# The European Library Standards Handbook

**Version 1.2 / December 2012**

**www.theeuropeanlibrary.org**
National Library of the Netherlands, NL 2509LK, The Hague

# The European Library Handbook

## Version 1.1


1 INTRODUCTION ..... 4

2 A LIBRARY DOMAIN AGGREGATOR ..... 5

2.1 Aggregation for Europeana ..... 5

2.2 A Network of Aggregators ..... 7

2.3 Aggregation beyond Europeana ..... 7

2.4 Rights Information Services ..... 8

3 SUBMITTING DATA FOR AGGREGATION ..... 9

3.1 Preparation ..... 11
3.1.1 Primary Contact, Questionnaire, Scheduling ..... 11
3.1.2 Sample Data, Analysis and Planning ..... 12

3.2 Ingestion ..... 12

3.3 Publication ..... 13
3.3.1 Publication in The European Library ..... 13
3.3.2 Publication to Europeana ..... 13

3.4 Updates and Deletions ..... 13

3.5 On-Demand Removal of Data ..... 14

3.6 Provision of the Europeana Data Model (EDM) for libraries ..... 14

4 OPERATIONAL REQUIREMENTS ..... 15

4.1 Information Requirements ..... 15
4.1.1 Information Requirements for The European Library ..... 15
4.1.2 Information Requirements for Europeana ..... 16

4.2 Technical and Metadata Requirements ..... 16

4.3 Administrative and Resource Requirements ..... 17
4.3.1 Content Acceptance by Provider ..... 18
4.3.2 Automatic Link Validation ..... 18

4.4 The European Library Aggregation Infrastructure ..... 18
4.4.1 Configuration Management (SugarCRM) ..... 19
4.4.2 Harvesting Infrastructure (REPOX) ..... 19
4.4.3 Unified Ingestion Manager (UIM) ..... 19
4.4.4 Normalisation, Enrichment and De-Duplication ..... 20

5 GOVERNANCE, LEGAL AND FINANCIAL ISSUES.....21

6 ANNEX.....24

6.1 ANNEX I: Glossary.....24

6.2 ANNEX II: Show Case.....27

6.3 ANNEX III: ESE Element Overview.....31

# 1 Introduction

The European Library is the library domain aggregator for Europeana. It helps libraries from across Europe to provide data to Europeana. As leader of the Europeana Libraries project, it has also created a robust aggregation infrastructure capable of handling thousands of libraries and millions of digital objects. For the Europeana Libraries project alone, the infrastructure aggregating some 5 million objects from 20 partners for Europeana.

This handbook is aimed at libraries which wish to submit data to Europeana and The European Library, and which wish to become sustainable partners of The European Library service. It clarifies both the process and the advantages of outsourcing specific data-provision tasks to an aggregator such as The European Library.

In brief, the process involves 3 steps:

1. A preparation phase in which the conformity with content strategies is checked, primary contact information is exchanged and sample data is processed. This phase is finalised by an agreed content ingestion plan.
2. An ingestion phase to process and enrich the data. This phase ends when the data provider validates and approves the enriched data.
3. A publication phase which makes the data available to Europeana and other services, depending on the agreements with the data provider.

Data can be provided in many different formats. The matrix in Table 1 (p. xxx) shows the various data formats in order of preference, from MARC21 via OAI-PMH (most preferred) to proprietary XML formats via FTP or HTTP (least preferred).

There are many benefits of this process for data providers.

- By using one aggregation system to reach complementary portals and services, content providers benefit from the widest possible exposure of their materials. Through Europeana they will reach the general public, including lifelong learners, school students, college students, special interest groups and expert amateurs. The Europeana Library portal will show off their objects to an audience of dedicated researchers, particularly in the social sciences and humanities. Providers can also provide material to a further variety of services, external to Europeana and The European Library.
- Application Programming Interfaces (APIs) will ensure that content can be discovered not only via Europeana and The European Library but also via a variety of other platforms, already embedded in the workflow and normal internet surfing patterns of users around the world.

The following pages expand on both the process of submitting data and the benefits of joining The European Library. The handbook also provides an overview of legal and financial issues to consider when providing content to The European Library and Europeana.

If at any point you need support or have questions about the aggregation process, please contact The European Library's Collections Team: collections@theeuropeanlibrary.org

# 2 A Library domain aggregator

This section clarifies the role of The European Library as the library-domain aggregator and draws a picture of current European initiatives surrounding the aggregation of cultural heritage material.

## 2.1 Aggregation for Europeana

As an aggregator to Europeana, The European Library helps individual content providers to submit cultural heritage data to Europeana. This is explained in Europeana's own Aggregators Handbook[1]:

> An aggregator in the context of Europeana is an organisation that collects metadata from a group of content providers and transmits them to Europeana. Aggregators gather material from individual organisations, standardize the file formats and metadata, and channel the latter into Europeana, according to the Europeana guidelines and procedures. Aggregators also support the content providers with administration, operations and training.

Generally, aggregation to Europeana is organised hierarchically (see Figure 1 below).

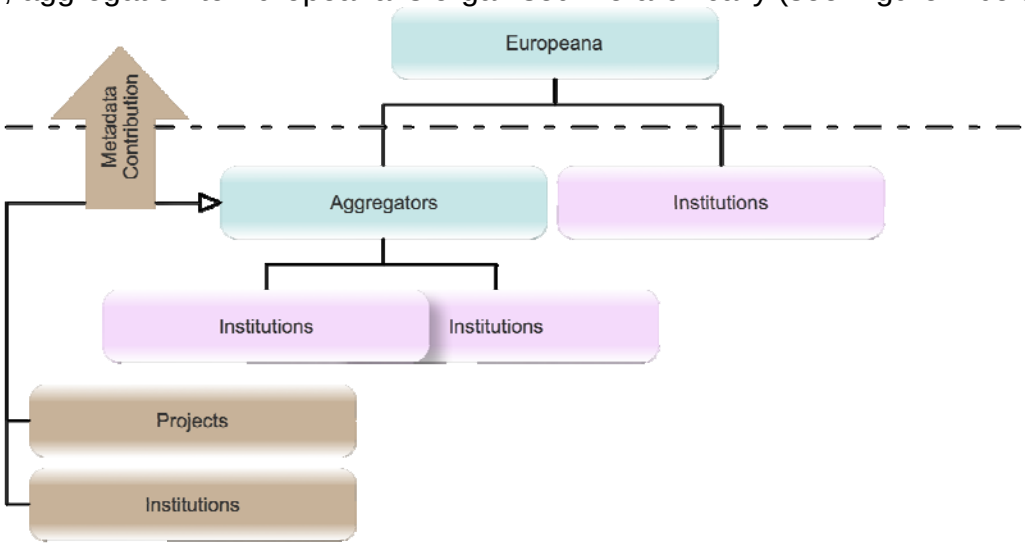



**Figure 1: Europeana Data Submission Pyramid**

This creates a model that can be expanded across Europe and beyond. It begins with institutions, which provide their data to an aggregator (project or organisation). The aggregator then processes the data and provides it to Europeana. In a minority of cases, an institution provides data directly to Europeana (usually because of limited representation of aggregators in countries, domains or themes).

Three aggregation types are currently shaping the aggregation landscape around Europeana – **Regional, Projects and Independent Organisations**. Each type of aggregator can further represent a geographic area (local, national or European), or a domain level (cross-, single, thematic).

[1] The Europeana Aggregator Handbook edition 2 is available at http://pro.europeana.eu/web/europeana-pro/provide_content

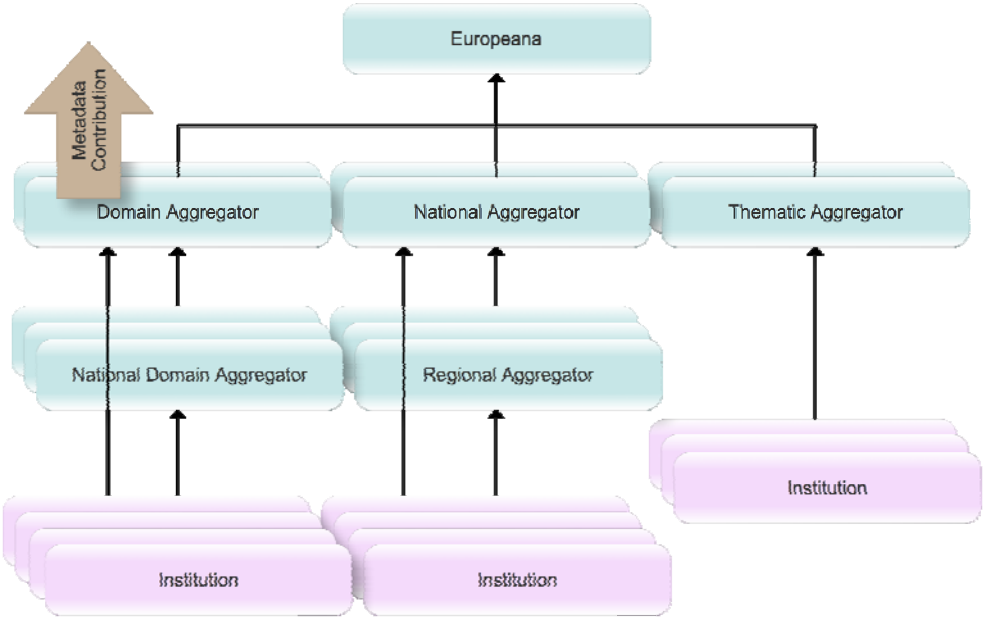



**Figure 2: A Hierarchy of Aggregators**

**Regional Aggregators:** Regional Aggregators bring together content from a specific geographic area. For example, there are many National Aggregators who aggregate from their national cultural heritage institutions. National Initiative Aggregators, are aggregators from one or several domains who have been appointed by their Ministry to take on the aggregation role in the country. National aggregators and National Initiatives are the first point of references for Europeana in accessing content from a particular country.

