/
RCE (Netherlands)	Archeologisch Basis Register (ABR) ABR Molens (Mills)
NPU (Czech Republic)	CZ_RETRO
UoY ADS (United Kingdom)	MEDIN subject NMR Monument Type MDA Archaeological Objects NMR Building Materials NMR Defence of Britain NMR Components NMR Maritime Craft Types NMR Maritime Cargo NMR Evidence NMR Archaeological Science NMR Event Types NMR Historic Aircraft Types RCAHMS Monument RCAHMS Object RCAHMS Maritime Library of Congress Subject Headings Getty Thesaurus of Geographic Names MIDAS Monuments MIDAS Periods
Provincie Limburg (Belgium)	AAT-Ned (Getty AAT, with Dutch translation) for the general concepts. Own lists with links to externally available lists for persons, places.
CG33 (France)	Thesaurus W (archival thesaurus), maintained by Archives de France : http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/thesaurus
Zavad Zara (Slovenia)	We use vocabulary, made by National Library
AIT (Austria)	external collections: ÖFOS, BIC Standard Subject Categories
FRS (Italy)	Some fields can be filled with a free text, while other ones use vocabularies. They are of two types: Open (implementable with new terms from authorized operator); Closed (related to norms of ICCD, Central Intitute for the Catalogue and the Documentation) e.g. luog_at_atto Tipologia luog_co_cott Tipo luog_do_dofa Autore

	Fotografia luog_do_dofp Formato di compressione luog_do_dofs Formato di memorizzazione luog_ev_evot Tipologia luog_in_infa Motivo Chiusura Temporanea luog_in_infq Motivo Chiusura Temporanea (EN) luog_og_ogtm Tipo Materiale luog_pb_pbcp Altri strumenti luog_pv_pvcl Località luog_se_seai Attività luog_se_sers Servizi luog_sp_spci Proprietà luog_sp_spvt Stile

Table 12. Standard controlled vocabularies – thesauri

Only 3 partners use SKOSified vocabularies and all three of them support that they include links to the IDs of each SKOS concept. Moreover, 4 partners out of 24 use a web service for vocabularies/thesauri in their system while 14 partners state that there exists a relevant authority for vocabularies in their country.

4.3 Geographical Information

Geographical information is an important part of the LoCloud content. About half content partners support a standard geographic reference system for coordinates. Reference systems used by partners are depicted in the following table (Table 13).

Content Provider	Geographic coordinate system
RCE (Netherlands)	WGS84, RD
NPU (Czech Republic)	SJTSK now, but we can transform it in WGS 84
VUKF (Lithuania)	LKS94 (national) and WGS84
UoY ADS (United Kingdom)	WGS84, OSGB, OSI
Provincie Limburg (Belgium)	Latitude - longitude
CUT (Cyprus)	WGS 84, Ordnance Survey
AHAI (Iceland)	EPSG : 3057
PrifUK KAEG (Slovakia)	WGS 84
DP (Ireland)	WGS 84, Ordnance Survey Ireland ING & ITM

Table 13. Geographic coordinate system

According to partners, only in 4 countries historical place names are accumulated in information systems (gazeteers etc.) of digitized heritage and/or digital humanities, while in 12 countries there are no such information systems. Only 5 partners maintain or use themselves an existing list of historical or local place names.

5. Implications for intermediary schemas

One of LoCloud's main objectives is to ensure interoperability between native content providers' metadata, the metadata stored in the aggregator repository and the metadata that will be delivered to Europeana. It became apparent in the previous chapter that there is a variety of metadata schemas and specifics among the content providers. These schemas need to be identified and then mapped to a set of intermediary schemas as suggested in D1.2: Definition of Metadata Schemas. This section reviews the options specified in deliverable D1.2 in the light of the findings of the current report on providers' content and metadata.

From the content providers workshops we noticed that most providers have or can more easily deliver their content in CARARE, LIDO, EAD or a form of extended Dublin Core. From the content providers' workshops feedback we received, the most appropriate intermediary schemas for delivery are CARARE for immovable objects, LIDO for movable (museum material). In the last content providers workshop EAD was introduced as a strong candidate intermediary schema as it was noticed that content providers have much archival material and would prefer to use EAD. EAD may pose a challenge regarding the MINT mapping tool as, although in theory it can be integrated with the tool, it is a complex hierarchical schema and has not previously been implemented in MINT. MARCXML was considered as a candidate intermediary schema, mainly for library material. However, although several libraries store their bibliographic content in MARC21, none are planning to deliver this content to LoCloud - they plan to submit special collections containing digital resources. These collections are described with extended Dublin Core and will be mapped to one of the three intermediate schemas according to their type. A suggestion was made that the possibility of a schema like EDM or ESE should be considered as some content providers can export content in ESE and EDM from their past involvement in projects that planned to submit content to Europeana. The following table shows the suggested intermediary schemas based on the metadata schemas content partners use to describe their collections and the object types contained therein (Table 14).

