

EUROPEANALOCAL

Technical Primer

Overview of Europeana Local Implementation Steps

**by Stein Runar Bergheim with contributions from Gordon McKenna, Gunnar Urtegaard,
Mary Rowlatt, Rastislav Rehak, Rob Davies and Robina Clayphan**

Revised version 16.12.2008

This document gives a brief overview of the tasks to be executed by Europeana Local Content Contributors with aid from Technical Partners and supporting material supplied by the European Digital Library Foundation and other Europeana related projects and initiatives.

Introduction

This document gives a brief overview of the tasks to be conducted by content provider who wishes to contribute his or her organization's content to Europeana Local throughout the project. The topics covered in this document may change with the development in technologies and the evolution of the Europeana service.

The following elements are covered briefly:

- **Digital collections:** the characteristics of digital collections and what they are in the context of the Europeana Local project
- **Repository installation:** the steps involved in choosing and installing a suitable repository technology
- **Metadata Transformation:** an abstract term covering metadata extraction, metadata normalization and metadata enrichment.
 - **Metadata Extraction:** 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
 - **Metadata Normalization:** the process of transforming attribute values from one notation to another. E.g. standardized way of expressing dates, transformation of coordinates etc.
 - **Metadata Enrichment:** automatically or semi-automatically processing of metadata with the purpose of improving the quality of *what*, *who*, *where* and *when* metadata
- **Repository Population:** the process of loading data into the chosen repository software based on custom metadata extracts or export of known standard formats
- **Metadata Harvesting:** the process of connecting to a repository, issuing a request for data and downloading metadata content as XML
- **Aggregator Repositories:** an explanation of what an aggregator repository is compared to a content contributor repository and what it can be used for
- **End-user Services:** a brief discussion on existing Europeana end-user services and potential future services utilizing Europeana and OAI-PMH infrastructure.

This document serves as an introduction to the Europeana Local Implementation Manual, a comprehensive document which forms the basis for technical workshops and which serves as a reference during implementation of OAI-PMH compliant repositories.

The Europeana Local Implementation Manual will explain in more detail the specifics of each of the above steps and thereby clarify any questions which may arise from this introduction. The document will be released prior to the technical workshops.