**Project Aggregators: These** aggregators are organisations which have joined a project consortium with a specific aim and purpose. They can either aim to aggregating within a specific theme or by single or cross-domain level.

**Independent Organisations;** Organisational Aggregators are independent organizations, which represent geographic and domain levels; thematic, single or cross-domain.

The European Library is an Organisational Aggregator for library content to Europeana and therefore has 3 main responsibilities:

1. To gather material from libraries.
2. To standardise the library domain formats and metadata formats to Europeana's preferred standard (**E**uropeana **S**emantic **E**lements (**ESE)** and the **E**uropeana **D**ata **M**odel (**EDM)** ).
3. To channel the data in ESE **or** EDM format to Europeana via OAI-PMH. Note that ESE and EDM are the defined metadata formats accepted by Europeana, and OAI-PMH is the preferred way of data exchange.

## 2.2 A Network of Aggregators

Figure 1 shows a fairly linear process but because a number of services are interested in metadata from the cultural domain, aggregation in the cultural sector has adopted a multi-level and multi-purpose community approach. In reality it is a network approach, as shown in Figure 3.

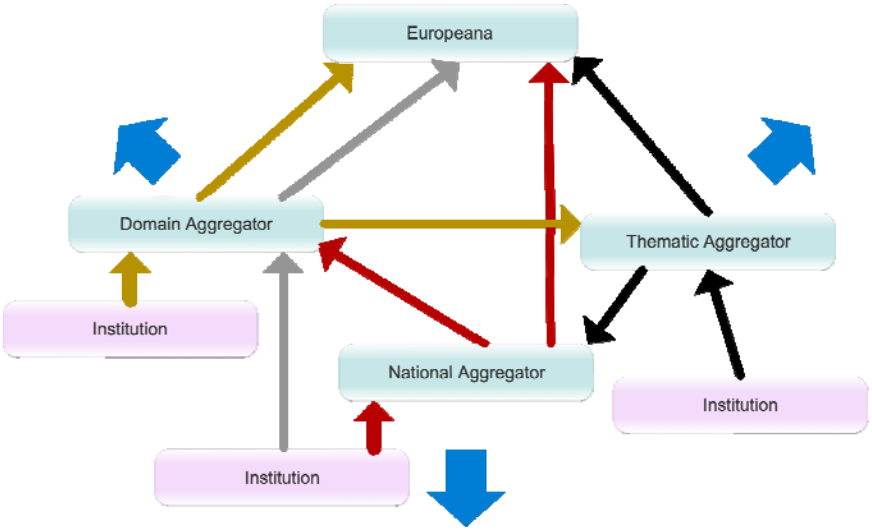



**Figure 3: Data contribution flow diagram**

Specialised and national aggregators aggregate metadata and content from data providers and institutions. This information is then provided to Europeana and to other projects such as research infrastructures and thematic portals.

Through this network of aggregators, metadata can be exchanged between institutions (with the permission of the providing institution). Providers have the ease of exclusively providing data to one aggregator. The task of further distribution is left to the aggregator. Since aggregators can then distribute data in big batches, this reduces the aggregation cost per record.

## 2.3 Aggregation beyond Europeana

The European Library has many strategic partnerships with fellow libraries and research organisations. These include:

- OCLC
- Serial Solutions (Summon) and Mendeley (reference management systems)
- DARIAH, CLARIN and EHRI (research infrastructures)

Data can be sent to these organisations as well as to Europeana.

The European Library has also recently forged new partnerships, with library groups such as LIBER and CERL.

Our experience with these partners (and data aggregation in general) is critical because formatting data for researchers is a complex task. It needs to stretch across domains (e.g. into archives and museums), and should meet the standards for various proprietary

formats. The European Library can provide that service efficiently, effectively and at a reasonable cost. Should standards change, we will be able to apply required metadata transformations to the data of multiple institutions at once, with only minor adaptations.

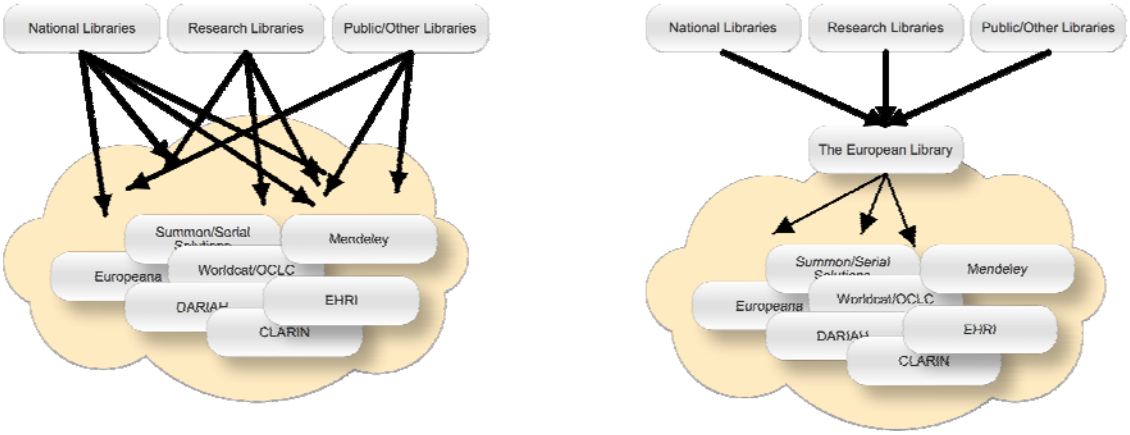



**Figure 4: Information Exchange without and with Aggregator Service**

## 2.4 Rights Information Services

One vision of the European Commission is to see more in-copyright material included in digital libraries such as The European Library and Europeana[2]. To facilitate this, our consortium of partners[3] has developed the ARROW system. It is a tool that helps libraries and other cultural heritage institutions to clarify the rights status of textual works. Once the rights are cleared for a particular work, it can be included in a digital library.

The ARROW system is currently available to determine the rights status of books published in France, Germany, Spain and the UK. This coverage will be extended to a further 12 countries through the ARROWplus project[4].

[2] http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1292
[3] Further information about the ARROW and ARROWplus projects is available at: http://www.arrow-net.eu/
[4] Further information about the ARROWplus project is available at: http://www.arrow-net.eu/what-arrow-plus

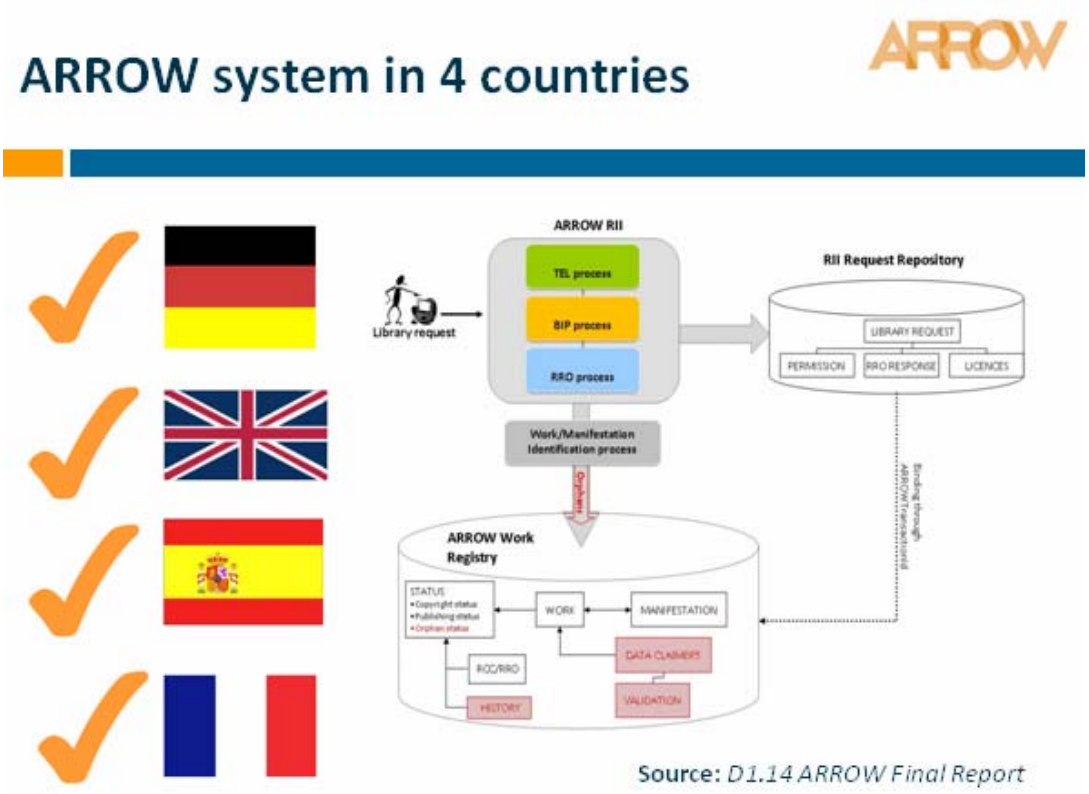



**Figure 5 :ARROW System**

As a key partner in the project, The European Library intends to integrate the ARROW system into the library-domain aggregation infrastructure and is investigating how use of the ARROW system can be made available to members of The European Library. Further information about using the ARROW system, including access to the test version of ARROW, is available from our team.

