DELIVERABLE

Project Acronym: Europeana Libraries
Grant Agreement number: 270933
Project Title: Europeana Libraries: Aggregating digital content from Europe's libraries

D6.5 The European Library Standards Handbook
Version 1.0

Authors:
Julia Brungs   The European Library
Sally Chambers   The European Library
Valentine Charles   The European Library / Europeana
Robina Clayphan   The European Library / Europeana
Louise Edwards   The European Library
Nuno Freire   The European Library
Friedel Grant   The European Library
Andreas Juffinger   The European Library
Chiara Latronico   The European Library
Magali Mermet   The European Library

Revision History

Revision	Date	Author	Organisation	Description
0.1	24.11.11	Valentine Charles, Chiara Latronico, Sally Chambers	The European Library	First draft outline of D6.5
0.2	02.12.11	Chiara Latronico	The European Library	Section 4-5: Draft
0.3	05.12.11	Andreas Juffinger	The European Library	Section 1-5: Draft
0.4	14.12.11	Chiara Latronico	The European Library	Section 4: Draft
0.5	15.12.11	Chiara Latronico	The European Library	Section 5: Draft
0.6	15.12.11	Friedel Grant	The European Library	Proofreading
0.6	19.12.11			Version sent for review
0.6	20.12.11	Shirley Agudo	The European Library	Proofreading
1.0	21.12.11			Final version
1.0	28.12.11	Valentine Charles	The European Library	Inclusion comments of appraisal group. Final version
1.0	29.12.11	Andreas Juffinger and Sally Chambers	The European Library	Adaptations to graphics and presentation. Executive Summary. Additions to the Legal and Finance section

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

TABLE OF CONTENTS

1 Overview

The European Library is a service of national and research libraries. It also acts as the library domain aggregator for *Europeana* and, in this role, supports data providers from the library domain across Europe in data-provision tasks. The *Europeana Library* gathers bibliographic records and metadata to digital objects from the 47 national libraries of Europe. With the *Europeana Libraries* project, a robust aggregation infrastructure is being developed, allowing the aggregation of hundreds of libraries and millions of records to digital and non-digital objects. The project is using this infrastructure to aggregate about 5 million objects from 20 partners for *Europeana*, as a pilot. The organisations LIBER and CERL are key partners in this project.

This document acts as a handbook for libraries providing data to *The European Library* and other services such as *Europeana*. Since the range of services available in the library domain is quite broad, this document is taking *Europeana* as the main example of redistribution of data from *The European Library service but there are others*.. Given the massive task of aggregation, this document clarifies the advantages of outsourcing specific data-provision tasks to an aggregator.

Providing a new collection to *The European Library* and then to *Europeana* via *The European Library* always involves a three-phase process: Firstly, 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. Secondly, during the ingestion phase, the data is actually processed and enriched. This phase ends with the acceptance and quality insurance by data providers whereby the transformed and enriched data is validated and approved. Thirdly, in the publication phase the data is then made available to *Europeana* and other services, depending on the agreements with the data provider.

As library domain aggregator, *The European Library* accepts various library formats being provided through different transport mechanisms and makes this data conform to *Europeana* and other research service specifications. The matrix in Table 1 shows the preference and accepted input formats, going from MARC21 via OAI-PMH (most preferred), over TEL/ESE via OAI-PMH and EDM, to proprietary XML formats via FTP or HTTP (least preferred).

Finally this document provides an overview of the legal and financial issues to consider when providing content to *The European Library* and *Europeana*. The Europeana Data Exchange Agreement (DEA) covers the rights of the metadata. It is worth noting that *The European Library* Partnership Agreement is currently being aligned with the DEA, and partners do not need to sign the DEA. The rights on the digital objects are untouched and not covered by these agreements. The digital object rights need to be provided on a record or collection level within the metadata.

For support and possible questions about the aggregation process, the main point of contact is *The European Library*'s Collections Team. It can be reached by e-mail:
collections@theeuropeanlibrary.org

2 Introduction

The *Europeana Libraries* project is working to build a robust aggregation model based on *The European Library*, which will make metadata to digital content from research and national libraries across Europe available on both *Europeana*[1] and the new *European Library*[2] portal. This handbook explains to data providers how this aggregation infrastructure can be used.

By using one aggregation system to reach complementary portals and services, content providers will benefit from the widest possible exposure of their materials. *The Europeana Library* portal will show off their objects to an audience of dedicated researchers, particularly in the social sciences and humanities. *Europeana* will reach the general public, including lifelong learners, students, special interest groups and expert amateurs. Providers do have the further possibility of choosing a variety of services, in addition to *Europeana* and *The European Library*, to which their material should be provided.

Application Programming Interfaces (APIs) from *The European Library* and *Europeana* will ensure that content can be discovered not only on these two primary portals but also via a variety of other platforms, already embedded in the workflow and normal internet surfing patterns of our audience.

Already, this new and innovative project is attracting significant interest from the library community. This handbook aims to provide sufficient information to libraries which wish to submit data to *The European Library* and others services such as *Europeana,* and also to become sustainable partners of the service.

The handbook first addresses the role of content providers in the project and the benefits of participating. It then examines the process of submitting data and the technical, organisational and financial requirements for the submission of data. Finally, the handbook outlines high-level legal aspects and lists a number of definitions and resources, which are useful for the overall development and sustainability of content providers.

3 A Library domain aggregator

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

3.1 Overview

The European Library has set up strategic partnerships with library and research organisations. Thanks to the aggregation infrastructure assembled via the *Europeana Libraries* project, data can be sent to these organisations as well as to *Europeana*.

---

[1] http://www.europeana.eu

[2] The website of The European Library – http://theeuropeanlibrary.org – will be relaunched in early 2012 with a new set of functions and a new design. This improved website is referred to as the new *European Library* portal.