Repository Implementation Workflow

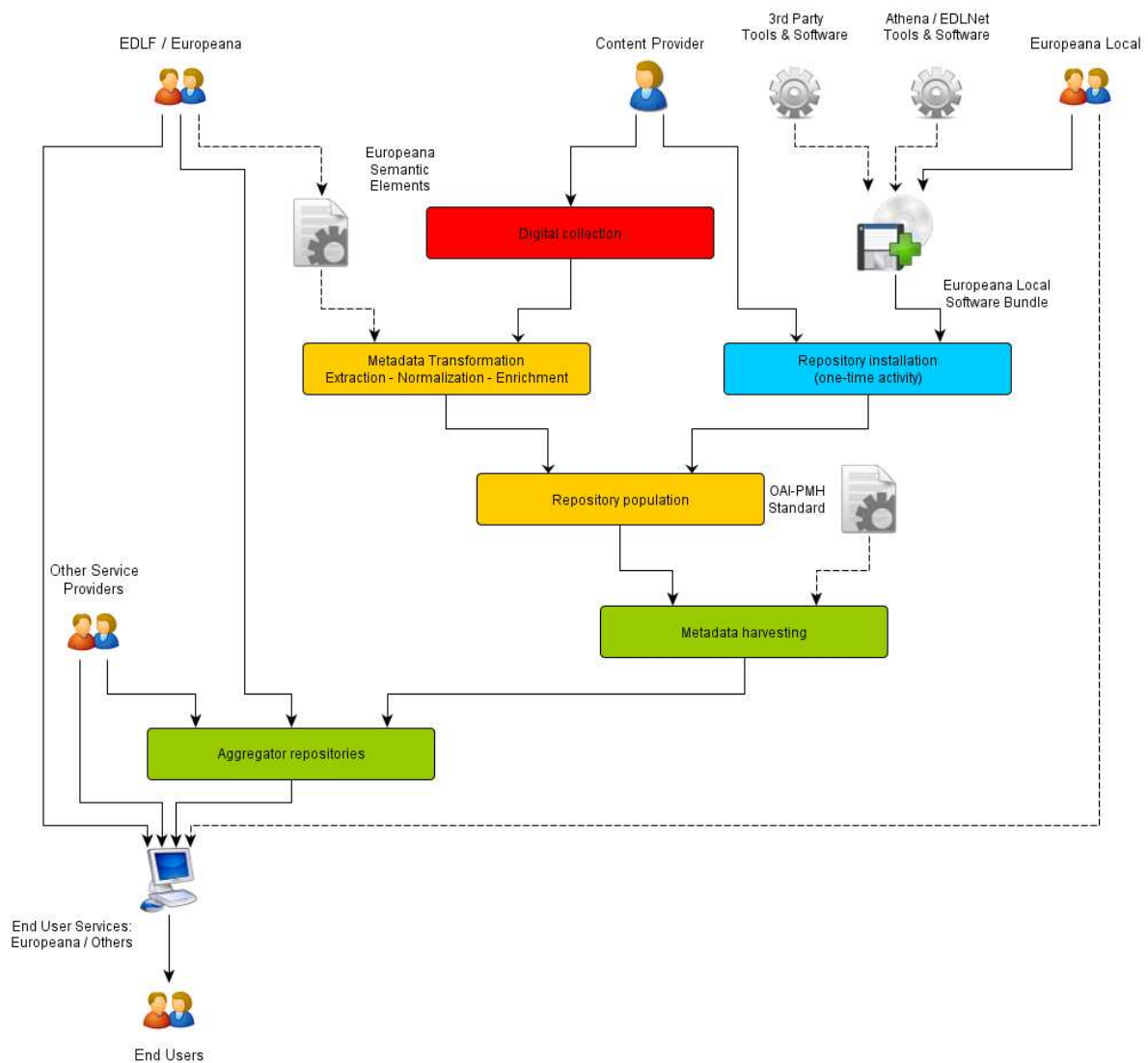


Diagram 1: The above diagram shows the process of implementing OAI-PMH compliant repository solutions on top of digital collections held by local content providers – thereby contributing content to the overall Europeana information infrastructure. The red boxes indicate existing collections, the blue boxes indicate one-time activities such as repository installation, orange boxes indicate tasks to be undertaken by content providers and green boxes indicate tasks to be executed by Europeana. Solid lines indicate direct participation, stippled lines indicate indirect contribution in the form of supporting documents, standards etc.

Digital collections

A digital collection is an abstract term used to reference the whole or part of a database. There is no universal definition of the term, but in Europeana Local it 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 with a uniform structure can hold more than one digital collection. However, from a technical perspective a digital collection is all 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.

This distinction is useful in the implementation process as more than one collection may be added to a local OAI-PMH compliant repository through one and the same metadata extraction/repository population operation. The workload of each content contributor partner largely depends on the number of different systems data have to be retrieved from. If an institution holds 10 collections in one collection management system it is a limited task – if an institution holds 10 collections in 10 collection management systems the task multiplies tenfold.

Repository installation

The first step on the way to implanting a local collection into the Europeana infrastructure is to install a repository software locally – or to enable and configure repository software which already exists in the local collection management system.

This task involves a number of steps:

- Choosing a suitable repository software (Europeana Local will make and offer support for a recommendation solution) based on existing software, operating system and platforms.
- Installing the repository software on local hardware
- Configuring the repository to work in connection with existing or embedded web server functionality and ensuring that the OAI-PMH web service protocol responds in a standards compliant manner to any requests from external harvesting repositories

This process will be covered in great detail in the EuropeanaLocal Implementation Manual to be circulated to content contributors prior to Europeana Local technical workshops in the winter of 2009.

Metadata Transformation

Metadata transformation is the process of moving data from its source collection management system, mapping to a target metadata profile, normalizing content attributes and enriching the metadata through various automatic and semi-automatic techniques.