# 3 Submitting Data for Aggregation

This section gives an overview of the process of submitting data to The European Library, which will then pass the metadata to Europeana. An overview and process diagram is shown in Figure 6.

The first step is to determine if the content provider is already a partner in a content delivery project to Europeana. If so, they do not need to complete the Europeana Data Exchange Information form. If they are not already a partner, then the form must be filled out and validated by Europeana. At that point, Europeana will assign the content provider to an aggregator (in this case The European Library). The aggregator will then carry out the necessary aggregation work and support the content provider through this process.

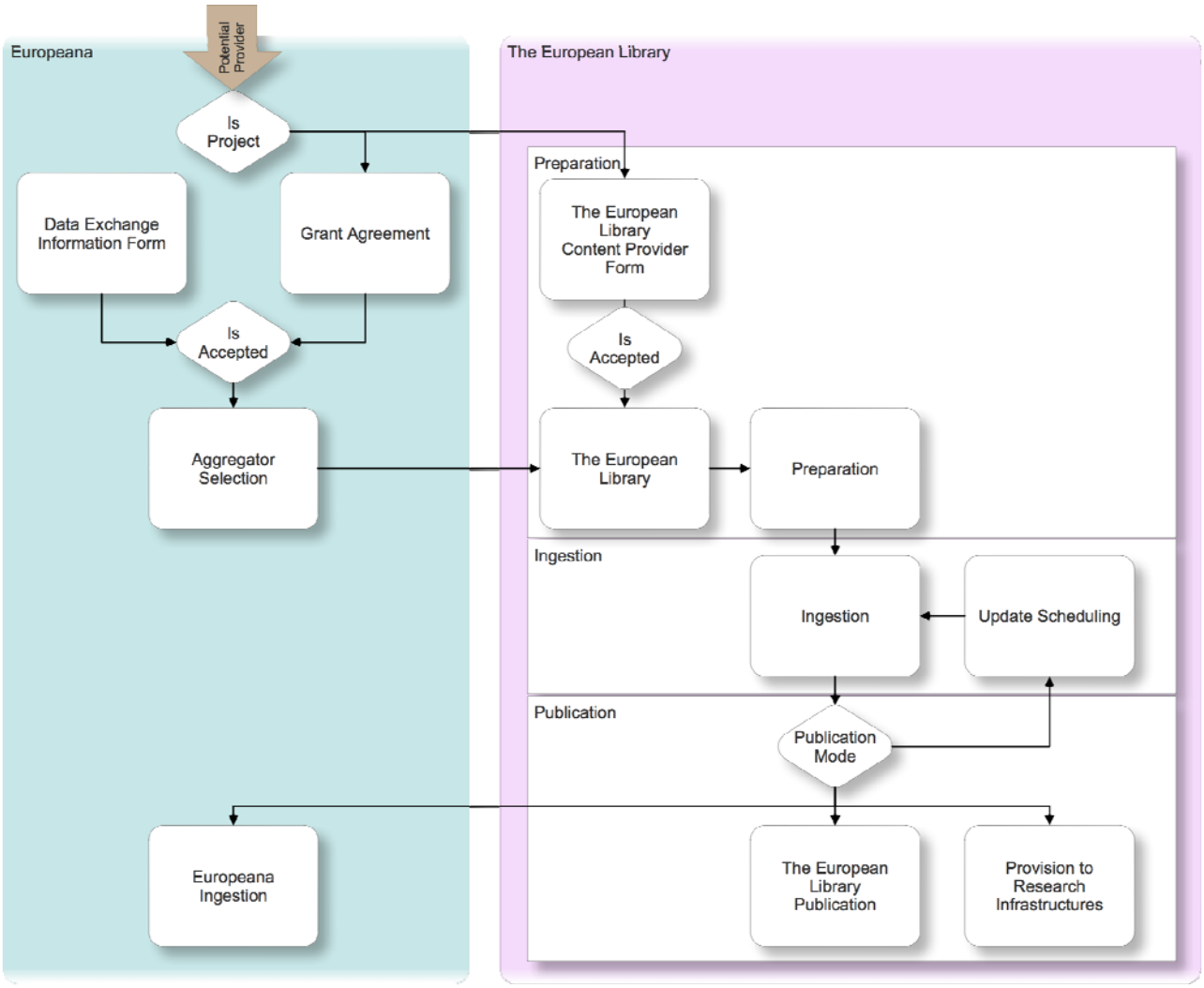



**Figure 6: Setup, Preparation, Ingestion, and Publication**

Content aggregation work is split into 3 main parts:

1. Preparation: all necessary information is collected and an ingestion plan is produced.
2. Ingestion: data is brought on site, normalized enriched and prepared for Europeana.
3. Publication: data is then made available to Europeana and the ingestion process at Europeana is triggered.

A more detailed view of these steps is shown in Figure 7. The steps are also described in detail in the following sections and a “visual” example is given in the Annex (Section 6.2).

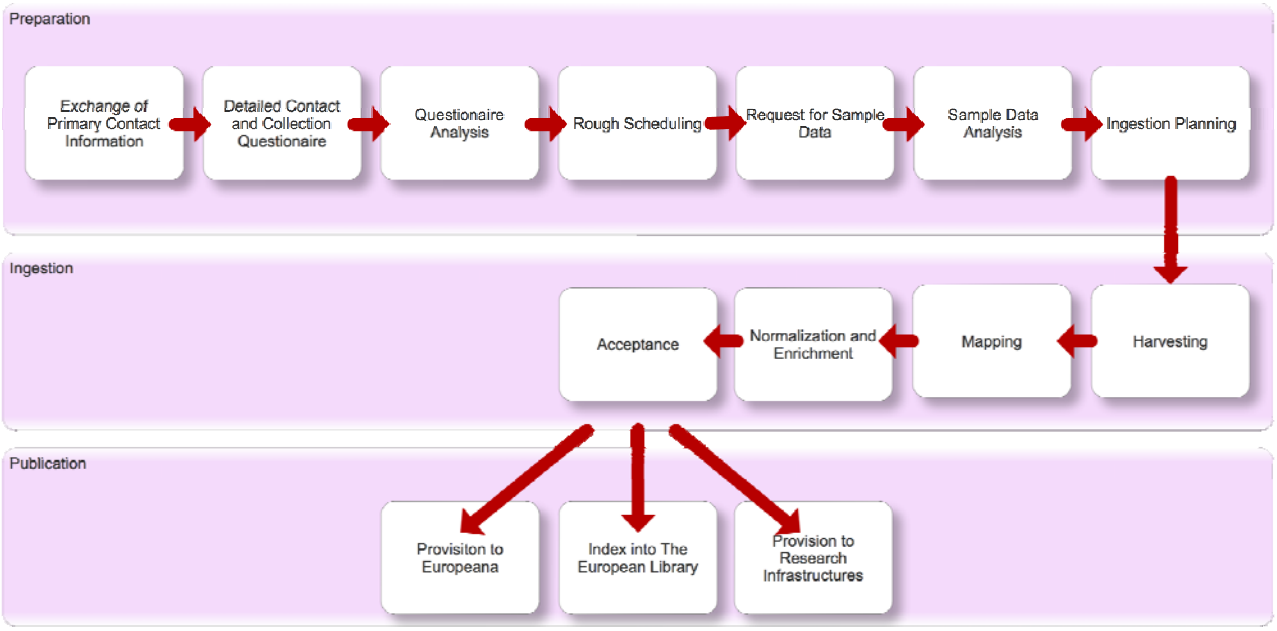



**Figure 7: Detailed Aggregation Steps**

## 3.1 Preparation

During the initial setup phase, the Collections Team will work with the content provider to identify possible ways of data provision, to agree on a metadata format and data transmission protocol and to set a schedule for the process. This preparation work can be split into 7 more detailed steps, which are outlined below.

### 3.1.1 Primary Contact, Questionnaire, Scheduling

After an aggregation agreement is signed between a content provider and The European Library or Europeana, the Collections Team will request primary contact details from the provider. Depending on the organisational structure, this might be a single contact or multiple contacts. Note that this is the primary contact and this person should be able to answer questions related to metadata and to data exchange, or should be able to delegate these tasks to appropriate people within the organisation.

Next, a questionnaire will be sent to the provider about the type and amount of data to be provided. It asks for details about metadata format, possible exchange protocols and other details concerning the content. It also includes technical, logistic and administrative questions such as:

- Preferred method of delivery and transport
- Digitisation status
- Quality of metadata
- Logistical questions such as the preferred time for harvesting, names of collections, number of records for collection, types of objects, and estimated update frequency
- Administrative information

Once the questionnaire is submitted, The European Library can provide a first schedule for further processing.

All information about providers, collections and communication between The European Library and content providers is logged in a customer-relationship management system. Please note that for the registry of collections on The European Library portal, we also collect information such as geographical coordinates, links to Wikipedia, other external resources, and various information related to collection categorisation information.

### 3.1.2 Sample Data, Analysis and Planning

Once a schedule has been set, providers will be asked to provide sample data via an agreed exchange protocol. This helps us assess the complexity, richness and quality of the metadata, and to plan the aggregation. Once this has been completed, both sides can agree on an ingestion plan. Each side should notify the other in good time when delays or unexpected issues are discovered.

The next step is to perform a sample harvest. This helps to further assess the readiness of the data exchange protocol. Based on these parameters, The European Library will incorporate the new collection into its content ingestion plan. The content provider will be informed and the current content ingestion plan can also be accessed online: http://www.theeuropeanlibrary.org/confluence/display/partner/Content+Ingestion+Plan
.