Some of the organisations with which there is a relationship include *OCLC*, *Serial Solutions*, and *Mendeley,* together with research infrastructures such as *DARIAH*, *CLARIN* and *EHRI. There is a strong partnership in place with LIBER and CERL..*

This long history of experience with data aggregation means that *The European Library* can use its accumulated knowledge to process partners' data efficiently and effectively. This is important because formatting data for researchers is a complex task. It needs to

This long history of experience with data aggregation means that *The European Library* can use its accumulated knowledge to process partners' data efficiently and effectively. This is important because formatting data for researchers is a complex task. It needs to interoperate with other domains (e.g. into archives and museums), and should meet the standards for various proprietary formats. *The European Library* can provide that service, from library standard formats to the mandatory metadata formats on the service side.

As an aggregator, *The European Library* can also accumulate knowledge for dozens of institutions to decrease costs significantly for all partners, thanks to the ability to apply required metadata transformations to the data of multiple institutions at once, with only minor adaptations.
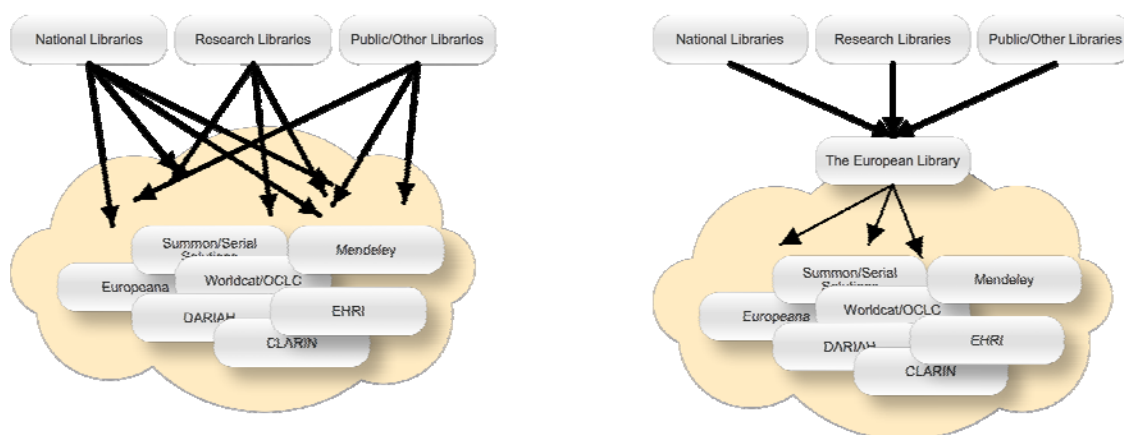

Figure 1: Information Exchange without and with Aggregator Service

3.2 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 .

Specialised and national aggregators aggregate metadata and content from data providers and institutions. This information is then provided to *Europeana* and to other projects, research infrastructures and thematic portals.
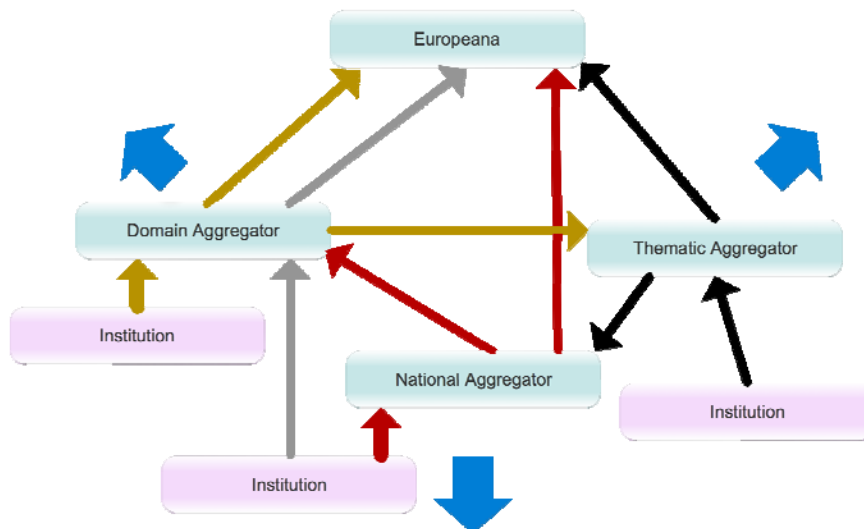
Figure 2: Data contribution flow diagram

Through this network of aggregators, metadata can be exchanged between institutions. This happens only with the permission of the providing institution. Such a network of aggregation does have the advantage that a provider can exclusively provide data to one aggregator, which then distributes the data to other aggregators on behalf of the provider. At the same time the data between aggregators can be exchanged in bigger batches, which generally reduces the aggregation cost per record.

## 3.3 The European Library: the library aggregator for Europeana

An aggregator to *Europeana* is a service organisation, which helps individual content providers to submit cultural heritage data to *Europeana*. The Aggregators Handbook edition 2[3] clarifies the role and core responsibilities of an aggregator:

> 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, standardise 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 as shown in **Error! Reference source not found.**. This creates a model that can be expanded across Europe and beyond.

---

[3] The Europeana Aggregator Handbook edition 2 is available at http://pro.europeana.eu/web/europeana-pro/provide_content
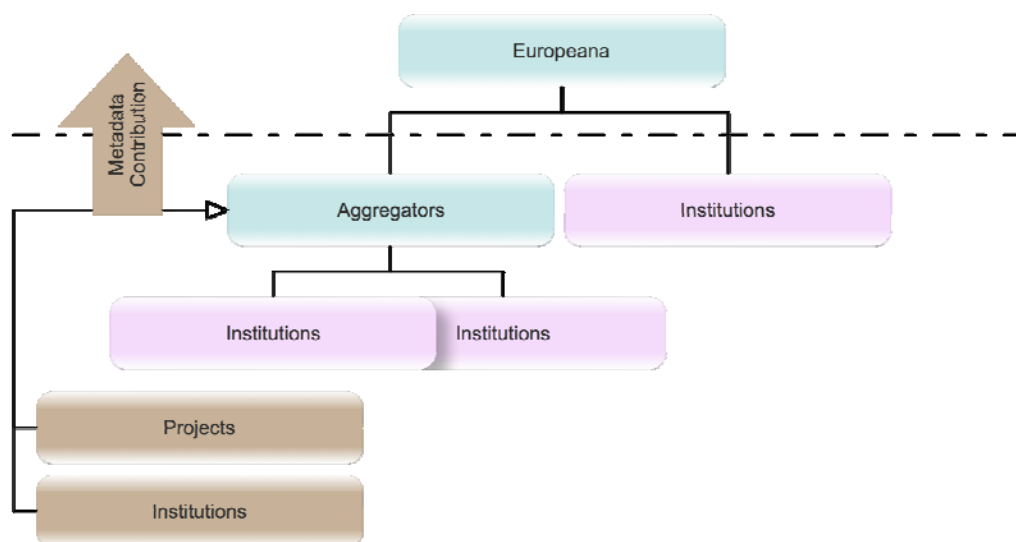
Figure 3: Europeana Data Submission Pyramid

