DELIVERABLE

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

D1.5: Requirement analysis

Revision: final

Authors:

Stavros Angelis (DCU)
Dimitra Nefeli Makri (DCU)
Dimitris Gavrilis (DCU)
Eleni Afiontzi (DCU)
Costis Dallas (DCU)
Adam Dudczak(PSNC)

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
O.1	01/11/2013	S Angelis	DCU	First draft
0.2	25/11/2013	S Angelis	DCU	Integration of comments from A Dudczak and C Dallas
0.3	29/11/2013	D Gavrilis	DCU	Final edit
1.1	01/06/2014	S.Angelis	DCU	Annex II added

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 deliverable is part of LoCloud WP1: Planning, preparation and requirements and aims to build on the previous work already done in WP1. D1.5: Requirement analysis presents the technical aspects of the user requirements that have been collected through a series of workshops and surveys and aim at facilitating the design of the technical infrastructure of the LoCloud project.

The structure of this document is as follows:

- Section 2 describes the methodology followed to identify and analyse user requirements. This deliverable builds on, and further analyses, information gathered during Task 1.4.
- In section 3, a classification of the content providers is made, by way of breaking down and presenting their individual characteristics; these characteristics include the type of collections these institutions hold, their IT expertise, and other important factors that enable the profiling of content providers and are significant for the development of tools and services in LoCloud.
- In section 4, user needs and requirements are presented; throughout the first stage of the LoCloud project content partners participated in various surveys and workshops providing important information about their needs; these needs are depicted here organized according to the user profiles defined in section 3, in the form of user stories.
- Section 5 assesses the relevance of the intermediate schemas identified in D1.2: Definition of Metadata Schemas on the basis of the high level characteristics defined in that deliverable (focusing on schema complexity and richness), and here further expanded by the results of the content provided workshops presented in D1.3: Content and metadata analysis
- In section 6 the types of possible incoming binary datastreams is presented.
- Finally, in section 7 and section 8 the various workflows and ingest points that have been identified are presented from the users' point of view.

Annex 1 presents the LoCloud content provider's profiles as identified from the results of the questionnaire survey

The deliverable was updated and Annex 2 added to include a short report and analysis of the state-of-the-art based on the results of the series of surveys carried out during the first year of LoCloud.

2. Methodology

In this section we describe the methodology followed in order to identify and analyse user needs and requirements as part of the requirement analysis work being done in LoCloud. This requirement analysis will constitute a basis for the development of the technical infrastructure and provide important information to assist technical partners of LoCloud in their significant effort in developing tools and services appropriate for the specific needs of this project.

The first step towards this direction was to identify metadata schemas appropriate to be used as intermediary schemas between the content ingested from content partners of LoCloud, and EDM, the target schema for content delivery to Europeana. To this end, a state-of-the-art study and a targeted survey regarding content partners' collections was performed, as presented in D1.2: Definition of Metadata Schemas.

Parallel to this task, a content and metadata analysis task was performed. This task was divided in two complimentary tasks: an online questionnaire survey, inquiring about content partner collections and contributed collections, and three content providers workshops. The aim of both was to gather information about content providers content and metadata, discuss and identify possible needs and requirements regarding components developed in LoCloud. Results where summarized and presented in D1.3: Content and metadata analysis.

By analyzing information gathered during the above tasks, it was possible to elicit and define specific content partners' profiles, as well as specific needs and requirements about the tools and services developed within LoCloud and the technical infrastructure in general.

3. Content provider profiles

In this section we identify possible content provider profiles. These profiles were based on the different nature of the entity holding the collections under consideration (e.g. small museums, libraries, personal collections, etc.), as well as their responses to the survey conducted as part of the work reported in this deliverable. From the online questionnaire survey performed as part of the content and metadata analysis task, some useful findings emerged, leading to the definition of content provider profiles. An important first finding was the identification of distinct types of collections in LoCloud, on the basis of the primary material they contain, as seen below (Table 1).

Types of collections
geophysical images of buried archaeological structures (buildings, chapels), archaeological reports
paintings and drawings, pieces of porcelain, engravings of various subjects, embroidered textiles
historic photos and pictures, local photos and documents
“an old family library”
movies, oral history and multimedia content
artefacts from museums
maps and historical cultural landscapes
textual digitalized documents and photographs about the beginning of railway, photographs of artifacts, trains, locomotives, buildings etc.
graphic content including surveys, pans and illustrations
images and photographs from Irish Monasteries

Table 1: Type of collections

As depicted in Table 1 collections present diversity both in content and in content types. The content varies from archaeological structures and archaeological reports to paintings, drawings, family libraries, maps etc. The content types also differentiate much as there are images, movies, audio and multimedia content, textual digitized content etc, and all these different types of materials should be acommodated by the technical infrastructure.

An additional characteristic concerned the information skills and information specialists in the staff of content providers, as summarized in the charts below.

The first chart shows how many of the content providers (23 in total) have in-house expertise in information management or IT systems, while the second chart shows how many of the content providers have in-house expertise in librarianship and information science, archival science or cultural heritage documentation.