Metadata extraction from local collection management system to local repositories

Once a local repository is installed and functioning, the process of extracting metadata to populate the repository can be initiated. This process can be simple or complex, potentially requiring a lot of manual customization work and entry-level programming – it all depends on whether the local collection management system used to maintain the digital collection offers any standard forms of export, e.g. MARC records, which can then be normalized and automatically be imported by the repository software.

If no export option exists, the alternative is most often to write simple SQL-scripts which connects to the data source (typically a SQL-compliant RDBMS) and reads metadata directly from the physical data structure. This requires extensive knowledge about the individual system and may require scripts to be modified as software gets upgraded. It is therefore preferable to use commercial repository add-on modules to existing collection management systems if such exists, to implementing custom ones.

Metadata will be transformed into an element set defined by the European Digital Library Foundation called Europeana Semantic Elements. This is an application profile based on qualified Dublin Core with the addition of a few custom elements to support extended functionality in the Europeana end-user service.

Metadata normalization

Metadata normalization is the process of harmonizing attributes which express the same information with different formatting and or notations and to ensure the data required to support Europeana functionality is present in each record. This is necessary to allow interoperability between content which originates in heterogeneous sources. The current Europeana prototype uses temporal metadata – dates – to visualize items along a timeline. This would not be possible unless dates were expressed according to a strictly defined format by all content providers.

Example: One collection may express dates like this “January 12, 2008” (Month dd, yyyy) while another expresses them like this “12/24/2008” (mm/dd/yyyy). To enable them to be used in the timeline function they will have to go through a normalization process to provide a common date format e.g. “2008” (yyyy).

Details on mapping and normalization will be found in upcoming versions of Europeana Semantic Elements and associated guidelines documents and the example above is merely meant to illustrate the process.

Metadata enrichment

Not strictly a part of the process of contributing content to Europeana, Metadata Enrichment can't be neglected if the goal is to provide content for good services – not just content.

Metadata enrichment is the process of trying to improve the quality of metadata both in the local digital collection but also once the metadata is to be used out-of-context and together with metadata from other sources including issues such as multi-lingual content, spatial references, temporal references, mapping to common vocabularies and identification of authorities.

Relevant standards, tools, and technologies that will be looked into throughout the project include:

- Dictionary description notations such as SKOS
- Translation utilities such as Google Translate, Babelfish etc.
- Indexing and keyword extraction utilities
- Location extraction utilities
- Date extraction utilities

Most metadata enrichment techniques involve some manual effort on the part of the content contributor. Most automated techniques are faulty and can lead to errors – the only way of making sure metadata has the proper quality is to make sure proper quality procedures are in place at the time of entry and maintenance through the collection management system.

Repository population

Once metadata has been extracted or exported from a source system, the next step is to setup automatic loading of the metadata into the repository. A schedule should be defined based on the

update frequency of the local digital collection. Some collections are finite and won't grow or change. In such cases – the loading of data can be done once – and no more.

Metadata harvesting

From this stage and onwards, the content has entered the realm of Europeana which is responsible for harvesting content and providing end-user services.

With all the previous steps being completed, harvesting repositories can now connect to your local repository, issue queries and download content – subject to being authorized from you as a repository manager.

Typically, the harvesting will function along the same lines as a search engine spider. An external web service will connect to your local repository web service once in a while (every night/every week/every month) and download all metadata – or delta sets containing the changes (inserts/edits/deletions) since “last visit”.

Aggregator Repositories

Aggregator repositories can in many cases be the same technology as the source repositories installed in your organization and mentioned above. However, these repositories will contain a web service requester which will scan through a list of source repositories according to a preset schedule.

This type of repository will get its content from harvesting source repositories rather than from local collection management systems. Such aggregator repositories include the underlying database of the Europeana Service and some national portals which base their services on harvesting via OAI-PMH but many more could – and most likely should – be added at regional/national level in order to justify the level of investments required to enable local content for the Europeana infrastructure.

End-user Services

The real benefit of enabling local content into the Europeana infrastructure comes through the exposure, cross-domain and cross-geography searches allowed by the Europeana end-user service. Here, your content will live alongside other content from numerous databases across Europe and may be rendered as result to any number of searches that you or anyone else never thought of performing.