As shown in the diagram, institutions provide their data to an aggregator (project or organisation), which then processes the data and provides it to *Europeana.* For reasons of limited representation of aggregators in countries, domains or themes, some institutions provide data directly to *Europeana*.

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 (local, national or European), or a domain level (cross-, single, thematic).

Regional aggregators: Regional aggregators are institutions aggregating content from a specific region only. National aggregators aggregate from their national cultural heritage institutions. National initiative aggregators are aggregator(s) single or cross domain, which have been appointed by their Ministry to take on the aggregation role in the country. National aggregators and national initiatives are first point of reference for Europeana in accessing content from a particular country.

Project aggregators: Project aggregators are organisations which have joined a project consortium with a specific aim and purpose. The project aggregators can be either aiming at aggregating within a specific theme or at single or cross-domain level.

Independent organisations: Organisational aggregators are independent organisations which have taken an aggregation role representing geographic and domain levels; thematic, single or cross-domain.

The European Library is the library domain aggregator for *Europeana* and therefore has three main responsibilities: firstly, to gather material from libraries; secondly, to standardise the library domain formats and metadata formats to *Europeana*'s preferred standard (**E**uropeana **S**emantic **E**lements or **ESE** and the **E**uropeana **D**ata **M**odel or **EDM**); and thirdly, 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.
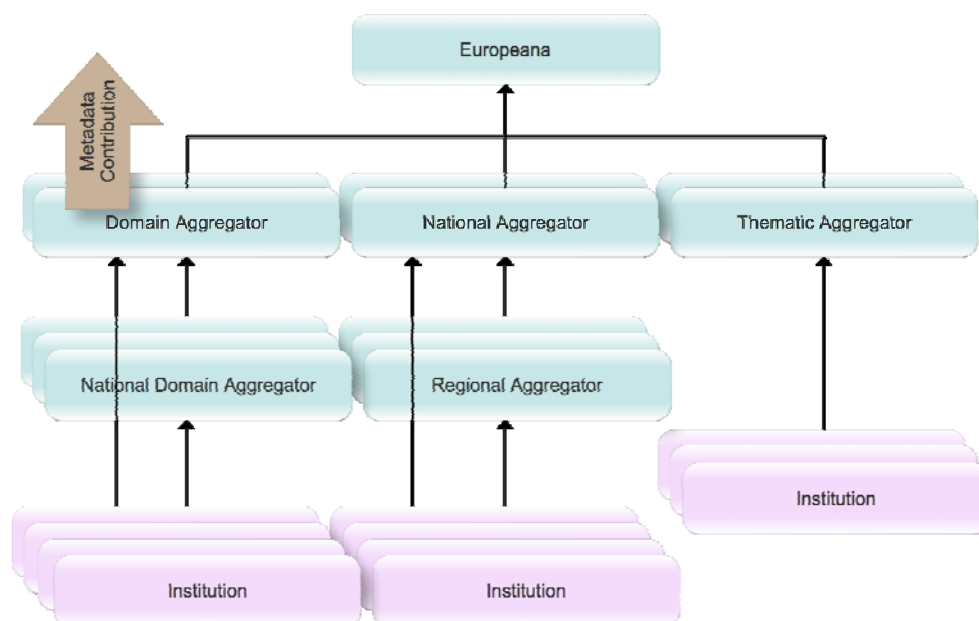
Figure 4: A Hierarchy of Aggregators

The massive task of aggregating content from thousands of institutions across Europe has necessitated the creation of an eco-system with multiple levels. Regional domain aggregators aggregate content to national domain aggregators. Higher-level aggregators then bring this content together before passing it on to *Europeana*. In some countries, institutions contribute directly to a national aggregator, as there are no regional aggregators.

3.4 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[4]. To facilitate this, we have the ARROW system. It is a tool that has been developed by a consortium of partners[5] with the aim of helping libraries and other cultural heritage institutions to clarify the rights status of textual works that they wish to include 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. Through the ARROWplus project[6] the coverage of the ARROW system will be extended to a further 12 countries.

