LoCloud

Local content in a Europeana cloud

D5.1: Periodic Report on content ingestion

Authors:
Gunnar Urtegaard (NRA)
Kate Fernie (MDR)
Runar Bergheim (AVINET)
Silvia Alfreider (NRA)

Version: Final

LoCloud is funded by the European Commission's
ICT Policy Support Programme

cip competitiveness and innovation framework programme 2007–2013

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 Introduction

LoCloud will put in place an infrastructure that will continue to increase the content available to Europeana. At the same time the project will enhance the skills, expertise and motivation required to support local institutions throughout Europe. Key content types to be made available through LoCloud include items and collections of high cultural value held at local or regional level, specific local collections held by libraries, museums and archives, local sound and film archives, public records held by archives, etc. LoCloud's role is thus to help to create the conditions, through for example, training, advice, providing suitable mechanisms whereby local and regional institutions are in a position to contribute their content to Europeana.

As described in the Description of Work the objectives of WP5 are to:

1. Monitor the achievement of the objectives of LoCloud and their impact on the user communities
2. Monitor and evaluate the amount, types and quality of metadata and content being provided to Europeana by LoCloud

This deliverable, D5.1, will focus on how to monitor the preparation and ingestion of metadata in its various phases in LoCloud. As central tool LoCloud will use a modification of the Events Log system used in Europeana Local.

2 Amount, types and quality of content

An important goal for LoCloud is to increase significantly the quantity of high quality digital content accessible through Europeana, provided by small and medium sized cultural institutions. Many small, local institutions have limited IT infrastructure and lack either the requisite staff skills in digitization and digital libraries. Despite this, their collections are important and interesting in the context of Europeana and its users. Increasingly these collections are being made available online and therefore have the capacity to contribute to Europeana.

LoCloud will explore the potential of cloud computing technologies for Europeana, both in the direction of an easier to use and a more efficient infrastructure and in the creation of a range of software services which will benefit content providers and users. For this purpose, by month 21, Athena RC and PSNC will make available a lightweight digital library system for small and medium sized institutions, delivered in the cloud and compatible with the standards established by MINT, MoRe and Europeana, based on their existing work in this field and suitable for deployment where local infrastructure is lacking.

For other institutions with access to IT infrastructure, tools will be specified and delivered by NTUA as cloud services to prepare content and metadata online in repositories and using the tools provided to support metadata harvesting, for example: the MINT metadata ingestion tool and OAI-PMH harvesters such as Repox. This will be available from month 16.

For small institutions with limited technical capacity, AVINET will specify and provide tools to prepare content online with embedded EDM metatags ready for capture by crawler services, to be incorporated in the LoCloud repository by month 21.

The infrastructure and tools described above will each be tested by a group of partners representing each of the above circumstances and adapted as necessary by Month 24. Following successful testing, the tools and systems will be implemented for live harvesting of the new content listed in the content table, utilizing the range of methods made available by LoCloud from Month 25 onward.