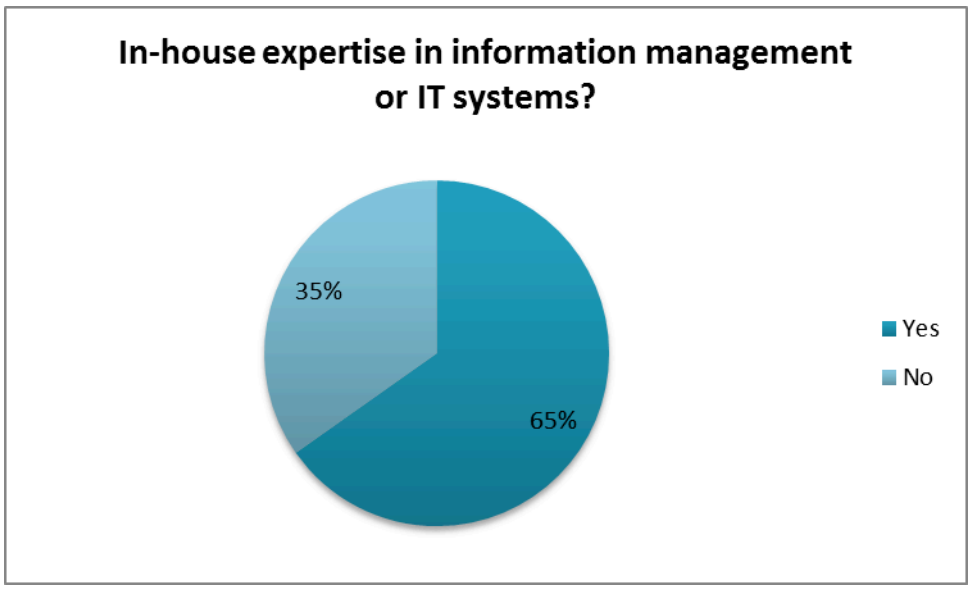

Figure 1: Expertise in information management or IT systems

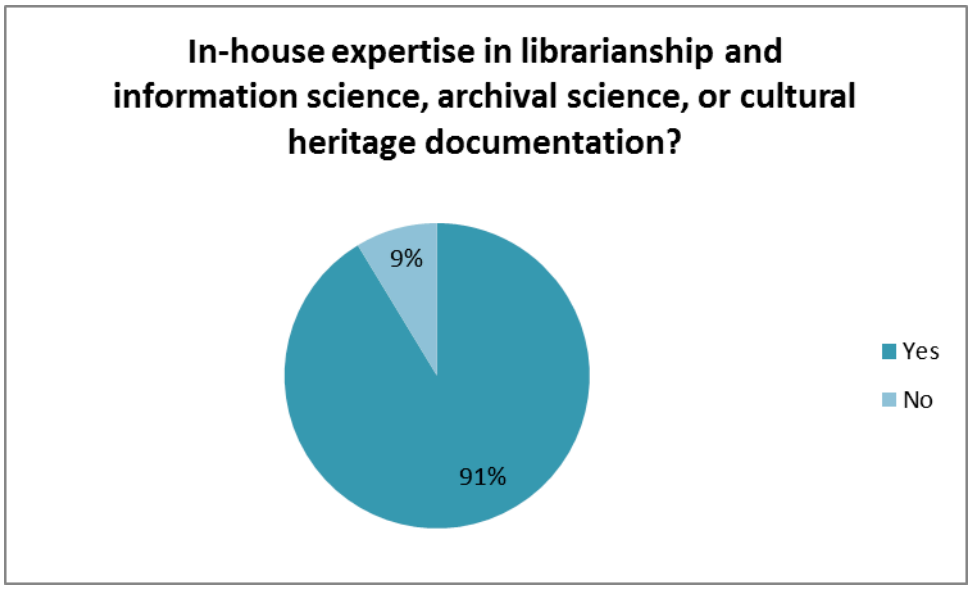

Figure 2: Expertise in librarianship and information science, archival science or cultural heritage documentation

These two figures show that more than half providers have in-house expertise in IT and the majority in librarianship, information science, archival science or cultural heritage documentation.

Another important characteristic concerned whether content providers are familiar with metadata and metadata schemas, or not. The results for the 23 given answers of the content providers are shown in the following figure.

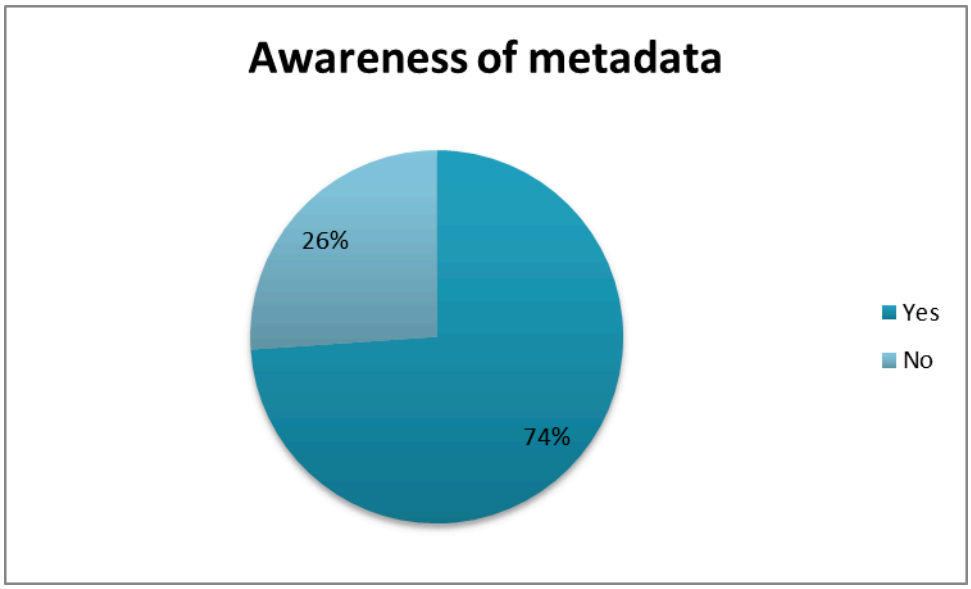

Figure 3: Awareness of content providers for metadata

Finally, an additional dimension concerned if some part of a content partners' digital collection or digital library is available through Europeana. About half providers already make their content available through Europeana as seen on the following chart.