---

[4] http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1292
[5] Further information about the ARROW and ARROWplus projects is available at: http://www.arrow-net.eu/
[6] Further information about the ARROWplus project is available at: http://www.arrow-net.eu/what-arrow-plus
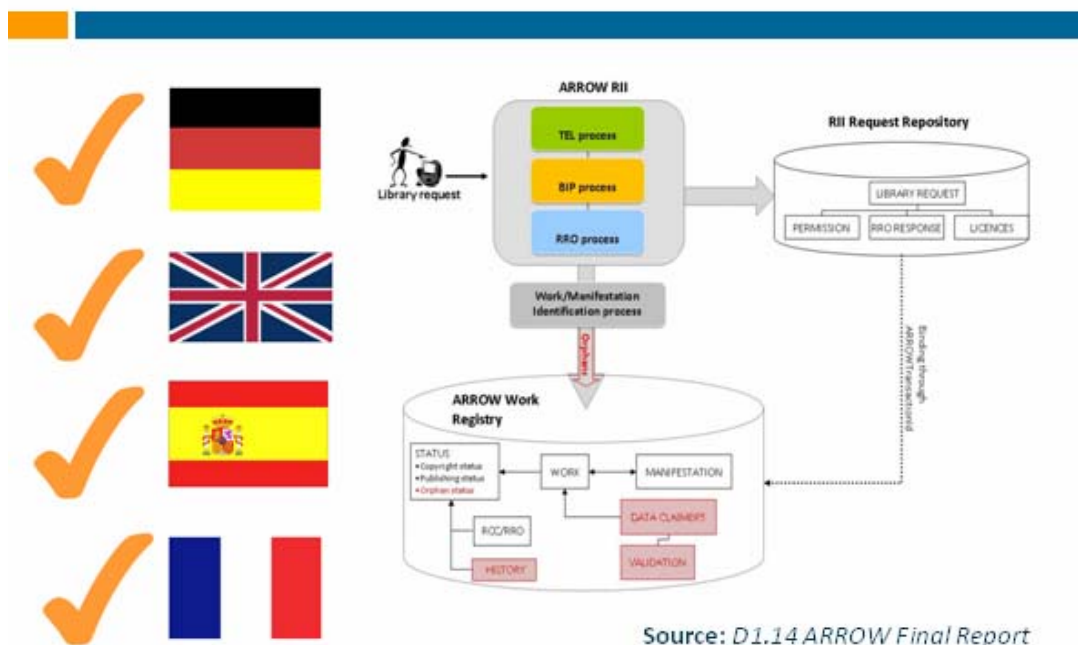
Figure 1 : ARROW System

As a key partner in the project, *The European Library* intends to integrate the ARROW system into the library-domain aggregation infrastructure. This means that partners of *The European Library* should be able to use the ARROW system as part of their membership of *The European Library*. Further information about using the ARROW system, including access to the test version of ARROW, is available from *The European Library* team.

4 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 other services. An overview and process diagram is shown in **Error! Reference source not found.**. A potential content provider may be a partner in a content delivery project to Europeana, in which case it is not mandatory to complete the Europeana Data Exchange Information form. If an organisation is not part of such a project, then it will fill in the Data Exchange Information form provided by Europeana. Europeana will then validate the information and inform the organisation about the decision made, based on the content strategy and rerouting alternatives. As soon as a potential content is assigned to an aggregator, then it will carry out the necessary aggregation work and act as a single point of communication.

The European Library, as the service for researchers, has a content strategy that reflects the information needs of academic communities. It will also carry out an assessment of the content for its relevance to the target audience.
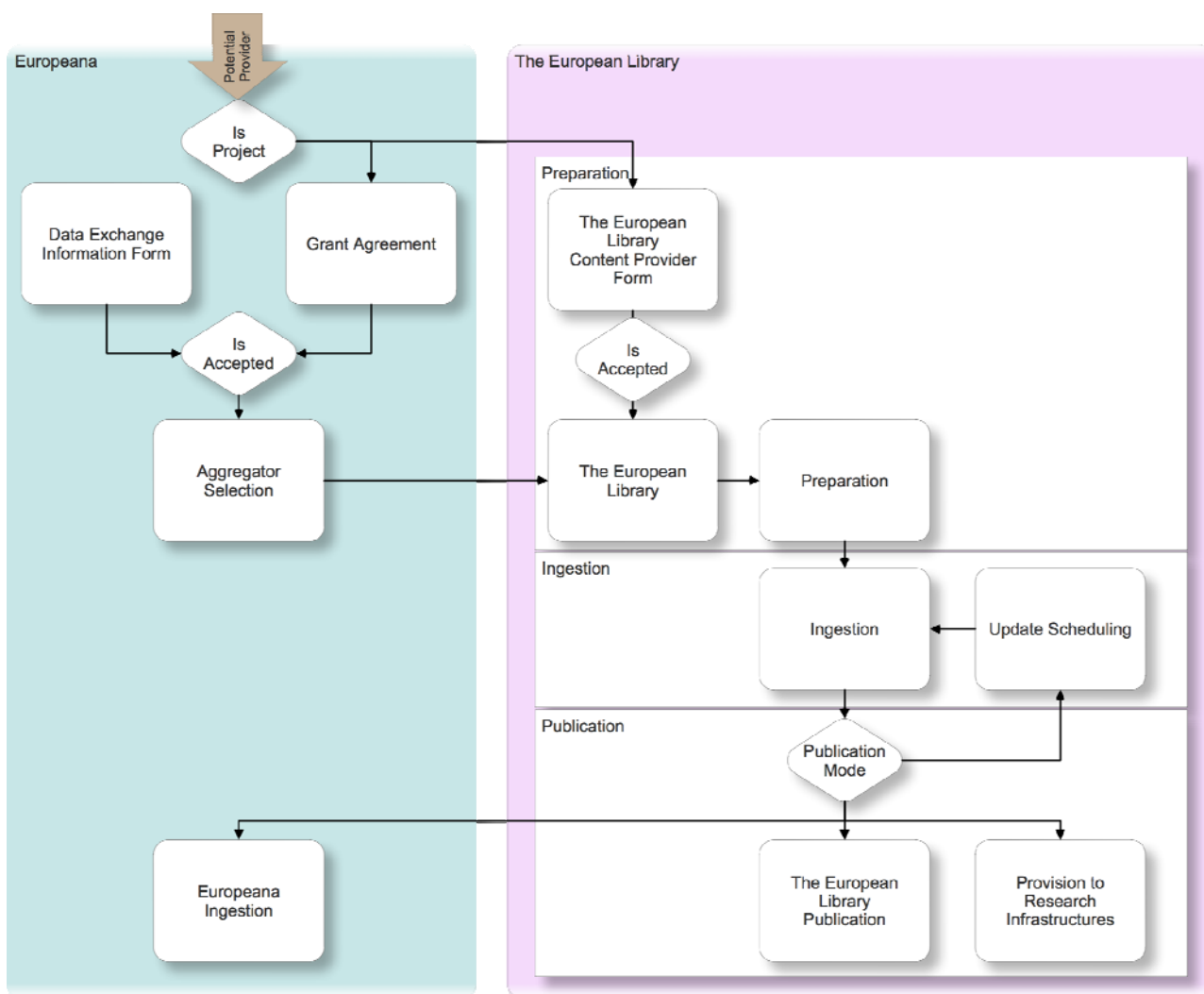
Figure 2: Setup, Preparation, Ingestion, and Publication

The aggregator then carries out aggregation and supports the content provider. The actual content aggregation work in *The European Library* is split into three main parts:

1. Preparation: all necessary information is collected and an ingestion plan is produced.
2. Ingestion: data is brought on site, normalised, enriched and prepared for others services such as Europeana.
3. Publication: data is then made available in *The European Library* portal and to other services.