Content Provider	CARARE	LIDO	EAD	ESE/EDM
Norsk Kulturrad (Norway)			✓	
PSNC (Poland)				✓
MECD (Spain)				✓
KUAS (Denmark)		✓	✓	
BJC (Romania)				✓
RCE (Netherlands)	✓	✓	✓	✓
NPU (Czech Republic)	✓			
VUKF (Lithuania)	✓	✓		✓
UoY ADS (United Kingdom)	✓	✓	✓	
IPCHS (Slovenia)		✓		
Provincie Limburg	✓	✓		

(Belgium)				
CG33 (France)			✓	
Zavad Zara (Slovenia)				✓
Future Library (Greece)				
FMNF (Portugal)	✓	✓	✓	✓
AIT (Austria)				✓
ABMR (Sweden)		✓	✓	
PSRL (Bulgaria)	✓	✓	✓	
BGB (Serbia)				✓
HU (Turkey)			✓	
CUT (Cyprus)	✓	✓	✓	✓
AHAI (Iceland)	✓			
PrifUK KAEG (Slovakia)	✓			
DP (Ireland)	✓	✓		
FRS (Italy)	✓			✓

Table 14. Recommended intermediary schemas

6. Conclusions

The online questionnaire survey and the content providers' workshops produced numerous valuable conclusions.

Collections among providers differ greatly in terms of size, number of items, level of metadata description object types etc. More than half of the content providers will submit items that belong to more than one collection held by their institution.

Some content partners will provide both native and third party collections, but there are cases of partners that will only provide third party collections (e.g. Cyprus University of Technology) or native collections are still under development (e.g. Future Library). Most third party content will belong to more than one collection. At this point there is still a lot of ambiquity regarding third party collections and their content; many content partners are waiting for the development of microservices in LoCloud before contacting small institutions. This means that although many content partners have already established their network of smaller providers, the information about the content contained in third party collections is in many cases not clear. Content partners stated that it is likely that there will be third party collections that contain minimum to no metadata description e.g. photographic collections.

There is a balance in the object categories content partners have and will provide to LoCloud. The initial data analysis indicates that 14 providers have movable objects, 13 immovable, 13 library materials and 11 archival sources. Most objects will be images, both thumbnails and full images and texts. Some audio and video material will also be submitted and a few 3D representations. Most objects will be complex and they will consist of more than one datastreams, including a metadata description in XML, thumbnail and full image, text etc.

Most content partners have the metadata they plan to submit in LoCloud openly accessible. Some providers impose restrictions to the full sized high resolution images of their content e.g. HU and provide them under additional licence, but they still provide an unrestricted access to a thumbnail image.

Several metadata schemas have been identified among content partners. These metadata schemas refer both to the metadata schemas the collections are described with as well as the metadata schemas collections can be exported to. Among these schemas the most common were CARARE, LIDO, EAD and several different extensions of Dublin Core. The elements most providers introduced in their native schemas aim to store information about the status of an object, rights information, spatial and temporal information and controlled vocabulary related information. There were content partners that pointed out that there is no established practice on a national level for describing collections and their national institutions don't share a common metadata schema for describing common types of objects. These partners view their participation in LoCloud as an opportunity to address this issue. Most providers don't have an XSD describing their schema and only some of them check their XML metadata for validity. About half content partners store their data directly in UTF-8 and all partners can export into unicode formats. Most providers include a Title and a Description element in the majority of their content.

Almost half the partners use controlled vocabularies and thesauri in order to complete information in various different elements in their collections. Most providers identify the importance of vocabularies and are really interested in the vocabulary services that will be developed in LoCloud; they however believe that most contributed collections won't include vocabulary elements and the

extra effort involved in enriching content with vocabularies will most probably discourage small providers from using them. Only two content partners use vocabularies available in SKOS.

Regarding geographical information approximately half content partners support a standard geographic system, with WGS84 being the most commonly used. During the content providers workshops content partners showed great interest towards geographic enrichment services.

The aim of this content survey and metadata analysis has been to guide and inform planning of the aggregation strategy, to provide feedback for the selection of appropriate intermediary schemas to be used in metadata mapping in LoCloud, and provide input for the technical partners to the design and development of appropriate micro-services for LoCloud.

References

Coburn, Erin, et al. "LIDO–Lightweight Information Describing Objects, Version 1.0." *ICOM International Committee of Museums* (2010). (http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf). Date accessed 30-09-2013

Europeana Semantic Elements Specification and Guidelines (http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5). Date accessed 30-09-2013

Fernie, Kate, Dimitris Gavrilis, and Stavros Angelis. "The CARARE metadata schema, v. 2.0." (http://carare.eu/cze/content/download/11454/98739/file/The%20CARARE%20metadata%20schema2.pdf). Date accessed 30-09-2013

Isaac, Antoine. "Europeana data model primer." (2011). (http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5). Date accessed 30-09-2013

Pitti, Daniel V. "Encoded archival description: An introduction and overview." (1999): 61-69. (http://www.dlib.org/dlib/november99/11pitti.html). Date accessed 30-09-2013