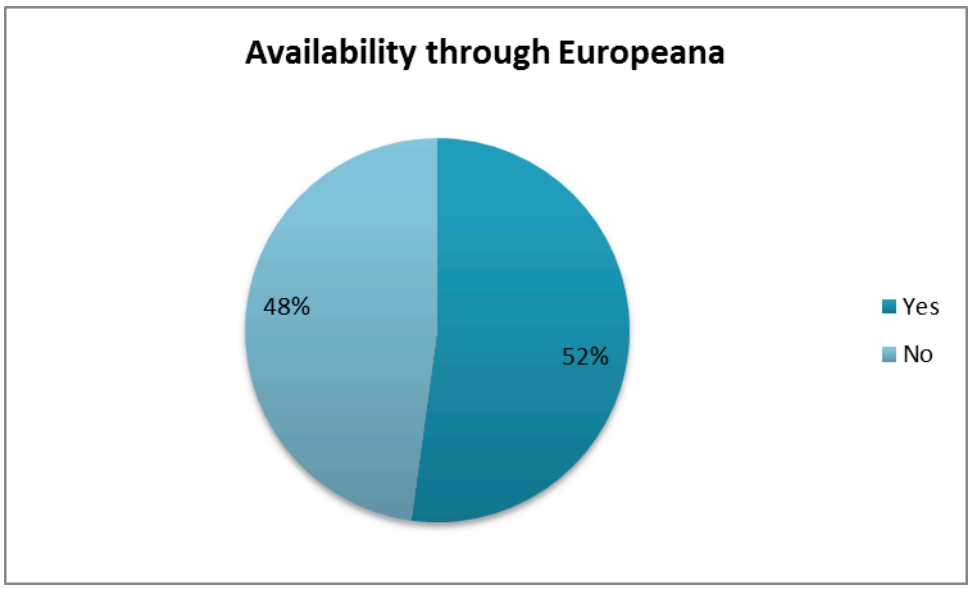

Figure 4: Availability of provider's collections or digital library through Europeana

The profiles of different types of holding institutions on the basis of these characteristics are summarized as follows.

3.1. Small memory institutions (including museums, libraries and local archives)

- **Type of collection**: photographs, newspapers (serials), books, museum objects and audiovisual content
- **IT skills:** Usually no IT staff or expertise at all. They can use the most widely known Web 2.0 services like YouTube or Flickr.
- **Knowledge about metadata**: Professionals working in these institutions are able to document properly their collections. The concept of metadata can be easily explained to

these users, even if they lack concrete knowledge about state-of-the-art metadata standards like EAD, CARARE or LIDO. With proper guidance this group should be able to provide high quality descriptions for their submitted objects. Depending on the type of institution (library, museum or archive) users in this group may more prepared to use specific metadata conventions, like hierarchical descriptive strucures (archives) or event-based descriptions (museums).

- **Have a working digital repository:** No
- **Already visible in Europeana**: No
- **Main motivation:** Share materials dedicated to local community history.

3.2. Existing Europeana content providers

- **Type of collection:** photographs, newspapers (serials), books, museum objects, 3D objects, inventories of monuments and historic buildings and audiovisual content
- **IT skills:** They have a limited IT expertise and hire part time IT staff. They have already participated in projects that helped them to create repository and deliver metadata to Europeana.
- **Knowledge about metadata:** They have knowledge about metadata standards like Dublin Core, Europeana Semantic Elements (ESE), CARARE and LIDO. Like small memory institutions (see 3.1), they are able to create high quality metadata.
- **Have a working digital repository:** Yes
- **Already visible in Europeana:** Most likely yes.
- **Main motivation:** On one hand, looking for a way to support the growing user base of their repository, and, on the other hand, looking for a way to cut costs of maintaining their own repository.

3.3. Large-scale national agencies or support organizations (e.g. NRENs, national aggregators)

- **Type of collection:** none
- **IT skills:** They have wide IT expertise, as they typically have to support other organizations in digital information work. They may also operate as aggregators directly cooperating with Europeana. Also, they may already have all the necessary hardware in order to create cloud-based solutions.
- **Knowledge about metadata:** They have significant expertise in metadata transformation and mapping to a range of schemas.
- **Have a working digital repository:** Several do.
- **Already visible in Europeana:** Yes in many cases.
- **Main motivation:** To offer new services to local institutions who wish to create their own digital archives.

3.4. Non-professional collections (municipalities, associations and hobbyists)

- **Type of collection:** mainly photographs and a-v content
- **IT skills:** Usually no IT staff or expertise at all. They can use the most widely known Web 2.0 services like YouTube or Flickr.
- **Knowledge about metadata:** No pre-existing knowledge. In some cases it would be necessary to explain why it is necessary to describe digital objects using structured metadata. This user group want may not be intrinsically interested in learning about

metadata, but they are very engaged with the content of their collections and have a strong motivation to promote it. They would be able to create uniform quality metadata records, assuming that they are assisted by appropriate tools, e.g., with guidance for the disambiguation of geo-spatial information, controlled vocabularies, and predefined mappings to advanced metadata schemas like EDM or LIDO.

- **Have a working digital repository:** No
- **Already visible in Europeana:** No
- **Main motivation:** Share materials related to their activities with Internet users and promote this document to increase visibility of institution.

3.5. Private collections owned by individuals

Similar to non-professional collections (see 3.3).

3.6. Ordinary Internet users

Ordinary Internet users are interested in finding interesting content, which would be easily accessible and discoverable through well-known services like Europeana. They do not possess technical knowledge but are well-versed in using services like YouTube, Google Maps etc.

4. User stories

User stories (scenarios) present user needs extracted from the analysis of the online questionnarie survey and the content provider workshops, organized according to the content provider profiles defined in the previous section. Requirements represented in these user stories will constitute the basis for the definition of tools and services for the content providers.

4.1. Small memory institutions (including museums, libraries and local archives)

#	I want to….	so that…
SM1	have access to a simple and friendly system,	I can manage my collection easily, export my data, add new information or update existing data.
SM2	share my digitized collections online,	I can promote local history and current affairs which are taking place in my region.
SM3	enrich and improve the data quality of my metadata,	the description of my collections to be more precise and it would be easier to find items from the collection.
SM4	make the metadata of my collections available under the terms of a Creative Commons Zero License,	it can be easily reused by Europeana and other services.
SM5	have access to geo-location enrichment services,	my collections have precise description of its spatial coverage.
SM6	implement historic place names micro services to my collections,	to be harmonized of names and enriched of content with gazeteer information.
SM7	make use of vocabularies services,	I can query and integrate a vocabulary in a local application.
SM8	have the opportunity to translate the content of a part of or all my collections,	my collections to be understandable in people of different nationalities.
SM9	use a crowdsourcing tool,	metadata of items from my collection can be enriched by volunteers.
SM10	keep statistics from the content of my collections,	I can check the completeness of description of my collections.
SM11	have a tool for sending content directly to Europeana,	my collections to be searchable from Europeana portal

4.2. Existing Europeana content provider

#	I want to….	so that…
CP1	share my digitized collections online,	I can promote local history and current affairs which are taking place in my region.
CP2	enrich and improve the data quality of my metadata,	my collections to be precise and completely updated.
CP3	have access to geo-location enrichment services,	my collections are completely located worldwide.
CP4	make the metadata of my collections available in accordance to Europeana rights policy,	no conflict in rights exists.
CP5	make use of vocabularies services,	I can query and integrate a vocabulary in a local application.