## 3.2 Ingestion

The European Library will contact the content provider about two weeks before the planned ingestion. This is to ensure that the content is ready and to clarify any last questions. The internal ingestion process is then split into three parts (see Figure 8), followed by acceptance and publication (see Section 3.3).

1. Harvesting – The transfer of metadata stored by providers to The European Library. The preferred way of data exchange (see Section 4.4.2) is metadata harvesting via OAI-PMH. The harvested data is then loaded into our raw data repository: a storage space for unmodified metadata, as provided by the content providers.
2. Mapping - The structural transformation of one metadata format into another. Potential library metadata format is mapped to an internal representation, which allows normalisation and enrichment.
3. Normalisation and enrichment -The alignment of provided metadata with common authority files and the enrichment of this data with links to external web resources, as well as between metadata records.

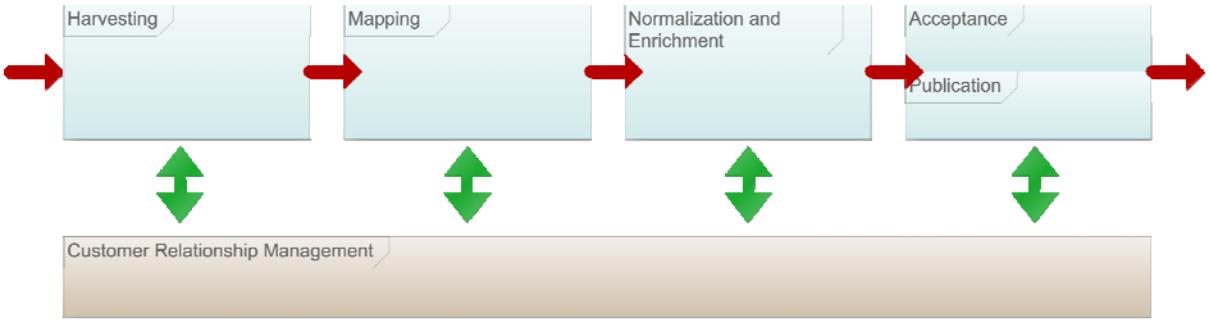



**Figure 8: Ingestion Steps**

As you can see in Figure 8. the data flows from left to right and all individual steps are documented and controlled in our customer management system.  An overview of the system and more technical details is outlined in Section 4.2.

## 3.3 Publication

In general, publication is a two-step process:

1. Acceptance phase - The metadata is available for partners only. Partners have the chance to validate and check the quality of the transformed and enriched metadata.
2. Publication phase - The metadata becomes available for a broader community and/or third party organisations, based on agreements with the data provider.

### 3.3.1 Publication in The European Library

Publication in The European Library has the following three targets:

1. Publication to end-users, the data becomes available in The European Library portal.
2. Publication to data services based on OAI-PMH for data exchange and search APIs.
3. Publication to research infrastructures (eg. Europeana) and as Linked Open Data (LOD). Both of these depend on agreements with the data provider.

The European Library works to both short term and quarterly timescales. The quarterly planning sets the rough planning structure for individual partners. Within any given quarter, data is selected and processed in batches that last 2 weeks. The amount of data processed depends on the availability of staff at The European Library and at the providing organisation.

### 3.3.2 Publication to Europeana

To ensure publication of data within Europeana The European Library will work with the data provider to format the data so it meets Europeana's specifications (ESE/EDM). The properly formatted data is then made available to Europeana in The European Library's OAI-PMH server and the Europeana ingestion team is informed about the newly available content.

The processing time of data in The European Library is between two and three weeks. After delivering the data to Europeana, it takes two to four weeks for the material to appear in the Europeana portal. As an aggregator, The European Library cannot guarantee availability in Europeana but in general we advise allowing a total of 8 weeks before the data is processed and made available in Europeana.

In general, Europeana publishes new data during the first week of each month. The publishing process is straightforward and no significant delays are to be expected.

## 3.4 Updates and Deletions

Any changes made by an institution in its repository, such as updates or corrections of datasets already live in The European Library and Europeana, should be communicated to the Collections Team. If the collection is frequently updated, an automatic update process can be agreed.

In cases where the data is made available to The European Library via OAI-PMH, the data is semi-automatically pushed through all necessary processing steps. Our incremental process only processes the updates. This significantly speeds up the process because only changed metadata needs to be reprocessed. It does, however, rely on persistent identifiers. If a collection does not have persistent identifiers, incremental processing will not be possible. The European Library will then have to delete and reprocess all records to ensure that only accurate data reaches the publication process. Europeana follows a similar update approach. This ensures that an updated collection always reflects the complete status of a collection at the time of harvesting.

## 3.5 *On-Demand Removal of Data*

In very specific cases, if required by a partner library, collections can be removed or taken offline from The European Library and Europeana services. The European Library can also delete single records. Up to two working days are required for records to be removed from all caches. Data aggregated to Europeana or other external services are beyond the direct control of The European Library. We cannot remove the records directly but we can liaise with these external services to proceed to the removal of collections and records.

## 3.6 Provision of the Europeana Data Model (EDM) for libraries

Europeana is currently moving from the Europeana Semantic Elements (ESE) to the Europeana Data Model (EDM). This change has implications for The European Library data provision to Europeana. The Europeana Library has been working within the project Europeana Libraries [5] on the Europeana Data Model from the library perspective. The project has specified how to model EDM for libraries materials (mainly monographs, multi-volumes works and serials) and has described how to use the EDM classes and their related properties[6].

EDM is a complex data model which differs from the other well-known library metadata formats in that it was designed from the start to be semantic web compliant. The delivery of EDM is a good way to identify gaps in library metadata and in the long term will allow the improvement of the library data quality in the Europeana portal.

The conversion to EDM is supported within the United Ingestion Manager (UIM) and will be continuously improved in order to follow the most recent EDM implementation done by Europeana. The European Library will in the future provide further guidance for providers on the data quality level which is expected to provide good EDM. The section in this Handbook on the information requirements will be updated as soon as Europeana will starts to actively receive data described according to EDM.

Further research work is carried out within The European Library Metadata Working Group to refine the EDM model for libraries.

[5] http://www.europeana-libraries.eu/
[6] D5.1 Report on the alignment of library metadata with EDM http://www.europeana-libraries.eu/documents/868553/1eade085-34ac-487f-82af-d5cd2545e619

# 4 Operational Requirements

This section discusses the operational requirements (technical, informational and administrative) for providers which have content being aggregated by The European Library. The overall aims of The European Library are to:

- Maximise the quality of metadata aggregated
- Keep information loss as low as possible
- Allow the highest degree of searchability
- Exchange data as efficiently as possible
- Minimise the costs for all involved organisations

## 4.1 Information Requirements

To support our operational goals in the future, as well as currently, we aim to collect as much information as possible about the objects being aggregated, even though this information may not yet be fully supported by the end-user services.

Besides these general aims, a number of requirements need to be fulfilled in the publication phase. These requirements vary, depending on the service for which the data has been aggregated..

### 4.1.1 Information Requirements for The European Library

The European Library's new research portal aims at higher-level scholars and researchers in the humanities and social sciences. After the approval of collections by the content strategy group, the data also needs to conform to the informational need of researchers. Special attention is brought to qualitative measures so that any research based on The European Library corpus can refer to the published content strategy and qualitative requirements.

With respect to metadata quality, The European Library automatically calculates a maturity level for each record. The maturity level is the weighted sum of metadata elements and their expected content length. Based on this maturity level, records are then classified into five groups: reject, weak reject, borderline, weak accept, and accept.

- Rejected records are not searchable in any of The European Library services, although the records are listed in statistics about collections and providers.
- Weak rejected and borderline records are only searchable via APIs and are available for data-mining services, but are not delivered services such as research infrastructures.
- Weak accepted and accept records are included in all services and data-mining services of The European Library and beyond. These records are included in content delivery to research infrastructures as well as in Linked Open Data services.

Content providers will be informed about the status of their collections. We aim to work with content providers to bring all records to at least weak acceptance level.

### 4.1.2 Information Requirements for Europeana

In order to fulfil the role of an aggregator, The European Library must adhere to the requirements for data completeness defined by Europeana. These are governed by the ESE and EDM specifications[7]. The European Library is responsible for making available to Europeana the following information packages:

- Metadata: descriptive and administrative metadata about the digital object.
- Persistent identifiers: for each single record provided.
- Links to resources: links to the digital object(s), the object in context and thumbnail.
- Rights information: about the metadata and the digital object.

Given that the Europeana service is based upon metadata only, the demand for the quality of metadata is higher than for other services. Europeana defines a number of mandatory fields in the ESE and EDM specifications. Please note that some of the fields are mandatory for textual objects only. Beyond that, it is highly recommended to provide language information with all textual metadata fields if the language differs from record to record.

The minimum requirement for the provided metadata is title and description, as well as contextual information (subject, type, coverage or spatial), and at least one link to a digital object. As mentioned above, textual objects also need language information. The other required fields such as data provider and rights information can be introduced at a collection level.

With regard to the rights information, the content provider needs to choose one of the seven Creative Commons licenses and, if none of them is applicable, one of the four Europeana rights statements. In order to choose the correct rights statements, please refer to the rights guidelines available at http://pro.europeana.eu/c/document_library/get_file?uuid=06e63d96-0358-4be8-9422-d63df3218510&groupId=10602

With regard to the preview, Europeana's main visual representation, it can be generated by Europeana using the provided link to the digital object. However, not all image formats are suitable for this process, and direct access to the image is mandatory to carry out this process. Note that an image embedded in a web page is not suitable to create a preview.