A more detailed view of these steps is shown in Figure 3. The involved steps are described in more detail in the following sections; a "visual" example is given in the Annex (Section 7.2).
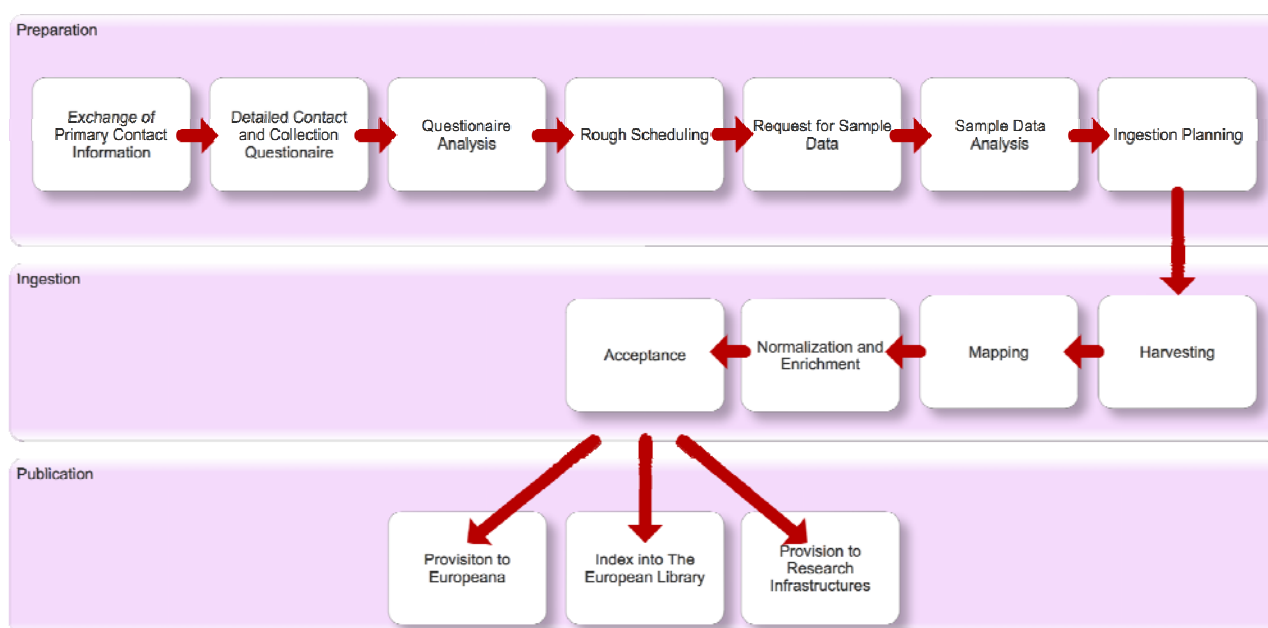
Figure 3: Detailed Aggregation Steps

For support and possible questions related to this workflow, the main point of contact is *The European Library*'s Collections Team. It can be reached by e-mail: collections@theeuropeanlibrary.org

4.1 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.

4.1.1 Primary Contact, Questionnaire, Scheduling

After an aggregation agreement is signed between a content provider and *The European Library*, 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 metadata-related and technical data exchange questions, or be able to delegate these tasks to appropriate people within the provider organisation.

Next, a questionnaire will be sent to the provider in order to clarify the type and amount of data to be provided. This questionnaire 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

Important aspects, such as status of collections, metadata, and access to digital content, must be clarified during this phase. As soon as the questionnaire is submitted, and open questions are answered, *The European Library* will provide a first schedule for further processing.

All information about providers, collections and communication between *The European Library* and content providers is maintained in an open-source 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 all kinds of collection categorisation information.

4.1.2 Sample Data, Analysis and Planning

According to the agreed aggregation plan, providers are asked to provide sample data via an agreed exchange protocol. This helps us to assess the complexity, richness and quality of the metadata. *The European Library* will then perform an in-depth analysis on the sample data, which will help us 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.

As soon as the most efficient way of data exchange is agreed, *The European Library* will 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 and inform the content provider about the ingestion schedule. The current content ingestion plan will be published under http://www.theeuropeanlibrary.org/confluence .

4.2 Ingestion

According to the content ingestion plan, *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 4), followed by acceptance and publication (see Section 4.3).

1. Harvesting - the process of bringing the metadata (stored at the provider side) on site. The preferred way of data exchange (see Section 5.4.2) is metadata harvesting via OAI-PMH. As soon as the data is on site, the data is loaded into our raw data repository, a repository of metadata with unmodified data as provided by the content providers.

2. Mapping - the process of structural transformation of one metadata format into another. During this step, the potential library metadata format is mapped to an internal representation, which allows normalisation and enrichment.