4.3. Non-professional memory archive

#	I want to….	so that…
NP1	describe my digital objects including only the necessary information,	a simple and easy to use metadata schema would be preferable.
NP2	use geo-location and vocabulary tools provided by a clear guidance,	quality metadata records could be created.
NP3	share materials related to my activities with internet users,	the visibility of my institution to be increased.
NP4	make use of a mapping tool,	my collections to be adopted to the used schemas like EDM or CARARE.

4.4. Large supporting organizations

#	I want to….	so that…
LO1	communicate with small partners in a more attractive way, explaining them the benefits of participating in the project,	more small partners can be motivated to share their materials.
LO2	cooperate directly with Europeana working as a national aggregator,	to be avoided to have duplications of records.
LO3	use services as for vocabularies, thesauri and geospatial information,	the metadata of digital objects of local institutions to be enriched.

4.5. Private collection owners

#	I want to….	so that…
CO1	have access to a simple to use and friendly system,	I can manage easily my collection, export my data, add new information or update it.
CO2	use a simple standard metadata schema,	my metadata to be described in a proper way.
CO3	use simple tools for enrichment of my metadata,	my collection to be completely updated.

4.6. Ordinary Internet users

#	I want to….	so that…
US1	find interesting content easily	I could use cultural heritage content to develop my work/business.
US2	make use of services like Europeana	I can access trustworthy information.
US3	help to enrich metadata of existing objects through crowdsourcing projects,	It would be easier to find them in the future.

5. Intermediate Metadata Schemas

A variety of metadata schemas are used by LoCloud content providers to describe their native collections, as was identified in D1.3: Content and metadata analysis. Metadata records should be delivered to Europeana in a uniform way and the interoperability between native metadata held by organizations and metadata used in Europeana has to be ensured. To this end, a set of suitable intermediate schemas was identified and suggested in D1.2: Definition of Metadata Schemas, taking into account existing metadata schemas used by the content providers and the Europeana Data Model schema.

When asked, content providers indicated that the most suitable metadata schemas that they could use to provide their content into are, primarily, CARARE, LIDO or EAD, and, possibly also, ESE (or EDM); just one provider indicated that they could also deliver content in UNIMARC XML. According to the survey of content providers, twelve (12) may deliver content in the CARARE schema, eleven (11) in LIDO, ten (10) in EAD, and eleven (11) in ESE/EDM. These four schemas were selected because they are appropriate respectively for handling information in the fields of archaeology and architecture, museum collections, historical archives, and other, more general information from the cultural heritage domain. CARARE, LIDO and EAD are schemas capable of expressing rich, detailed information. ESE, a qualified Dublin core schema, is relatively less expressive in terms of information structure, but many content providers decided to include it as they are are familiar with it, and established mappings their existing primary schema may already exist; also, there are already reliable mappings between ESE and EDM.

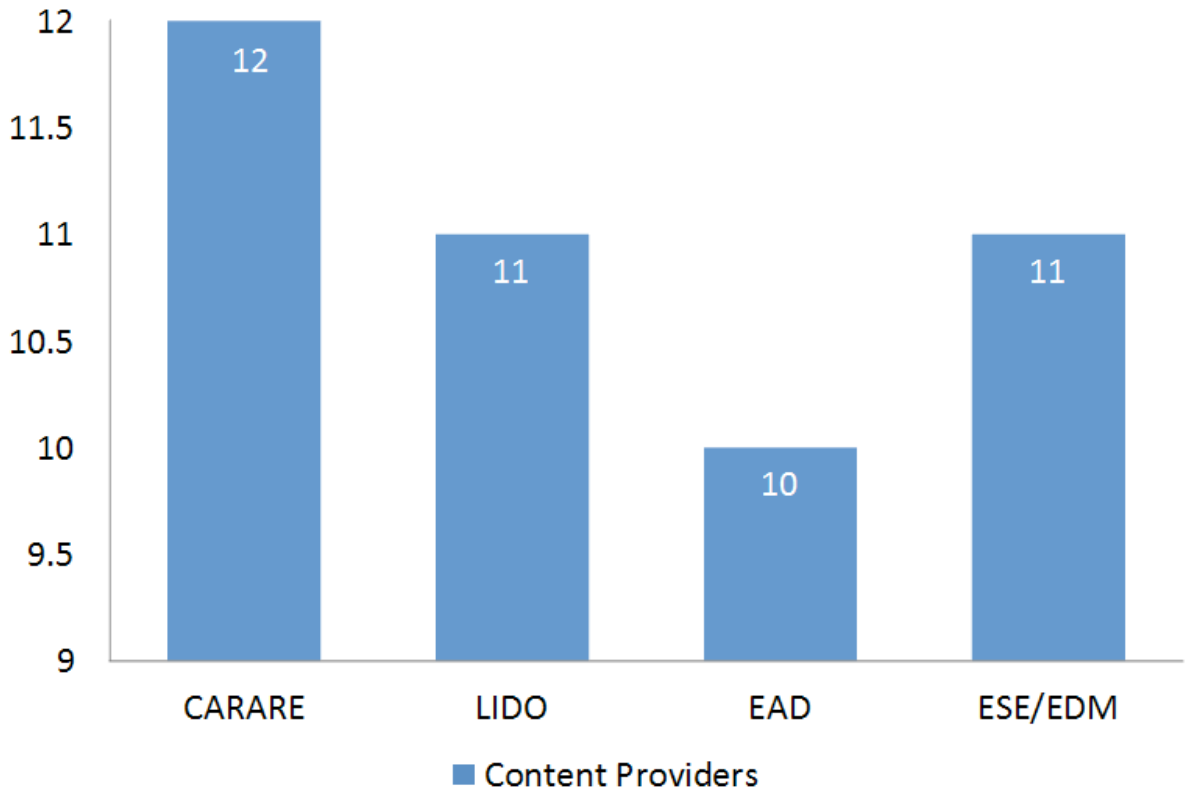

Figure 5: Metadata schemas reported as suitable for content delivery

The schemas mentioned above are mostly appropriate for the following types of material:

1. Immovable objects (e.g. monuments, archaeological sites, features and finds, shipwrecks, etc.) – CARARE Schema
2. Movable objects (e.g. museum objects, movable artefacts, etc.) – LIDO
3. Archival material (e.g. information contained in finding aids, corporate records, personal papers, etc.) EAD
4. General cultural items - ESE

CARARE Schema is a rich and extensive schema, capable for rich description of multilingual objects. It is appropriate for immovable heritage assets as mentioned above, with over 230 elements and 5 different element-related attributes. CARARE records consist of a top level element that wraps 4 main kinds of entities: i) Heritage Asset, ii) Digital Resource, iii) Collection Information and iv) Activity. These entities may extend to multiple levels of depth, thus allowing for the capture of rich information. CARARE Schema is expressed in XML. It has already been mapped to EDM and has been used extensively for data delivery to Europeana.

LIDO schema is also a rich and extensive schema intended to describe museum objects like artworks, artefacts, technology objects, etc. LIDO is an application of the CIDOC CRM, in the sense that it provides an explicit format to represent content in a standardized way. It is a schema intended to support the full range of descriptive information about museum objects, and supports multilingual information with approximately 200 elements and several different attributes. LIDO consists of 4 main elements: i) Object WorkType, ii) RecordID, iii) RecordSource and iv) Title. A LIDO record is divided into two groups of information: i) descriptive information and ii) administrative information. LIDO is expressed in XML, and has been extensively used for delivery of content to Europeana. Also LIDO has been mapped to both ESE and EDM.

EAD is a hierarchical schema appropriate for archival material, expressed in XML. It is capable of expressing deep nested hierarchies, and providing context and provenance information. The EAD DTD specifying the elements of the EAD schema contains 146 elements. A mapping between EAD and EDM exists, although there is no clear evidence if EAD content has already been delivered to Europeana. Also there is a possible complication, as, though this seems feasible, EAD has not been used with the MINT tool in the past.

ESE is a simple schema based on Dublin Core (ESE is an application profile of DC). It was developed as the main operative scheme by Europeana, but is now replaced by EDM. All ESE records which were in Europeana have now been transformed into EDM. ESE consists of 28 elements and is appropriate for diverse kinds of cultural objects and many providers express a preference towards it on the grounds that they have used it in the past, and they already have ready mappings and content already expressed in ESE.

6. Types of binary datastreams

The types of the binary data streams identified fall into the following categories:

- Documents (MS Word, OpenOffice, etc.)
- Spreadsheets (MS Excel, CSV, etc.)
- Images
- Movies
- Audio
- Zip archives
- Spatial files (e.g. KML files containing places, coordinates)
- Other data formats (e.x. XML, RDF, other binary data)

Regarding average file sizes, no accurate data exist, as there is a very diverse set of organization types and sizes. Furthermore, no accurate assumptions can be made regarding use of provider repositories. An attempt has been made though to produce a way of estimating an average file size per page for each of the identified document types. This can be seen on the table below:

	File Type	**AVG Size per page**
Document	MS Word	15 KBytes for MS word per page
	PDF	100 KBytes for PDF containing text + images (per page)
Spreadsheet	MS Excel	6 KBytes per page
Images	TIFF	65 KBytes per page
Movies	Raw uncompressed 1080p	7 GBytes per minute
Audio	MP3	3.5 Mbytes per song
Zip	-	-
Spatial	KML	<1 KByte
Other	-	-

Table 2: Estimation of average file size per item/page

[sources from: 1, 2]

7. Ingestion paths

One of the most critical issues impacting technical requirements concerns the method of ingestion of metadata from primary collections into Europeana that will be supported. Decision about this issue will determine a substantial part of the technical architecture and its complexity. Ingestion methods were one of the discussion topics in the three workshops. Although it might seem straightforward, this consideration is quite complex, as many providers have their own views and established practices on how to deliver content for ingestion. This view is mostly shaped by their experience, i.e.:

- Larger content providers have an already established digital collection or database, conforming to a domain-specific or institution-specific schema with rich information about collection items. They would typically need to export and map their metadata using a tool like MINT, so that it is then ingested into the LoCloud aggregator, enriched using LoCloud microservices, and delivered to Europeana. Large content providers may require a certain level of control on the way their content is enriched and aggregated for delivery to Europeana.
- Some small and medium content providers may not have an already established database; they may have machine-readable metadata for their objects, but effectively need an application that would allow them to prepare metadata in a form that does not require them to worry about schemas or mapping. This could be either in one of the intermediate schemas (i.e., CARARE schema if their collection concerns archaeological and location-based heritage assets, or LIDO if it concerns artefacts or artworks), or even directly in EDM. In some cases, content providers will use that application as their primary information system, supporting basic documentation, management, retrieval and display/presentation of items in their collection. Small and medium content providers may expect the mapping and enrichment process to be conducted transparently and automatically.
- Medium sized content providers would either fall into the first category or would expect a plugin that will export their data to an aggregator.

Finally, several content providers have already delivered metadata to Europeana in previous projects (in ESE, CARARE or LIDO). Some have ongoing arrangements with aggregators to provide content to Europeana through them. These providers anticipate that they will continue to use the same method of providing content to Europeana in the LoCloud project.

Regarding the methods of ingestion, the following alternatives that conform to the scope and DoW of the LoCloud project have been identified:

IngestionMethod	**Description**
LoCloud aggregator (UI)	Allows users to provide metadata to Europeana through the LoCloud aggregator, by using the MINT and MORE cloud systems. The user utilizes the web based UI to harvest content or to upload XML files with metadata.
LoCloud aggregator (API)	Allows content upload through the LoCloud aggregator API by directly tying native repositories to the aggregator. An API key must be

	required.
LoCloud aggregator plugins	Allows providers to install plugins for their systems that can directly ingest to LoCloud. A plugin must be provided for each system and well-known repositories such as: DSpace, Wordpress, Omeka should be provided.
Lightweight Digital Library	Allows providers to create and manage metadata of their collections using the functionalities of the Lightweight Digital Library, and directly ingest metadata to the LoCloud aggregation infrastructure and to Europeana.
Existing aggregator	This is a special case for providers who already use an existing aggregator to provide content to Europeana. In this case providers are able to continue sending content through the particular existing aggregator, under the following conditions: 1. The EDM element "edm:provider" should have the value "LoCloud" in order to identify records 2. The content provider should have a mechanism of counting, and reporting in a timely fashion, the number of records thus provided Records provided through this path will not benefit from content enrichment services called by the LoCloud aggregator.