## 4.2 Technical and Metadata Requirements

The core requirements are best described along two dimensions: Firstly, "what" kind of data to exchange – the metadata format – and secondly "how" to exchange this data – the transport mechanism.

The European Library accepts a range of metadata formats via a set of transport mechanisms to accommodate the diverse and growing infrastructure in libraries' data provision. This should lower the costs at the provider side as much as possible. In order to tackle the quality of metadata at the same time, we prefer to exchange data in the richest

[7] Please find the requirements:
for ESE (Europeana Semantic Elements Specifications v3.4) at http://pro.europeana.eu/technical-requirements
for EDM (Europeana Data Model v5.2.2) at http://pro.europeana.eu/edm-documentation

export format possible. Table 1 lists the requirements matrix. The combinations which tick the most boxes are the preferred ones.

| | OAI-PMH | FTP (provider) | FTP (aggregator) | Z39.50 | Others (HTTP, Email) |
|---|---|---|---|---|---|
| MARC21 | XXXXX (5) | XXXX (4) | XXXX (4) | XXX (3) | XXX (3) |
| UNIMARC | XXXXX (5) | XXXX (4) | XXXX (4) | XXX (3) | XXX (3) |
| Other MARC | XXXX (4) | XXX (3) | XXX (3) | XXX (3) | XXX (3) |
| MODS | XXXX (4) | XXX (3) | XXX (3) | XXX (3) | XXX (3) |
| EDM | XXX (3) | XXX (3) | XXX (3) | XXX (3) | XXX (3) |
| DC | XX (2) | XX (2) | XX (2) | XX (2) | XX (2) |
| DCX/TEL/ESE/DMF | XX (2) | XX (2) | XX (2) | XX (2) | XX (2) |
| TEI | XX (2) | XX (2) | XX (2) | XX (2) | XX (2) |
| EAD | XX (2) | XX (2) | XX (2) | XX (2) | XX (2) |
| Others | X (1) | X (1) | X (1) | X (1) | X (1) |

**Table 1 Transport and Metadata Requirements Matrix**

The supported transport mechanisms are OAI-PMH[8], FTP[9], Z39.50[10] and HTTP[11]. In exceptional cases we also support all kinds of other means of data exchange. The difference between FTP (provider) and FTP (aggregator) is simply which organisation runs the FTP server. The European Library does have an FTP server to which the provider can upload batches of data, but if the data provider prefers to host the data, The European Library can also harvest data from an FTP server hosted by the provider. The exchange protocol Z39.50 is only supported if the implementation supports harvesting by identifiers For regular updates, the set of transport mechanisms is reduced to: OAI-PMH, FTP, and HTTP.

In addition to the listed formats, The European Library also supports container formats like METS, where the embedded descriptive metadata should comply with a format listed in the table. To support customisable metadata formats like TEI and EAD, as well as non-standard formats, the documentation needs to be available in English and additional metadata support needs to be available during the mapping phase.

## 4.3 Administrative and Resource Requirements

For efficient aggregation, all communication should be kept as effective and brief as possible. In practice, this means that the defined contact person on the provider side should be able (during the two week transfer period) to answer questions from The European Library promptly and to get in touch with relevant people, if necessary. Note that questions about rights and legal issues usually involve directors and therefore we strongly recommend clarifying the rights statement beforehand. When aggregating to Europeana, The European Library will approach the provider a few weeks before the scheduled

[8] http://www.openarchives.org/pmh/
[9] http://www.ietf.org/rfc/rfc959.txt
[10] http://www.loc.gov/z3950/agency/
[11] http://www.ietf.org/rfc/rfc2616.txt

ingestion. As mentioned above, The European Library works in batches of two weeks and, in the case of a delay, the continuation of work on a collection may need to be moved by several weeks.

Some data providers have the resources to support The European Library with mappings from their own metadata format to Europeana format (ESE/EDM). As described in Section 4.4, The European Library uses XML style-sheet transformations for metadata mapping and can therefore incorporate provided XSL definitions to a large degree.

### 4.3.1 Content Acceptance by Provider

As shown in Figure 8, after complete ingestion the data arrives in an acceptance phase. During this phase, the data provider needs to validate the data mapping, normalisation and enrichment. Depending on the data format and collection type, The European Library will apply different workflows during processing.

### 4.3.2 Automatic Link Validation

To ensure access to the digital objects or the catalogues at the provider side, The European Library will perform link validation on a sample of the records. Depending on the size of the collection, 500 to 10,000 records are uniformly drawn at random from the whole collection and submitted to the link-checking system. The link checker will then validate up to three links (thumbnail, digital object, and catalogue link) by issuing a HTTP HEAD request. If the web server on the provider side does not support HTTP HEAD, then a HTTP GET request is issued. Note that the validation system implements the robot's exclusion protocol and does not issue more than two requests per second. The HTTP-Client header value of the link checker is "The European Library: Validation".

Based on the outcome of the sample set validation, we can estimate the percentage of broken links in the whole collection with a confidence level of 99%. If the percentage of broken links is greater than 3%, we will need to validate all records in a collection. The list of broken links will be provided to the data provider and further steps will be agreed.

## 4.4 The European Library Aggregation Infrastructure

This section focuses on the technical details of the aggregation infrastructure and is presented as supportive information. The overall system architecture, including The European Library portal and some of the APIs, is shown in Figure 9.

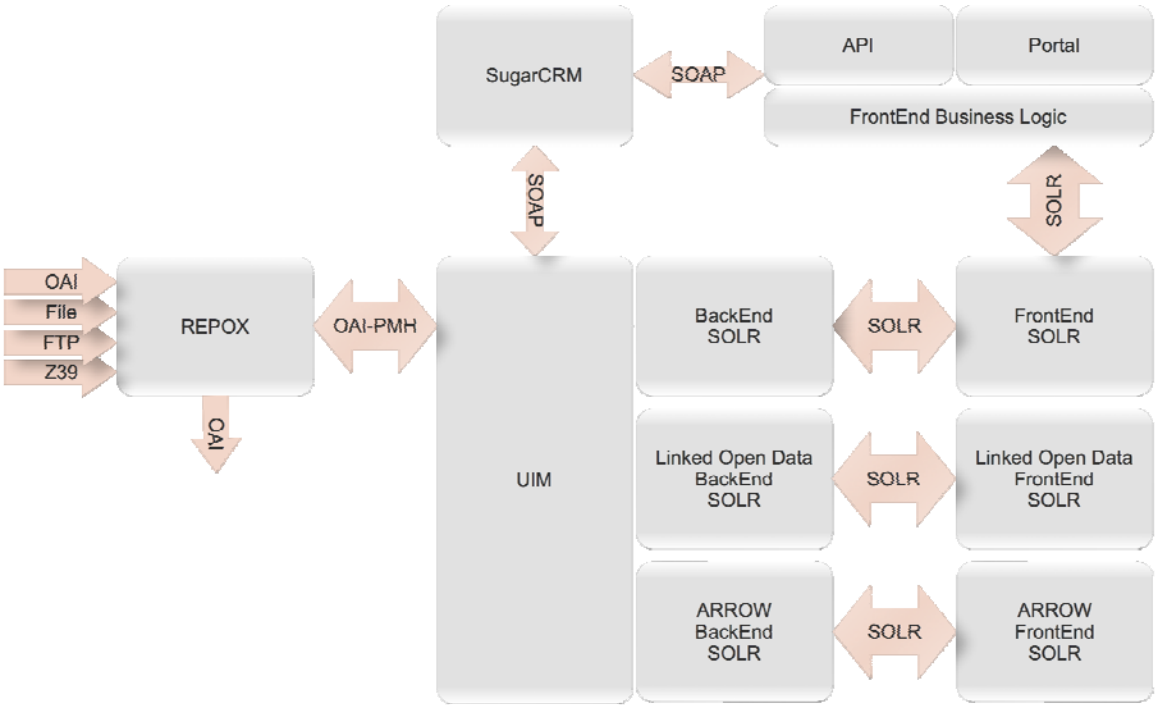



**Figure 9: System Architecture of The European Library**

### 4.4.1 Configuration Management (SugarCRM)

SugarCRM – our customer relationship system – provides a complete overview of all information about a provider, a collection and the operational status. It is where configuration information across all systems is stored, alongside the status of information being processed and any contact between The European Library and the provider.

### 4.4.2 Harvesting Infrastructure (REPOX)

The European Library uses the REPOX[12] Data Aggregation and Interoperability Manager for all kinds of harvesting tasks. The REPOX system implements an OAI-PMH client and supports incremental harvesting based on the OAI-PMH parameters 'from' and 'until'. The REPOX system further supports harvesting from local and remote files via the FTP protocol and also supports Z39.50 harvesting. As used in The European Library, the REPOX system behaves agnostically with respect to metadata format and only requires well-formed XML as input.

### 4.4.3 Unified Ingestion Manager (UIM)

The Unified Ingestion Manager is a joint development between Europeana and The European Library and serves as the central platform for all ingestion work. Besides moderating the distribution of configuration information between the systems (SugarCRM, REPOX, UIM, and APIs) it serves as the underlying framework for any enrichment and normalisation plug-ins.

[12] http://repox.ist.utl.pt/

The UIM is based on OSGI[13] specification and therefore implements a component and plug-in system. The European Library has implemented a number of normalisation and enrichment plug-ins. These plug-ins process the data in so-called 'workflows'.