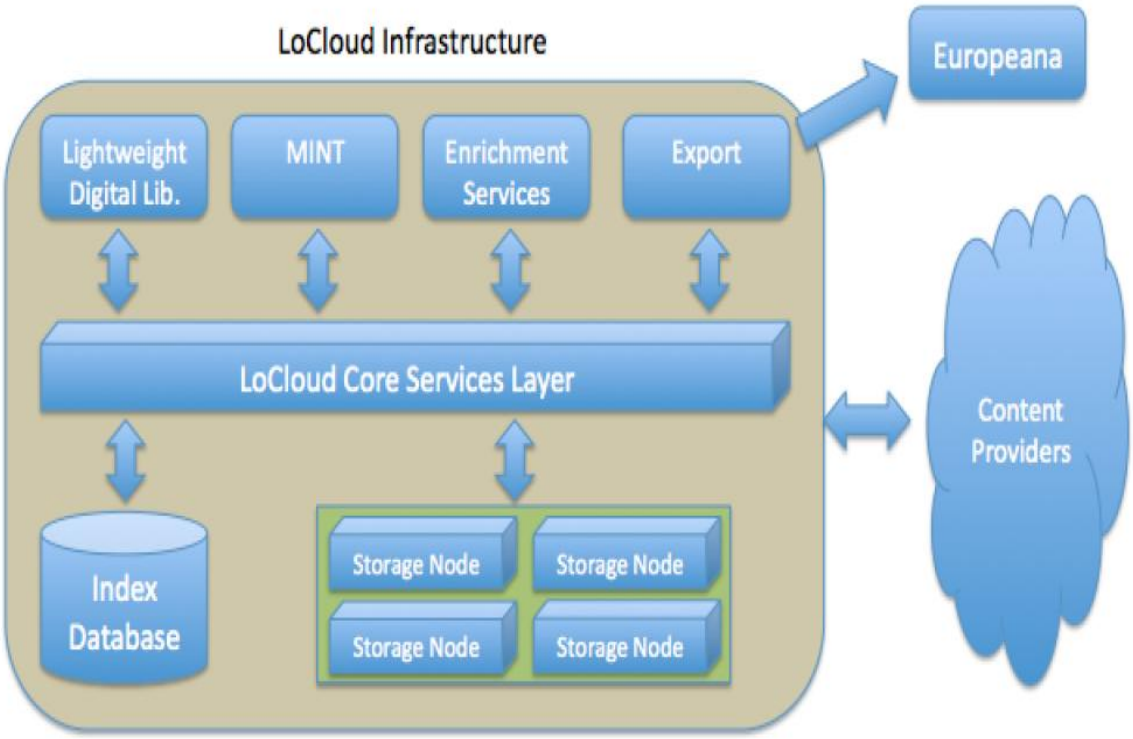

Illustration of the cloud based infrastructure and services layer of LoCloud.

The expected progress in terms of amount and types of content is 4 million items by end of year 3. This deliverable will describe how we intend to monitor progress through year 2 and further in year 3. No content is expected to be added in year 1.

Table 1: Some Examples of planned digital content that will be made accessible and re-usable by the end of the project:

Provider	Quantity and type	Subject matter (topic or theme that content is about).	Language	Format	Existing Metadata	IPR
VEKAM - Turkey	820 Postcards / Photos	Postcards and photos that reflects various views of Ankara including buildings, monuments, streets, archaeological remains etc.	TR	JPEG,TIFF PNG	Local metadata standard (Dublin Core based).	for non-commercial and commercial use (for commercial use payment is required)

VEKAM - Turkey	180 Maps and plans	Maps and plans of the Ankara	TR	JPEG	Local metadata standard (Dublin Core based).	for non-commercial and commercial use (for commercial use payment is required)
ADS – University of York	>1400 Text	Grey Literature – Archaeological Reports	English	PDF	ADS (Extended DC, plus MIDAS)	ADS – T&Cs, for non-commercial use with attribution
ADS – University of York	c.450 RD metadata	ADS archived collections RD metadata	English	XML	ADS (Extended DC, plus MIDAS)	ADS – T&Cs, for non-commercial use with attribution
ADS – University of York	134 Volumes (c. 4000 articles/files)	Proceedings of the Society of Antiquaries of Scotland	English	PDF	ADS (Extended DC, plus MIDAS)	ADS – T&Cs, for non-commercial use with attribution
ADS – University of York	130 Text	CBA Research Reports`	English	ADS (Extended DC, plus MIDAS)	ADS – T&Cs, for non-commercial use with attribution	PDF
ADS – University of York	96 Volumes (c 960 articles/files)	Surrey Archaeological Collections.	English	ADS (Extended DC, plus MIDAS)	ADS – T&Cs, for non-commercial use with attribution	PDF
Institute for Spanish Cultural Heritage	15.000 photographs	Buildings and archeological sites	Spanish	JPEG/TIFF	ACCESS database	CC0 1.0
SG for State Museums	20.000 museum objects	Different collections of museums: plans, maps, photographies, memories and archaeological material….	Spanish	JPEG/TIFF/PDF	DC, ESE, METS	CC0 1.0
SG for Library Coordina-tion	15.000 maps 60.000 photographs	Local plans and maps People and landscape and building sites	Spanish	JPEG/TIFF/PDF	MARC21, ESE, DC, METS	CC0 1.0

3 Content survey and metadata Analysis

During the planning stage of LoCloud in WP1, the Athena Research Centre, in association with content partners, made a first evaluation of the content to be ingested from the partners and began to identify new content that may become available during the life of the project, as 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 networks.