Table 3: Ingestion methods and description

8. Workflows

The workflows that will be supported by the system have to take into account the different methods of ingestion as well as the diverse, and possibly distributed, nature of the LoCloud services. Especially because of the latter, the existence of a workflow engine and an execution environment for these services is imperative in order to ensure the quality of the content to be delivered and the minimization of errors. The various workflows must be able to adapt to the following conditions:

- The type of the content provider (small sized, large institution)
- The services to be utilized (e.g. less and more simple for simple schemas like ESE and more and more sophisticated for complex schemas like CARARE)
- The point of ingest (e.g. no mapping required when ingesting from LDL)

The various workflows are presented and organized in the following sections:

8.1. Ingestion through LDL

When ingesting through the LDL application, content will be automatically mapped to one of the intermediate schemas. The enrichment services will be available through MORE.

8.2. Ingestion through MINT

When ingesting through the MINT application, content will have to be mapped using MINT's mapping functionality into one of the intermediate schemas and then stored in the LoCloud storage infrastructure. The enrichment services will be available through MORE.

8.3. Ingestion through existing aggregators

Content providers will be able to continue to ingest content through existing national aggregators under the conditions and with the limitations summarized in Table 3.

8.4. Ingestion of Wikimedia content

When ingesting Wikimedia content, it will be automatically mapped to one of the intermediate schemas and then stored in the LoCloud storage infrastructure. The enrichment services will be available through MORE.

8.5. Enriching content using the LoCloud aggregator (MORE)

After the ingest process, content will be stored within the LoCloud storage layer and in one of the intermediate schemas. An enrichment phase will be provided by MORE and will be implemented using the available micro-services. Each one of these micro-services will create a new enriched datastream that will replace the existing record (by creating a new version of the existing record). The microservices will support one of many of the intermediate schemas and will affect one or more parts of them (e.g. a micro-service will affect only spatial information). A mechanism of streamlining these microservices into specific workflows will be provided and the user will be able to overview/control the whole process.

8.6. Delivering content

The content delivery process will comprise of an easy way of creating a publication to Europeana with a “click of a button”. The content to be published will be available as a Set on an OAI-PMH server. Users (content providers) shall be able to inspect the content themselves.

9. Conclusions

This deliverable presented the technical aspects of the user requirements that have been collected through a series of workshops and surveys during the planning and preparation stage of LoCloud. The methodology followed to identify and analyse user requirements was briefly described. This requirement analysis constitutes the basis of the technical infrastructure as it provides rich information about users' profiles, their needs and their content. Firstly metadata schemas appropriate to be used as intermediary schemas for content delivery to Europeana were identified. Parallel to this task, existing content and metadata was analyzed.

According to this information user profiles were defined, taking into account important characteristics like i) the type of content providers' collections and content, ii) expertise in information managemend and IT skills, iii) in-house expertise in librarianship, information science, archival science or cultural heritage documentation, iv) awareness of metadata and v) content availability through Europeana.

Based on these characteristics the following profiles of different types of holding institutions were defined: i) Small memory institutions, ii) Existing Europeana content providers, iii) Large-scale national agencies or support organizations, iv) Non-professional collections, v) Private collections owned by individuals and vi) Ordinary Internet users.

User requirements in the form of user stories where extracted from the analysis of the online questionnarie survey and the content provider workshops and were organized according to the user profiles above. These requirements are the basis for the definition of tools and services for the content providers.

In Locloud there is much diversity of content as well as a variety of metadata schemas which are used to describe this content. In this deliverable the relevance of the intermediate schemas identified in D1.2: Definition of Metadata Schemas is assessed, and further expanded by the results of the content provide workshops presented in D1.3: Content and metadata analysis. The most suitable schemas to be used as intermediary, as indicated by the providers, were i) CARARE, ii) LIDO, iii) EAD and iv) ESE/EDM. These schemas are matched with the most relevant types of material in content provider collections. The types of expected binary datastreams were identified and categorized on this basis.

One of the most critical issues impacting the technical requirement is the method of ingestion of metadata from primary collections into Europeana. According to the feedback from the content providers' workshops and in conformance with LoCloud's scope and DOW the following alternatives were proposed: i) LoCloud aggregator (UI), ii) LoCloud aggregator (API), iii) LoCloud aggregator plugins, iv) Lightweight Digital Library and v) existing (national) aggregators.

Finally a set of workflows to be supported by the system was defined. These workflows take into account the different methods of ingestion as well as the diverse nature of the LoCloud services. The workflows will be able to adapt to the following conditions: i) the type of the content provider, ii) the services to be utilized and iii) the point of ingest. According to these conditions the proposed workflows are: i) ingestions through LDL, ii) ingestion through MINT, iii) ingestion of Wikimedia content, iv) enriching content using MORE and v) delivering content.

References

1. http://help.netdocuments.com/file-sizes/
2. http://www.filecatalyst.com/todays-media-file-sizes-whats-average
3. http://www.carare.eu/bul/Resources/CARARE-Documentation/CARARE-metadata-schema
4. http://www.lido-schema.org/schema/v1.0/lido-v1.0-schema-listing.html
5. http://www.europeana.eu/schemas/ese/
6. http://pro.europeana.eu/edm-documentation
7. http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki
8. C. Papatheodorou, C. Dallas, et. al, "A new architecture and approach to asset representation for europeana aggregation: The CARARE way". Metadata and Semantic Research, pp. 412-423, ISBN: 978-3-642-24730-9

Annex I - Results from questionnaire (content provider profiles)