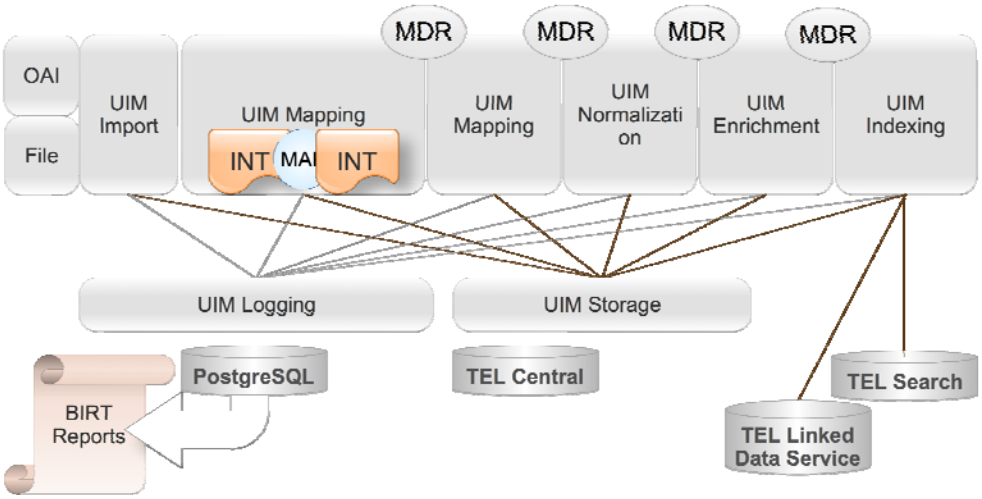



**Figure 10: Architecture Unified Ingestion Manager**

Based on a workflow specification (e.g. "Load and Enrichment") the UIM performs the tasks of loading the data from REPOX into the UIM Storage, mapping and parsing the metadata and then applying the normalisation and enrichment plug-ins.

### 4.4.4 Normalisation, Enrichment and De-Duplication

The European Library aligns all provided metadata with identifiers of the cross-library authority file and thereby normalises the provided information. For example, the name of each person and organisation is normalised and aligned with the appropriate record from the authority file, and the identifiers are added to the metadata record.

In all cases, such normalisations and enrichments are clearly marked as being added by the aggregator, to ensure the separation of provenance. This normalisation step demands a combined central authority file, which The European Library helps to build. The maintenance and automatic alignment of authority files is a highly complex task and the subject of current research and scientific discussions, which goes beyond the context of this document.

In addition to the normalisation steps with authority files, The European Library enriches the metadata with GeoNames[14] and the Ortelius Thesaurus[15]. This provides additional contextual information for places of publications and the disciplinary affiliation of objects aligned with research information systems and EuroCRIS[16].

Due to the nature of the library domain, where each book usually exists in several copies or editions, it is necessary to de-duplicate the provided metadata. For this, The European

[13] http://www.osgi.org/
[14] http://www.geonames.org/
[15] http://cordis.europa.eu/cerif/src/future.htm
[16] http://www.eurocris.org/

Library has developed a metadata-clustering algorithm, which groups records into manifest groups (in the sense of FRBR[17]).

# 5 Governance, legal and financial issues

*Governance*

The governance of The European Library is based on a formal agreement between CENL (Conference of European National Librarians), LIBER (Association of European Research Libraries) and CERL (Consortium of European Research Libraries). The main governing body, The European Library Management Committee, has 10 members - 6 from CENL, 2 from LIBER and 2 from CERL. The Chair and Treasurer are CENL representatives, the Vice-Chair is from LIBER. It is the job of the Management Committee to oversee the development of The European Library.

*Legal*

The legal basis for every library's membership of The European Library is the Partnership Agreement between CENL, the owner of The European Library, and each library or consortium of libraries. A new Partnership Agreement runs from 1st January 2013.

The Partnership Agreement uses The European Library Handbook as a major point of reference. The Handbook will therefore always be kept up to date and its purpose is to guide libraries in their provision of metadata and content to The European Library, as well as providing general information about our services.

The European Library has a policy to distribute member libraries' metadata as widely as possible and it has approval from the CENL Board to release a bibliographic dataset under a CC0 licence. The Partnership Agreement makes it clear that it is up to libraries which metadata they wish to include under a CC0 licence and which metadata they wish to remain restricted.

The Partnership Agreement is fully aligned with the Europeana Data Exchange Agreement[18], which has been signed by CENL. In signing the Partnership Agreement, libraries give The European Library the necessary approval for us to forward metadata to digital content to Europeana, all within the bounds of the Europeana Licensing Framework.

For Europeana and The European Library, being metadata services, it is crucial to clarify the rights situation around the metadata, because this significantly impacts the operational side of our work. It is worth noting that the granularity for rights on metadata is always agreed on a collection and provider level.

For end users, it is also crucial to know what one can do with the digital object, once it has been downloaded from the provider's website. An end user might only be interested in objects which can be reused and republished. The European Library and Europeana also therefore maintain the rights and privileges about the digital objects. Given that the situation

[17] http://archive.ifla.org/VII/s13/frbr/
[18] http://pro.europeana.eu/data-exchange-agreement

may vary between objects in a single collection, this information is maintained on a record level within the metadata. Europeana also provides filter functionality for public domain material as well as other rights statements.

*Membership fees*

Membership of The European Library is open to individual libraries and consortia.

There are eleven fee bands and each member library will be placed in one of these. The library's exact fee is calculated on the basis of six criteria; five of these relate to national economic indicators and one is based on the size of the library's resources. Each criterion has a weighting, adding up to a total of 100%. The six are as follows:

- Gross national income per capita $ (average for the period) (15% weighting)
- Purchasing power parity $, a figure similar to GNP but factoring in the cost of living (50% weighting)
- Gross national income $ (average for the period) (10% weighting)
- Gross domestic product $ (10% weighting)
- Population of the country (5% weighting)
- Staff (full-time equivalents) of the library (10% weighting)

Significant discounts are available to consortia, based on the number of libraries, and whether there is a single source from which we can harvest the data, thereby reducing the amount of processing required.

**Document history**

| Revision | Date | Author | Organisation | Description |
|---|---|---|---|---|
| 1.0 | 21.12.11 | Valentine Charles, Sally Chambers, Chiara Latronico, Andreas Juffinger | The European Library | First version |
| 1.0 | 15.12.11 | Shirley Agudo, Friedel Grant | The European Library | Proofreading |
| 1.1 | 04.12 | Valentine Charles | The European Library | New edition |
| 1.2 | 12.12 | Valentine Charles, Louise Edwards | The European Library | Addition of a section on EDM and update of the lagal and financial section |

Acknowledgements
Thanks are due to the following for their expert contribution to this document:
Julia Brungs, Sally Chambers, Valentine Charles, Robina Clayphan, Louise Edwards, Friedel Grant, Andreas Juffinger and Chiara Latronico from The European Library office.

# 6 Annex

## 6.1 ANNEX I: Glossary

This annex sums up definitions used across this document and clarifies the use of specific terms. The following terms are thereby agreed vocabulary used by The European Library.

[Def.1] **Catalogue**: A concept related to the Metadata of one or more Collections of a library, independent of the digital or non-digital nature of the related Contents. This is a vague concept, as sometimes it might be used to mean a set of Metadata (thus, an information entity) but it also might be used to mean the system that manages the creation, editing and storage of that Metadata (in those cases it is more correctly named, in the library domain, as the "Cataloguing System", but it is also common to see that system simply named "Catalogue" for the same purpose).

[Def.2] **Collection:** An intentionally-defined set of **Content,** compiled under a specific policy. This is a common concept in the library domain, so it is used here with the same meaning as in that domain.

[Def.3] **Content**: The digital objects that can be accessed through Metadata. Content is typically held on Data Providers'/Aggregators' sites. Content is usually defined by its individuality and cultural, intellectual or artistic expression. Content has a reference to an individual object of the real world or is born digital. Examples: photographs, books, letters, films, paintings, television, etc. Note: In online delivery, Content excludes the peripheral packaging/platform.

[Def.4] **Contextual Resources**: Catch-all term for resources which help to provide context for the **Content** and make it possible to enrich the services to be developed by the **Service Providers** (such as Europeana). Data such as linked data, ontologies, vocabularies, thesauri, classifications, taxonomies, etc.

[Def.5] **Data**: Catch-all term including Metadata, Thumbnails, Audio and Moving image previews. **In the scope of this document, this concept also includes, by default, Full-text Data** [Def.15].

[Def.6] **Data Aggregator**: Organisation that collects, formats and manages Data from Data Providers before making that available to Service Providers (such as Europeana).

[Def.7] **Data Collection:** The **Data** corresponding to a specific **Collection**.

[Def.8] **Data Export Task**: This is a task of harvesting a Data collection from the TEL Aggregator by a Service Provider.

[Def.9] **Data Ingest Task**: A task of harvesting a Data Collection from a Data Provider.

[Def.10] **Data Provider**: Organisation that makes **Data** available to a Data Aggregator (such as the TEL Aggregator) or a Service Provider (such as Europeana).

[Def.11] **Data Provider Record**: A Data Provider Record is a generic concept to name all the structured information the TEL Aggregator maintains about a Data Provider. That concept comprises all the descriptive and contact information, as well as the information about all the Data that the Data Provider is willing to provide for Data Harvest Tasks.

[Def.12] **Data Schema**: A description of the structure of a specific Data.

[Def.13] **Enriched Data**: Data that has been subject to a process of Enrichment, Normalisation or Transformation.