This first evaluation was done by completing an online questionnaire survey in August 2013, and by organizing three Workshops during September 2013 in Copenhagen, York and Madrid. The examples in table 1 above show how the survey identified 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.

At the workshops the content providers' content and metadata was discussed and their needs and extraction requirements were identified. The results show that content providers have a large variety of content as summarized in Deliverable 1.3: content and metadata analysis. During the workshops the partners also discussed the ingestion methods. In Deliverable 1.5 it emerges that the situation is quite complex, as many providers have their own views and established practices on how to deliver content for ingestion.

1) 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.
2) 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.
3) Medium sized content providers would either fall into the first category or would expect a plugin that will export their data to an aggregator.
4) 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.

Starting at Month 16 as part of WP5, MDR and NRA will begin monitor and evaluate the amount types and quality of metadata and content provided to Europeana by the LoCloud partners.

As a central tool LoCloud will use a modification of the Events Log system used in EuropeanaLocal. In the context of this tool, the term event relates to the individual collection and is intended to document what is done with each digital collection relevant for LoCloud from the time the collection is entered into the content survey in the Event Log from partners or from new content providers.

Data about each event will be added to the online event log. Some manual reports will also be needed to provide information about lessons learned and to identify good practices. These reports can be uploaded to the event log and made available to other partners. Each partner will be responsible for their own reporting in the Event Log and also for reporting about new collections from new providers added later in the project.

The quality of metadata and content should be handled differently. Techniques will be developed by a combination of technical reports and manual reporting. Once we have metadata uploaded to the Locloud infrastructure from different providers, the metadata will be analyzed to find out more about quality. Methods and reports will be developed. These methods can be applied to individual collections or all the metadata in the LoCloud infrastructure.

Several discussions about monitoring the content and metadata workflow have taken place, including at the LoCloud General meeting in London, November 2013 and during the Project Management Board Skype call meeting in January 2014. The partners concluded that it is important to establish a service that will give access to updated information about progress in each country or region and participation from the different domains, and will provide the possibility to trace each collection from local content provider to Europeana. Quality issues were also discussed.

The group stated that it is important to get key information equally from all partners/regions. The system established to collect this information must not be too difficult or involve too much work. It should give flexibility for those who wish to delve deeply into particular questions and share their thoughts with the rest of the project partners.

The working group thus concluded that the monitoring and evaluation of the amount and types should be handled by a combination of analyzing metadata in the LoCloud infrastructure and data added by partners about each collection in the Event Log.

The monitoring and evaluation work should follow each of the phases involved in the passage of a collection from a content provider to Europeana, and it should document the work done in each phase. This approach also creates a stronger link between the technical work done with functionality in the repositories and reporting facilities at collection and repository level. The partners concluded that most of the data needed about amount and types could be collected by developing reports from the different phases in the lifecycle of a collection starting by entering data about collection in the online content survey in the Event Log.

4 Lifecycle of a collection

The Partners in LoCloud are aware that it is important to follow the lifecycle of a collection from LoCloud provider to Europeana and provide important data from each phase in the lifecycle. Each collection candidate for provision to Europeana through LoCloud must be registered in the Event Log. The online content survey in the Event Log will contain important information about each collection and must be maintained and kept available for new content providers and new collections throughout the project. Important documentation about each collection are made available to the project and the evaluation group through this tool.

There is no universal definition of the term collection. In LoCloud collection is used about the content listed in the original project application – as well as content from any new data sources associated with the project throughout the execution period.