Content provider	Type of collection	IT Skills	Knowledge about metadata
PSNC (Poland)	movies, oral history, pictures and multimedia content, collections from very small institutions with city structures and oral history	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of customized Dublin Core
KUAS (Denmark)	Art museums and local history museums.	In-house expertise in librarianship and information science, no IT expertise	No information
BJC (Romania)	Local photos and documents, newspapers and local history books, Library documents from County Public Libraries, Archive documents from memorial house	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of ESE Schema
RCE (Netherlands)	Historical Cultural landscapes, Archaeological reports, Controlled vocabulary of Dutch archaeology. Shipwrecks collections and landscapes.	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of EDM and CARARE
NPU (Czech Republic)	new collection of archaeological sites, the GIS location and digital resources (photographs atc.)	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of CARARE
VUKF (Lithuania)	Collection of hillforts (texts, geodetic data, digitised and digital photos, aerophotos, etc.), Collection of castles	In-house expertise in information management or IT	Aware of metadata, use of CARARE

	and fortified sites (texts, geodetic data, digitised and digital photos, aerophotos, etc.)	systems, in-house expertise in librarianship and information science	
UoY ADS (United Kingdom)	Unpublished archaeological field reports, resource discovery metadata for all 450+ of its existing collections, metadata for the c. 4000 PSAS reports, dating from 1851 to the present. metadata for around 2,500 artefacts (most with images, but not all) held in museums, metadata for the collection totalling about 300 images from small museums/county, 424 images (and reports in PDF, CAD plans in DXF and a variety of other file types) from the Southampton City Council.	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of Extended Dublin Core and CARARE
IPCHS (Slovenia)	No information	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	No information
Provincie Limburg (Belgium)	Monuments, photographs, not a good system to collect.	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of Spectrum and LIDO
CG33 (France)	Archival and documents, textual, postcards, maps and cards. Local history society archive, museum, local environmental and historical preservation association	In-house expertise in librarianship and information science, no IT expertise	Aware of metadata, use of EAD
Zavad Jara (Slovenia)	Collections, related to local history, contributed by various organisations.	No expertise in librarianship and information science, no IT expertise	Aware of metadata, use of Extended Dublin Core
Future Library (Greece)	A collection of digital material of local content, digital stories (video and	In-house expertise in information	No information

	audio), local pictures of cultural and historical value, local texts of cultural and historical value.	management or IT systems, in-house expertise in librarianship and information science	
FMNF (Portugal)	Textual digitalized documents and photographs about the beginning of railway in Portugal; Museum: Photographs of artifacts, trains, locomotives, buildings etc. about the beginning of railway in Portugal.	In-house expertise in librarianship and information science, no IT expertise	Aware of metadata, use of LIDO for museums, EAD for archival collection, CARARE for immobile heritage collections
AIT (Austria)	Heterogeneous (archives, images, library materials, numismatic, archeological images, theatre texts, performance	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of ESE and EDM
ABMR (Sweden)	Collections from a local museum of medicine history. Parchment and paper collection of letters. Collection of photographs from Ånge municipality. Photo Collection related to the school of Kubikenborg.	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of Spectrum
PSRL (Bulgaria)	No information	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of Dublin Core
BGB (Serbia)	No information	No expertise in librarianship and information science, no IT expertise	No information
HU (Turkey)	library and archival collections, mostly archival	In-house expertise in librarianship and information science, no IT expertise	Aware of metadata, use of Extended Dublin Core

CUT (Cyprus)	Images and Books which belongs to the local archive of the Limassol Municipality - 3D Icons which belongs to the Church of Cyprus - AudioVisual materila which belongs to the CyBC	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of EDM
AHAI (Iceland)	Mainly excavation material	In-house expertise in librarianship and information science, no IT expertise	No information
PrifUK KAEG (Slovakia)	Geophysical images of buried archaeological structures (buildings, chapels)	No expertise in librarianship and information science, no IT expertise	No information
DP (Ireland)	Leo Swan Aerial photo collection with archaeology images, image collection of range of graphic content including surveys, pans, illustrations and photographs generated over the last 21 years. Images and photographs from Irish Monasteries	In-house expertise in information management or IT systems, in-house expertise in librarianship and information science	Aware of metadata, use of Dublin Core
FRS (Italy)	Paintings from 16th to 20th c.; over 600 pieces of porcelain by Italian and European manufacturers; 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. 3000 historic photos, including photos of local monuments and historical events, as well as 130 maps, ranging from the 17° to the 20° c. An old family library, initiated in the late 18th century, which now includes about 30.000 items, including 1500 ancient volumes. It also has an original library catalogue from 1802.	In-house expertise in librarianship and information science, no IT expertise	Aware of metadata, use of Dublin Core

Table 4: Type of collection – IT skills – Knowledge about metadata

Content provider	Digital Repository	Availability through Europeana
PSNC (Poland)	A digital repository exists, where collection is stored in dLibra-based digital library.	No
KUAS (Denmark)	A digital repository exists, called Regin.	No
BJC (Romania)	A digital repository exists, called Greenstone software.	Yes, through EuropeanaLocal Romania
RCE (Netherlands)	micrsoft MDB, ExpoLab:MatriXML, OpenText DMS ArcGIS (ESRI) RNA (sematic Network) Adlib Beeldbank	Yes, through CARARE
NPU (Czech Republic)	3 databases: 1) Information System on Archaeological Data consists of the the State Archaeological List (SAL) of the Czech Republic and the database of Significant Archaeological Sites. 2) Geographical Information System 3) Metainformation System	Yes, through CARARE
VUKF (Lithuania)	A database and portal, which serves as digital library/archive.	Yes, through CARARE
UoY ADS (United Kingdom)	A bespoke collections management system.	Yes, through CARARE
IPCHS (Slovenia)	No information	No information
Provincie Limburg (Belgium)	A complex custom-built (since 2006) system that would currently be called 'aggregator' in Europeana contex, consisting of several modules.	Yes, through Erfgoedplus.be
CG33 (France)	Archival software named Pleade from AJLSM	Yes, through EuropeanaLocal
Zavad Jara (Slovenia)	A digital repository exists, called KAMRA.	Yes, through KAMRA
Future Library (Greece)	A digital library does not exist.	No
FMNF (Portugal)	For the archive collection we use the software Fortis; for the museum collection we use the software inPatrimonium.	Yes, through EuropeanaLocal Portugal
AIT (Austria)	OAI PMH	Yes, through