[Def.14] **Enrichment**: A process that generates **Enriched Data** from **Raw Data**. It can consist of adding machine-generated new attributes to **Records** (such as linking to authority files, geographic data, etc., making use of **Contextual Resources**); in this case the values assigned to the attributes can consist of data (such as a textual string or a temporal date) or a URI to an external entity. In the particular case of this project, **this also comprises the building of search indexes from the full-text.** Other kinds of processes of **Enrichment** are **Transformation** and **Normalisation.**

[Def.15] **Full-text Data**: **Data** in the form of text representing literal transcriptions of written or spoken words from the **Content**. This is a new class of **Data** to be considered, related to the and thus not covered (and so not to be confused) by the concepts of **Contextual Resources** ([Def.4]) or **Metadata**.

[Def.16] **Mapping:** An expression of rules to convert **Data** structured according to a source **Data Schema** into new **Data** structured according to a target **Data Schema**.

[Def.17] **Metadata**: Metadata is information about Content, describing its characteristics to aid in its identification, discovery, interpretation and management. Metadata is given to Europeana and drives discovery of Content held at the Data Provider's/Aggregator's site. Metadata are usually facts or fact-like information, containing little individual artistic/creative expression. Examples: Bibliographic or filmographic data, temporary and spatial data, etc.

[Def.18] **Normalisation**: A kind of **Enrichment** in order to make the **Data** conformant with its declared **Data Schema**. This might comprise, for example, adding missing mandatory attributes or the normalisation of values (e.g. the normalisation of date values to ISO 8601 compliant strings).

[Def.19] **Preview** : A reduced size or length audio and/or a lower resolution visual representation of Content, in the form of one or more images, text files, audio files and/or moving image files.

[Def.20] **TEL Aggregator**: The **Data** aggregator system realised by the European Library Aggregation Infrastructure under the responsibility of The European Library (and sometimes also mentioned in the Definition of Work, DoW, as the European Library Aggregation Infrastructure).

[Def.21] **Transformation**: A kind of **Enrichment** by applying a set of **Mapping** rules to **Raw Data** in order to produce new **Enriched Data** structured according to a target **Data Schema**. It is important to stress that a **Transformation** only uses the Raw Data "as it is", which might imply the need of **Normalisation** to ensure that the **Enriched Data** is fully conformant with the target **Data Schema.**

[Def.22] **UIM: Unified Ingestion Management tool**, also called **Ingestion Control Panel**, represents the extensible framework to manage the whole ingestion process.

[Def.23] **URI**: Uniform Resource Identifier, URLs (Uniform Resource Locators) are URIs.

[Def.24] Raw Data: The Data the TEL Aggregator collects from the Data Providers.

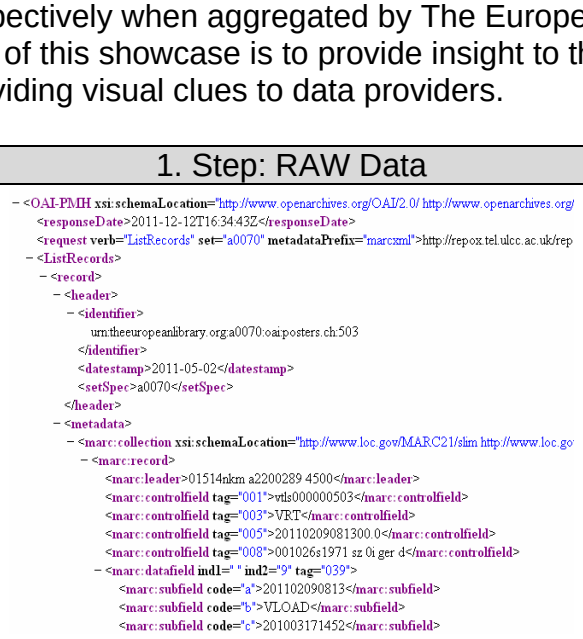
[Def.25] **Record**: The unit of **Metadata** concerning a single **Content** object. Europeana has strict definitions for records and objects. A record is, for Europeana, the unit of metadata concerning a single content object. On the other hand, an object with no

metadata information cannot be considered a record and cannot be accepted in Europeana.

## 6.2 ANNEX II: Show Case

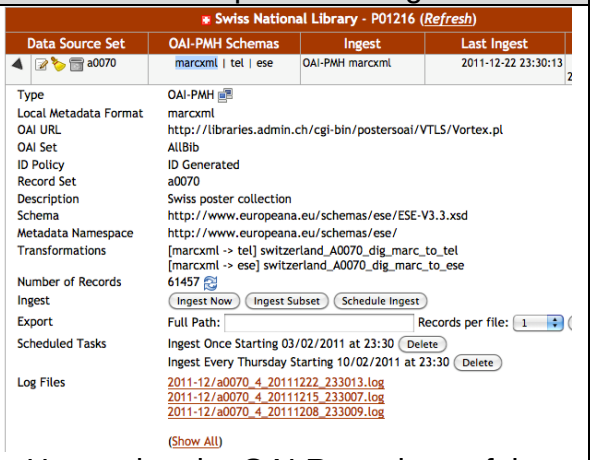
The following series of images reflects the tools, and states of a collections, or single record respectively when aggregated by The European Library and provided to Europeana. The aim of this showcase is to provide insight to the complexity of the process as well as providing visual clues to data providers.

### 1. Step: RAW Data




Collection and records provided in MARC 21 via an OAI-PMH repository. (Section 4.2)

### 2. Step: Harvesting

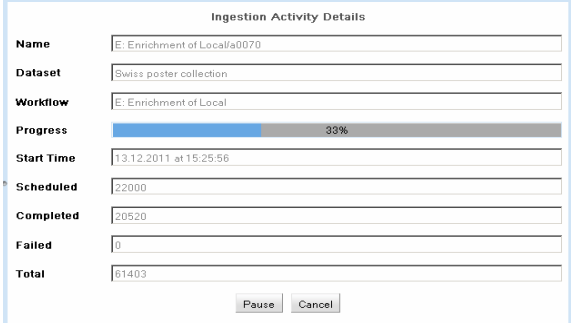



Harvesting the OAI Repository of the data provider with our REPOX System to bring the RAW data on site.

### 3. Step: Ingestion

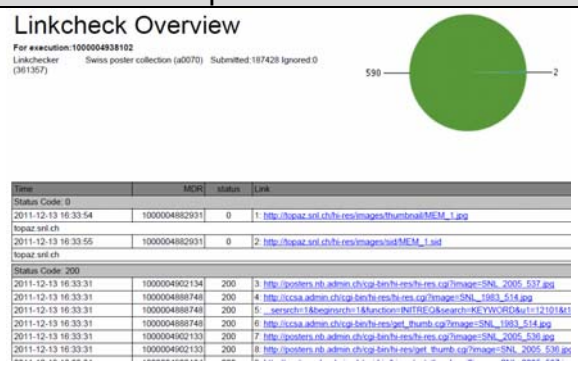
The data is then loaded into the enrichment, and ingestion tool UIM.