One database/dataset with a uniform structure can hold more than one digital collection in a more traditional use of the term. However, from a technical perspective, a digital collection is all the data which can be extracted from the same set of tables in one database, carrying the same set of attributes – without heed to thematic divisions based on content types, themes etc.

The following activities are important milestones to document at collection level. At the point:

- When the collection is added to the content survey in the Event Log.
- When metadata is extracted about all items in a collection and mapped to a registered LoCloud intermediary schema using the MINT tool. (i.e. the process of "reading" metadata from local collection management systems, mapping them to a target metadata profile and writing them into a format readable by the chosen repository technology).
- When metadata is normalized. (i.e. the process of transforming attribute values from one notation to another. E.g. a standardized way of expressing dates, transformation of coordinates etc. This may apply to some collections or only some items in each collection, but may be not to all).
- When the normalized metadata is ingested into the LoCloud repository.
- When metadata is enriched. (i.e. automatically or semi-automatically processing of metadata with the purpose of improving the quality of what, who, where and when metadata. This may also apply to some collections or items, but may be not to all).
- At the point when collection are harvested from the LoCloud aggregator by Europeana. (i.e. the process of connecting to a repository, issuing a request for data and downloading metadata content as XML).
- When the collection is added to Europeana services and is available to end-users.

5 Establishing the LoCloud Event Log

The Event Log will be a simple, but powerful tool to monitor amount and progress. At the same time the Event Log will be open to all partners and content providers as a place to add comments on framework, tools, methods and standards being involved in the project.

The content survey contains the important information about the collection, such as the amount of items, objects, thumbnails, metadata formats, technical formats etc. The event log will collect and document work done on this particular collection from the original content provider to Europeana and any problems, questions and considerations done by the provider or the aggregator. Users may add manual reports to each event about technical issues, the use of EDM, normalization problems etc.

The Event Log will be a simple database with suitable reporting and analyzing functions. The services will be available from the LoCloud web site. When a content provider or aggregator has done work on one or several collections, he/she will connect to the Event Log, select the collection involved and add information from a predefined list of events. At a minimum this can be just some very simple core facts that will take a few minutes to add, sufficient to document the action and enter amount of items and objects and date the work was done. This will be enough to maintain data about amount and progress.

If the content provider or the aggregator encounters any problems or wishes to share lessons learned and ideas about improving the processes, this can be added as a manual report and thus being made available to the rest of the partners in the project.

The aggregator of a collection from a partner will connect to the Event Log, search for this particular collection and at the simplest just add data about extraction and mapping, and the number of items and objects effected. Mapping, normalization and enrichment will only affect metadata, and thus different numbers will be reported.

Information to be added about each event:

Info	**Comments**
Type of event	Selected from listbox
Date	Automatic
Number of items handled	Normalization, mapping and enrichments only effects metadata and items, and a selection of items
Number of objects	Only effects extraction and harvesting
Name of person responsible	Elin Østevik
Email of person responsible	Elin.ostevik@sfj.no
Issues related to the event	Short comments
Lessons learned	Comments
Uploaded documents	Manual reports and documentation

The technical partner, Avinet, will set up the Event Log as a prototype by May 2014 and as a fully functional service by July 2014 (Month 16), in time for monitoring the amount of data transported

from the local and regional providers to Europeana. A handful of partners responsible for the first collections to be harvested will test the Event Log.

Example LoCloud Event Submission Form:

Example LoCloud Event Submission Form

Type of event:	Harvesting
Date:	15.06.2014
Number of items handled:	
Number of objects:	30.000
Name of person responsible:	
E-mail of person responsible:	
Issues related to the event:	First attempt failed due to network time-out issues
Lessons learned	Must time the harvesting so that it doesn't coincide With other types of heavy network usage
File upload	Browse ….

Submit event

Fig 1: Example of event submission form

6 Reports from the event log

The Event Log will be an important tool for collecting and sharing information about the amount of content and progress in LoCloud. Reports can be selected by country, region, type of content, time span, amount etc. The number of collections will not be very large, even at the end of the project.