3. Normalisation and enrichment - the provided information is aligned with common authority files and enriched with links to external web resources, as well as between metadata records.
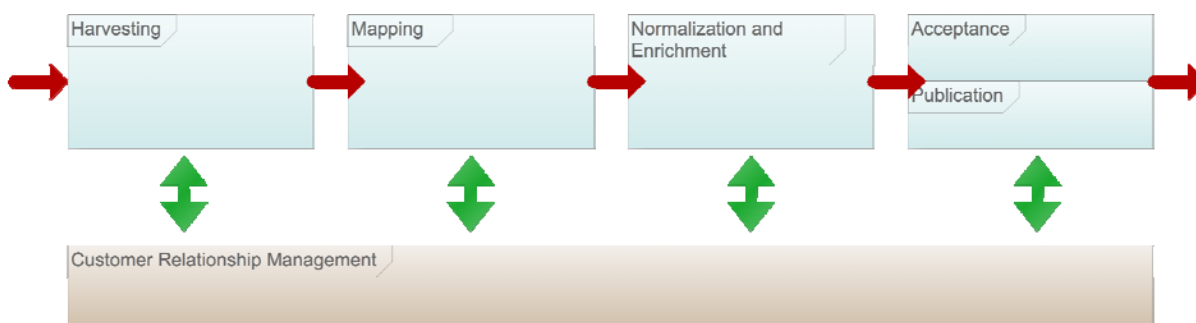
Figure 4: Ingestion Steps

The data flows from left to right, and all individual steps are documented and controlled in our customer management system. For an overview of systems and more technical details, see Section 5.2

4.3 Publication

Publication is the process of delivering the information (depending on the agreements with the data provider) to other external services such as Europeana, as well as making it available to *The European Library*'s user community. Furthermore, the data becomes available via APIs and Linked Open Data (LOD) services.

Generally, publication is a two-phase 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 the agreements with the data provider.

4.3.1 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 and as Linked Open Data (LOD).

The European Library works on a basis of short-term planning and longer-term quarterly planning. The quarterly planning sets the rough planning structure for individual partners. Within any given quarter, data is selected and processed in batches that last two weeks. The amount of data processed depends on the availability of staff at *The European Library* and at the provider.

4.3.2 Publication to other services: the example of Europeana

Publication to Europeana means the provision of metadata, which meets Europeana's guidelines to Europeana. As an aggregator, *The European Library* ensures, together with the data provider, that the data conforms to the specifications of Europeana (ESE/EDM). As soon as the data has been made available to Europeana in *The European Library* OAI-PMH server, the Europeana ingestion team is informed about the newly available content.

Currently, Europeana follows a monthly batch publishing process. All added and updated content is published during the first week of each month. Given that *The European Library* has vast experience in providing content to *Europeana*, the publishing process is straightforward and generally no significant delays are to be expected.

The processing time of data in *The European Library* is between two and three weeks. After delivering the data to Europeana, it takes between two and 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 eight weeks before the data is processed by *The European Library* and made available in Europeana.

4.4 Updates and Deletions

If an institution has made changes in its repository, such as updates or corrections of datasets already live in *The European Library* and *Europeana*, this 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 process is fully incremental and the data is semi-automatically pushed through all necessary processing steps. Note that an incremental process allows the processing of updates only. That significantly speeds up the updates because only changed metadata needs to be reprocessed. The matching of records is thereby based on identifiers, which works only if persistent identifiers are in place. 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.

4.5 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. In any case it can take up to two working days 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* and therefore no clear indicator can be made here of the removal procedure for collections and records.

5 Operational Requirements

This section discusses the operational requirements, which imply technical, metadata, information and administrative requirements for providers aggregated by *The European Library*. The aim of *The European Library* is thereby to maximise the quality of metadata aggregated, to keep the information loss as low as possible, to exchange the data as efficiently as possible and to minimise the costs across all involved organisations (data provider, *The European Library*, *Europeana…*).

5.1 Information Requirements

Besides the general aim of maximising the quality and richness of aggregated metadata, a number of requirements need to be fulfilled in the publication phase, depending on the service for which the data has been aggregated. Metadata should be provided to *The European Library* and to other services, always with the intention of allowing the highest degree of searchability. Given that it is

impossible to foresee all kinds of information needs, it is generally recommended to provide as much information as possible, even though it might be unused or not fully supported by the end-user services yet.

5.1.1 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 approval of collections by the content strategy group, the data also needs to conform to the information 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 to research infrastructures and the like.
- 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 linked open-data services.

Content providers will be informed about the status of their collections and, together with the operational team of *The European Library*, it should be the aim to bring all records to at least weak acceptance level.

5.1.2 Information requirements for Europeana

In order to fulfil the role of an aggregator, *The European Library* has to adhere to the requirements for the data completeness defined by *Europeana*. In this way the data requirements 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 preview.
- 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

---

[7] Please find the requirements for ESE (Europeana Semantic Elements Specifications v3.4) and for EDM (Europeana Data Model v5.2.2):
http://pro.europeana.eu/documents/51031/487157/Europeana_Semantic_Elements_Specifications_v3.4.doc and
http://pro.europeana.eu/documents/866205/13001/EDM_v5.2.2.pdf

only. Beyond that it is highly recommended to provide language information with all textual metadata fields if it 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 a preview, *Europeana's* main visual representation, *Europeana* can generate one 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.

5.2 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 at the same time the quality of metadata, it is preferred to exchange data between data provider and aggregator in the richest export format of the deployed library management system. Table 1 lists the requirements matrix, where the combinations, which ticked the most boxes, are the most 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. Z39.50 as an exchange protocol is only supported if the implementation supports harvesting by ID. 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.

5.3 Administrative and Resource Requirements

It is crucial for efficient aggregation to keep all kinds of feedback loops as short as possible. This effectively means that the defined contact person on the provider side should have enough spare resources during the agreed period of two weeks to answer questions promptly or, if necessary, to get in touch with relevant people to answer requests from *The European Library*. Note that questions about rights and legal issues usually involve director level, and therefore we heavily recommend clarifying the rights statement beforehand. In case of aggregation to *Europeana*, *The European Library* will approach the provider weeks before the scheduled ingestion. As mentioned above, *The European Library* works in batches of two weeks, and it is very likely that in the case of a delay, the continuation of work on a collection may need to be moved by several weeks.

Some of the data providers do have the resources to support *The European Library* with mappings from their own metadata format to E*uropeana* format (ESE/EDM). As described in Section 5.4, *The European Library* uses XML style-sheet transformations for metadata mapping and can therefore incorporate provided XSL definitions to a large degree.

5.3.1 Content Acceptance by Provider

As shown in Figure 4, after complete ingestion the data arrives in an acceptance phase. During this phase the data provider is requested to validate the data mapping, normalisation and enrichments. Depending on the data format and collection type, *The European Library* will apply different workflows during processing.