		EuropeanaLocal
ABMR (Sweden)	CollectiveAcess (collectiveaccess.org) and a local database solution called Theodor.	No
PSRL (Bulgaria)	Digital library is developed by Public Library - Varna team.	Yes, through Public Library - Varna
BGB (Serbia)	A digital repository exists, where collection is stored in dLibra-based digital library.	No
HU (Turkey)	A digital repository exists, called MIDAS Otomation System based on Dublin Core.	No
CUT (Cyprus)	No digital library for collections of the small content providers	No
AHAI (Iceland)	Use of File Maker PostgreSQL	Yes, through CARARE
PrifUK KAEG (Slovakia)	No digital repository	No
DP (Ireland)	A digital repository exists, based on Dspace.	No
FRS (Italy)	A digital repository exists, called SAMIRA.	No

Table 5: Digital repository – Availability through Europeana

Annex II – Report and analysis of the results of the Locloud surveys and the state of the art

As part of LoCloud WP1: Planning, preparation and requirements, a series of activities were carried out to inform the LoCloud content action plan and to develop the aggregation infrastructure. These included:

- A review of the state-of-the-art in cloud based services relevant to the needs of small and medium institutions;
- A survey of metadata schemas relevant for use in LoCloud;
- A survey of content and metadata amongst organisations participating in LoCloud.

Together the results of these surveys informed the requirements analysis reported in this deliverable.

This annex provides a systematic report of the complete analysis of the results of the survey and the state of the art.

State of the art monitoring and situational analysis

D1.1 reported on the monitoring of the state-of-the art of cloud computing and made an assessment of aspects of the cloud relevant to the needs of the project and to small and medium sized institutions. The survey found that:

There are many advantages to cloud computing which could be taken advantage of by cultural and heritage institutions. Cost-effectiveness and access to resources beyond the abilities of the individual institutions are among the primary advantages.

There is high awareness and willingness to participate in cloud-based development from the heritage institutions and agencies voicing their opinion in this report. However, both a literature review and our own situation reports show a hesitancy to delve into the new service offerings on their own. Hesitancy is mainly based on lack of knowledge and skills. The aim of the LoCloud project - to create a Best Practice Network that will support institutions in making their content and metadata available to Europeana by using a cloud-based technology infrastructure – therefore is very relevant at this point in time.

There are a number of SaaS providers providing services for the cultural sector. Some of the commercial vendors of collections management systems offer cloud-based versions of their software, and in the library domain the OCLC offers a number of relevant services. However, none of these come with plug-in aggregation tools for Europeana. There still is a need for online tools with a very low barrier to entry, which are suited to the needs (and budgets) of smaller local and community museums. This is the window of opportunity for the LoCloud project.

The LoCloud community consists mainly of small and medium-sized institutions, which lack financial and intellectual resources to create, purchase, regularly update and maintain software tools and services for digital heritage use. Therefore the scope of LoCloud project is to create these required Cloud computing based micro-services and Cloud computing tools in the area of metadata interoperability, content aggregation and harvesting, informational infrastructure, multilingual

controlled vocabularies, historical place names, geolocation and metadata enrichment, usage of Wikimedia applications and professional networking.

However, there are possible risks, which must be taken into consideration, and the infrastructure must be designed with security in mind and considered a priority for every application, service, and network solution which is provided. LoCloud is monitoring partners' present and future needs in view of cloud-computing and the possibilities of developing a shared vision of infrastructure. The LoCloud team is also monitoring what is emerging from Europeana Cloud in the way of infrastructural thinking and will take it into account in recommendations to the CH partners in LoCloud.

Definition of metadata schemas

D1.2 reported on a review of metadata schemas in use amongst LoCloud partners and in prior projects, such as Europeana Local, Athena and CARARE. D1.2 made recommendations on intermediary schemas to be used in LoCloud as intermediaries to EDM, and work needed to support their implementation by LoCloud.

One of the LoCloud's main objectives is to ensure interoperability between the native metadata held by heritage organizations and the metadata used by Europeana. This deliverable describes the identification of the metadata schemas to be used within the project as intermediaries to Europeana Data Model (EDM), the schema introduced by Europeana to improve on ESE, its first data model.

EDM is a comprehensive, general-purpose model which accommodates the range and richness of community standards such as LIDO for museums or EAD for archives. The current implementation of EDM by Europeana is part of a roadmap which covers the transition from the initial ESE schema to full implementation of EDM. It is likely that the detail of EDM implementation will change over the next four years.

Following past experiences on domain aggregation services, such as CARARE, Athena and MIMO projects, LoCloud will bridge the gap between the metadata schemas in use by their content partners and the data model being developed by Europeana by establishing or utilizing intermediary schemas and then mapping these to the metadata standard implemented by Europeana. This mapping has enabled harvested metadata to be transformed to EDM for supply to Europeana.

The benefits of this approach are:

- providing an intermediary schema accommodates the need for guidance which is relevant to the needs of a community and the particular characteristics of their data;

- mapping the intermediary schema to EDM builds in flexibility and accommodates any changes by Europeana as it phases in the full EDM model. This approach will be continued and developed in LoCloud.

A survey was carried out amongst LoCloud content partners to undertand their preferences regarding intermediate schema, and further information was gathered during the content survey (described below). Based on the analysis of these results the following intermediate schemas were recommended for LoCloud:

1. For material which is moveable (like museum items, etc.), the use of the LIDO schema as intermediate schema.

2. For material which is territory based (like monuments, archeological items, etc.), the use of the CARARE schema (versions 1.x and 2.x) as intermediate schema.

3. For archive materials (collections of manuscripts, etc), the use of EAD as intermediate schema.

4. Use EDM as intermediate schema for those providers who are currently exporting metadata following EDM to make mappings to the current implementation of EDM by Europeana.

5. Expand the aggregation workflow in LoCloud and provide mappings to automatically convert the information from ESE to EDM. This option will be useful for those providers who are currently exporting metadata following ESE.

Content and Metadata Analysis

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.

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 were presented in D1.3. 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.

The content survey gathered information 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.

As a result of the information gathered from partners on their metadata, the metadata schemas recommended for uses as intermediaries to EDM the content and metadata survey suggested that most providers can more easily deliver their metadata in CARARE, LIDO, EAD or a form of extended Dublin Core.