### 4. Step: Normalization & Enrichment



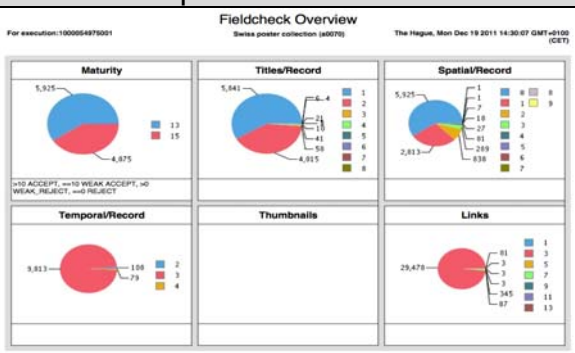

A enrichment and normalization workflows are then applied on a collection level

### 5. Step: Link Validation

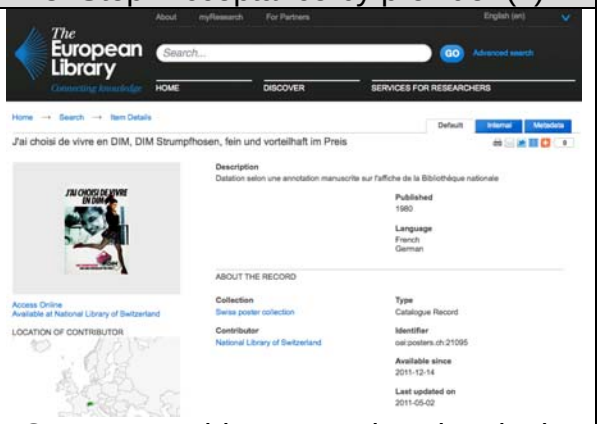
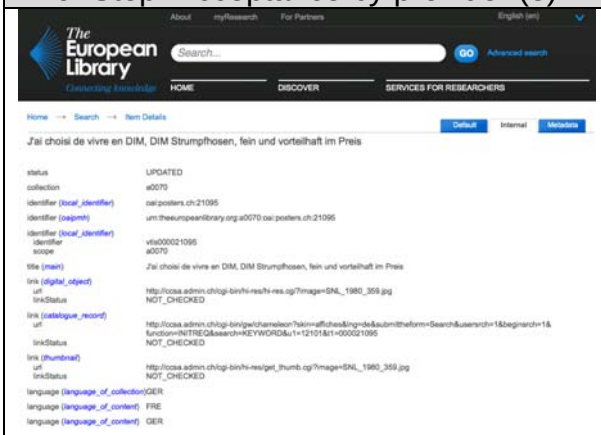
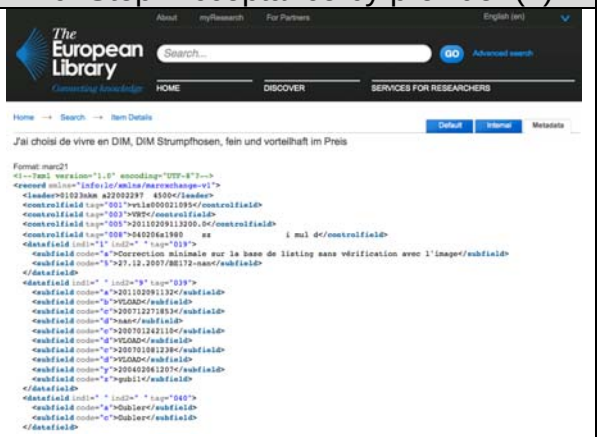
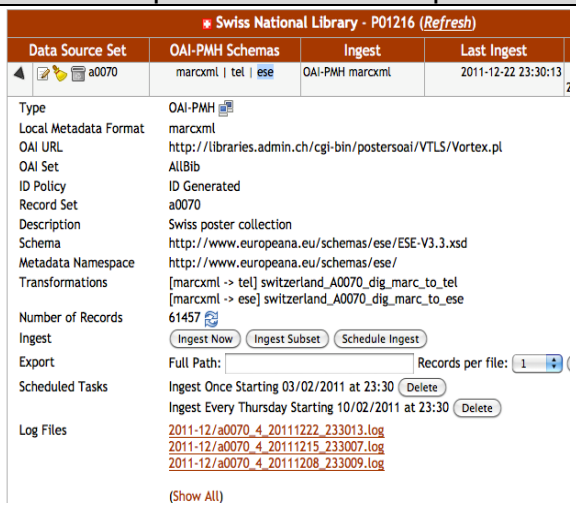
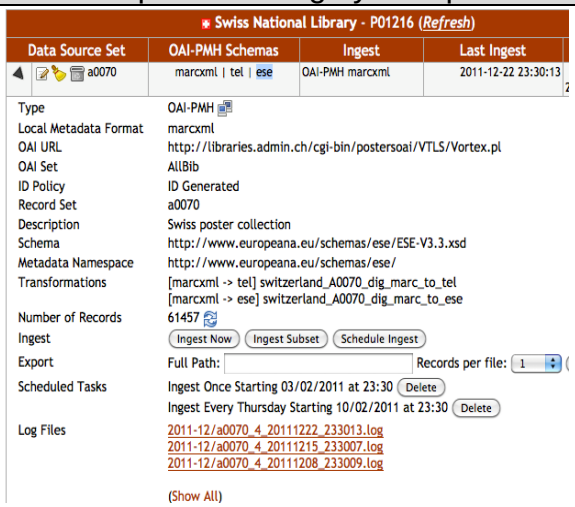


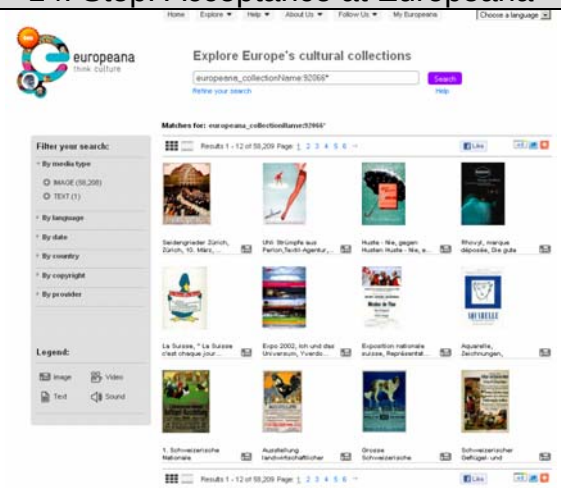
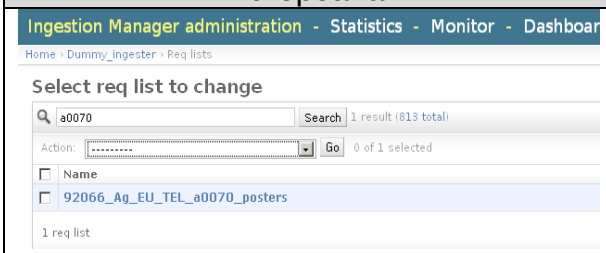
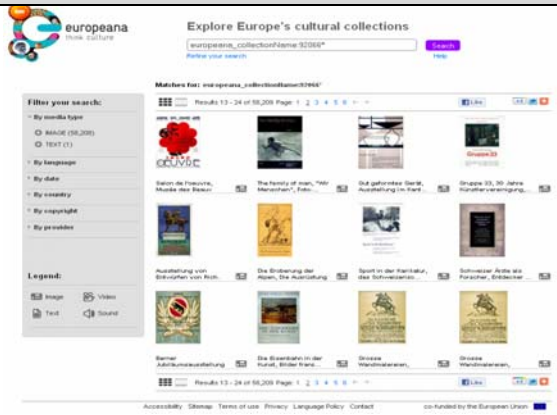
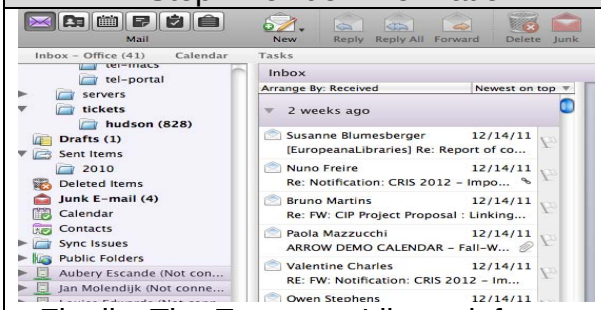

A link check report is then created to validate the provided links.

### 6. Step: Content Validation




Furthermore a content report is created to assess the completeness and maturity of the data (full collection).

| 7. Step: Acceptance by provider (1) | 8. Step: Acceptance by provider (2) |
|---|---|
| The European Library provides then a link to the preview the complete collection. |  Content provider can review the single records from an end user perspective. |
| **9. Step: Acceptance by provider (3)** | **10. Step: Acceptance by provider (4)** |
|  Content provider can review the single records in the internal view, representing all enrichments and normalization steps. |  Content provider can also access the provided record in the original metadata format. |
| **11. Step: Provision to Europeana** | **12. Step: Harvesting by Europeana** |
|  The approved data is then provided to Europeana via our OAI Repository in one of the Europeana formats (EDM or ESE) |  Europeana then harvests the ESE data from The European Library, via the REPOX system. |

| 13. Step: Enrichment at Europeana | 14. Step. Acceptance at Europeana |
|---|---|
| <br>Europeana further uses the UIM to load and enrich the data with cross domain thesauri etc. |  <br>Europeana provides the data then on their acceptance portal to The European Library for final validation. The Europeana Ingestion team does also final verifications during this acceptance phase. |
| **15. Step: Thumbnail Caching at Europeana** | **16. Step: Publication at Europeana** |
|  <br>Europeana does then a complete link check and starts to cache thumbnails. |  <br>In parallel the collections are published on a monthly basis to the Europeana portal. |
| **17. Step: Provider Information** | |
|  <br>Finally, The European Library informs the provider about the publication of the data in Europeana | |

## 6.3 ANNEX III: ESE Element Overview

For more details about the Europeana Semantic Elements (ESE), please consult the guidelines at
http://pro.europeana.eu/technical-requirements

| Mandatory Elements | Recommended/Preferred Elements |
|---|---|
| 1. dc:title<br>2. dc:description<br>3. dc:language<br>4. europeana:dataProvider<br>5. europeana:isShownAt<br>6. europeana:isShownBy<br>7. europeana:provider<br>8. dc:subject or dc:type or dc:coverage or dcterms:spatial<br>9. europeana:rights<br>10. europeana:type | 1. dcterms:alternative (this is a refinement of dc:title)<br>2. dc:creator<br>3. dc:contribuitor<br>4. dc:date (recommended ISO 8601: YYYY-MM-DD)<br>5. dcterms:created (refinement of dc:date)<br>6. dcterms:issued (refinement of dc:date)<br>7. dcterms:temporal (refinement of dc:coverage)<br>8. dc:publisher<br>9. dc:source<br>10. dcterms:isPartOf (refinement of dc:relation)<br>11. europeana:object<br>12. dc:identifier |
| **Additional/Optional Elements** | **Elements supplied by Europeana** |
| 1. dc:format<br>2. dcterms:extent (refinement of dc:format)<br>3. dcterms:medium (refinement of dc:format)<br>4. dc:rights (about intellectual property rights, access rights for the digital object)<br>5. dcterms:provenance<br>6. dc:relation<br>7. dcterms:conformsTo (refinement of dc:relation)<br>8. dcterms:hasFormat (refinement of dc:relation)<br>9. dcterms:isFormatOf (refinement of dc:relation)<br>10. dcterms:hasVersion (refinement of dc:relation)<br>11. dcterms:isVersionOf (refinement of dc:relation)<br>12. dcterms:hasPart (refinement of dc:relation)<br>13. dcterms:isReferencedBy (refinement of dc:relation)<br>14. dcterms:references (refinement of dc:relation)<br>15. dcterms:isReplacedBy (refinement of dc:relation)<br>16. dcterms:replaces (refinement of dc:relation)<br>17. dcterms:isRequiredBy (refinement of dc:relation)<br>18. dcterms:requires (refinement of dc:relation)<br>19. dcterms:tableOfContents (refinement of dc:relation)<br>20. europeana:UGC<br>21. europeana:unstored | 1. europeana:country<br>2. europeana:language<br>3. europeana:uri<br>4. europeana:usertag<br>5. europeana:year |