5.3.2 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 (preview, digital object,

---

[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

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, it is then possible to assess the percentage of broken links in the whole collection with a confidence level of 99 percent. This means checking about 3,500 records in a collection totalling 25k records, or 4,000 records in a collection of 250k, allowing the assessment of the broken links with a very high confidence. If the percentage of broken links is greater than 3 percent, then it becomes necessary to validate all records in a collection. The list of broken links is provided to the data provider and further steps are agreed.

5.4 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 5.
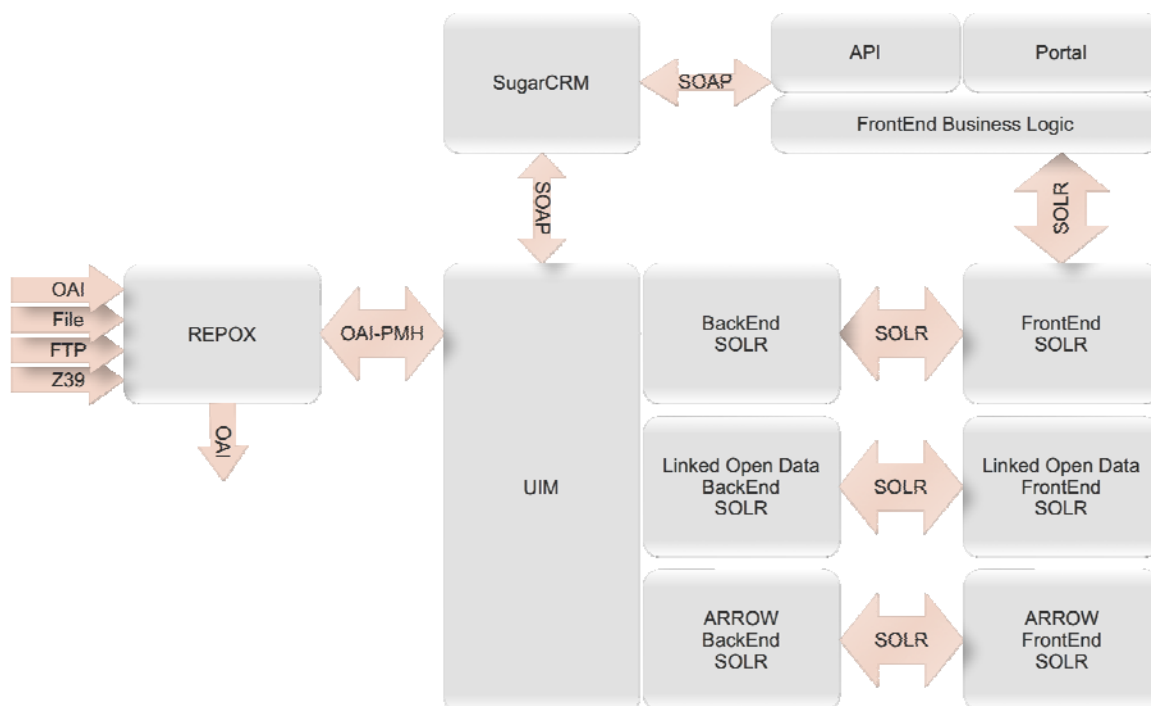
Figure 5: System Architecture of The European Library

5.4.1 Configuration Management (SugarCRM)

SugarCRM is used as the central store for configuration information across all systems and therefore becomes the single point of reference for any administrative information. Furthermore, all status information and processing of relevant information is collected in the CRM system, so that the system provides a complete overview of all information about a provider, a collection and the operational status.

5.4.2 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.

5.4.3 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.

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 which process the data in so-called workflows.
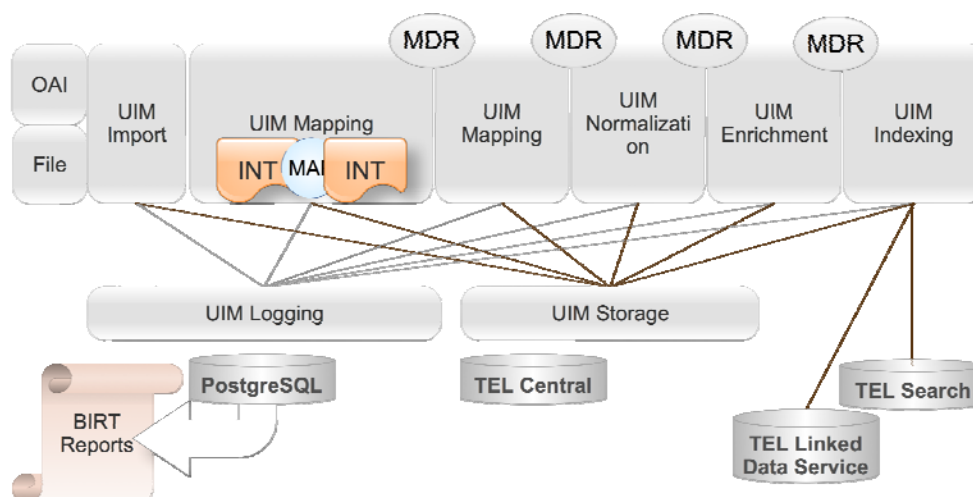
Figure 6: Architecture Unified Ingestion Manager

Based on a workflow specification, e.g. "Load and Enrichment", the UIM then 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.

5.4.4 Normalisation, Enrichment and De-Duplication

The European Library aligns provided metadata with identifiers of the cross-library authority file and thereby normalises the provided information. 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. Note that, in all cases, such normalisations and enrichments are clearly marked as being added by the aggregator, to ensure the separation of provenance. The aforementioned normalisation step demands a combined central authority file, which *The European Library* helps to

---

[12] http://repox.ist.utl.pt/
[13] http://www.osgi.org/

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

6 LEGAL AND FINANCIAL ISSUES

On behalf of *The European Library*, the Conference of European National Librarians (CENL) has signed a Data Aggregator Agreement with Europeana. CENL is therefore responsible for having a legal agreement with the data providers it aggregates. For the national libraries, this is *The European Library* Partnership Agreement.