The Event Log will document what happened to each collection. Search and reporting facilities available for all partners will be developed. The Event Log will enable many interesting analyses of the data to be performed, for example:

- The Event Log will show at any given time the amount of items and objects being extracted from local providers and the amount added to Europeana. This will be a report organized by country and updated on a daily basis by the system collected data from the Event Log
- How many collections, items and objects are entered into the content survey in the Event Log at any given time? (Can easily be arranged by country, domain etc)
- How many digital objects from a chosen number of collections are extracted and mapped to Europeana by a certain date.
- How many items were affected by metadata normalization and enrichment from a country, a region, certain types of collections etc in a given period of time
- How many item and objects are harvested into LoCloud repository by any date, country, region, provider type (museum, archive, library)
- Items and objects harvested by Europeana by end of year 2, end of year 3.
- The number of providers and collections involved in LoCloud at any given time.
- The number of manual reports added, from which country, provider, etc

7 Event log and progress reports

The Description of Work indicates that a progress report on content ingestion (D5.1) should be delivered every 6 Months, starting from year 2, showing progress based on a predefined set of key metrics. The Event Log will be the main tool for creating data for these reports. Data will be downloaded from the Event Log and analyzed and results reported in D5.1. The manual reports added to the event log will also constitute an important resource to be used in these progress reports.

The manual reports added to the Event Log will give us access to experiences collected by providers and partners and aggregators throughout the project.

Information to the partners about the Event Log

By June 2014, when the Event Log is established, a manual will be ready and distributed to all the partners giving them information about the Event Log and how to use it as a tool to document own actions and get access to lessons learned by others.

This manual will describe in detail how to use the Event Log, why it is established and how it will benefit the project and document the work done.

8 Metadata quality

During the workshops the issue of metadata quality was discussed. What is good metadata and how can it be measured? It is important that the metadata contributed to LoCloud should be of as high a quality as possible. Thus, we should have a way to measure their quality.

The diversity of the metadata being made available to LoCloud by partners makes it hard to compare quality or to use global metrics. For example, different criteria apply to free-text metadata than to controlled vocabularies.

The institutions that are participating in LoCloud are using the following metadata schemes in their repositories: Dublin Core, Extended Dublin Core, EDM, ESE, CARARE, LIDO, EAD, TEI, and local schemas (MAG, ARUODAI).

All of the metadata provided by partners will be mapped to LoCloud's specified intermediary schemas and transformed to EDM for provision to Europeana. This process allows for a number of quality measures to be evaluated, such as:

- Completeness of mandatory and recommended elements; support for search facets
- Currency of links
- Accuracy of geographical coordinates (for example, points lie within the expected region when mapped)
- Availability of descriptive text and its length
- Availability of titles or captions and their length
- Use of controlled vocabularies
- Use of international controlled vocabularies recommended by Europeana
- Provision of rights statements
- Availability of thumbnails or previews
- etc

The group (= NRA, AVINET, MDR Partners, NTUA, DCU, IPCHS, UDE) will monitor the metadata quality and produce a report based on the results, focusing mostly on what metadata ideally ought to be present in new collections in order to produce the best results.

9 Conclusions

This deliverable summarizes the work of the Evaluation Working Group to date. The group is made up of representatives from the following partners: NRA, AVINET, MDR Partners and UDE (UDE is the work package leader).

The work presented here focuses on the creation of an Event Log which will be used to monitor and evaluate the amount, types and quality of metadata and content which are being contributed by LoCloud partners to Europeana. The log will be used to collect evaluative data at key points in the content lifecycle.

This data and the reports generated through the Event Log will provide an invaluable source of user feedback to Europeana and other interested parties, which can be used to inform both modifications and further developments to the systems and processes they make available to content providers.

Over the next period the group will review the operation of the Event Log, based on practical experience from its use as data begins to be harvested from the first group of partners. They will use the data from the log to complete the first evaluation report (D 5.1) and will consider how best to conduct the impact study.