In September 2011, Europeana Foundation officially adopted the Europeana Data Exchange Agreement (DEA)[18]. The DEA, which is part of the Europeana Licensing Framework, will replace the previous Data Provider and Data Aggregator Agreements. *The European Library* Partnership Agreement is being aligned with Europeana's Data Exchange Agreement.

For content providers, whose metadata is aggregated by *The European Library* and provided to Europeana as part of a project, but are not covered by *The European Library* Partnership Agreement, for example, research libraries in the Europeana Libraries project, it is not necessary for the organisations to sign the Data Exchange Agreement directly with Europeana, for the duration of the project. The project's Grant Agreement fulfils the role of the legal agreement in these instances.

With respect to aggregation, it is important to highlight the difference between the rights on the metadata from the rights on the digital object itself. For Europeana, being a metadata service only, it is crucial to clarify the rights situation around the metadata, because this does have significant impact on the operational aspects when processing metadata. It is worth noting that the granularity for rights on metadata is always agreed on the collection and provider level. For end-users, it is 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

---

[14] http://www.geonames.org/

[15] http://cordis.europa.eu/cerif/src/future.htm

[16] http://www.eurocris.org/

[17] http://archive.ifla.org/VII/s13/frbr/

[18] http://version1.europeana.eu/web/europeana-project/newagreement/

and republished. Therefore, Europeana is also maintaining the rights and privileges about the digital objects. Given that different situations might apply for different objects in a single collection, this information is maintained on a record level within the metadata, and Europeana also provides filter functionality for public domain material as well as other rights statements.

As part of the Europeana Libraries project, the Business Planning Group, made up of representatives of CENL, CERL, LIBER and Europeana, are reviewing and revising the current business, financing and governance model of *The European Library*. It is anticipated that these revised models will be implemented from January 2013, following the completion of the Europeana Libraries project.

7 Annex

7.1 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-upon 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, Images, 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 **Error! Reference source not found.**, 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 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 for **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.

7.2 ANNEX II: Show Case

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

1. Step: RAW Data	2. Step: Harvesting
Collection and records provided in MARC 21 via an OAI-PMH repository. (Section 5.2)	Harvesting the OAI Repository of the data provider with our REPOX System to bring the RAW data on site.
3. Step: Ingestion	4. Step: Normalisation & Enrichment
The data is then loaded into the enrichment, and ingestion tool UIM.	Enrichment and normalisation workflows are then applied on a collection level.
5. Step: Link Validation	6. Step: Content Validation
A link check report is then created to validate the provided links.	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 then provides a link to the preview of 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 normalisation steps.	Content provider can also access the provided record in the original metadata format.

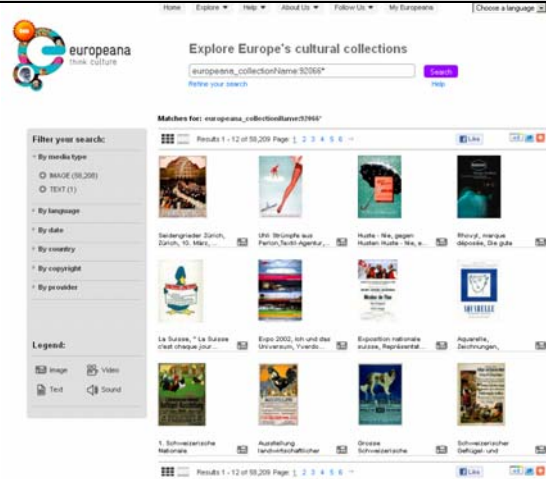| 11. Step: Provision to Europeana	12. Step: Harvesting by Europeana
The approved data is then provided to Europeana via our OAI Repository in one of the Europeana formats (EDM or ESE).	Europeana then harvests the ESE data from The European Library, via the REPOX system.

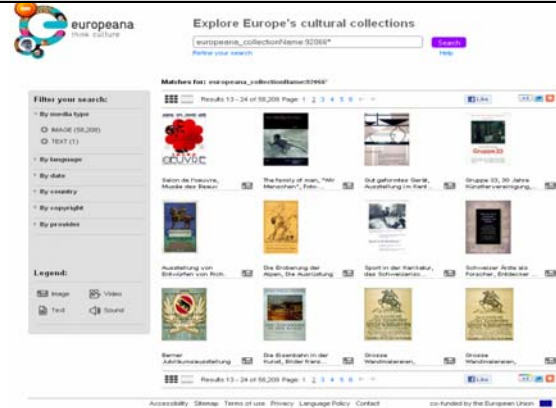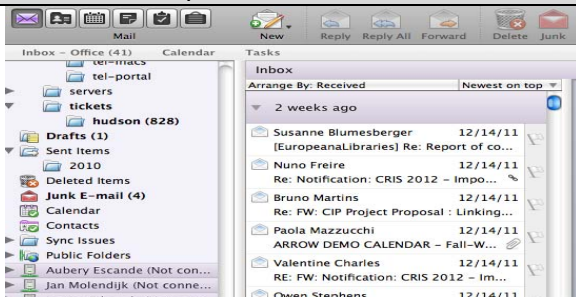| 13. Step: Enrichment at Europeana	14. Step. Acceptance at Europeana
Europeana further uses the UIM to load and enrich the data with cross domain thesauri, etc.	Europeana then provides the data on their acceptance portal to *The European Library* for final validation. The Europeana Ingestion team also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

7.3 ANNEX III: Definition of a Record

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*. Please see the definitions of Record, Metadata and Content in the [Glossary](#).

[Def.26] **Record**: The unit of **Metadata** concerning a single **Content** object.

[Def.27] **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.28] **Content**: The digital objects that can be accessed through Metadata. Content is typically held on Data Provider's/Aggregator's 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.

7.4 ANNEX IV: ESE Element Overview

Q: Which fields in ESE are mandatory, which are optional, which are preferred? Which are the minimum requirements to the data?

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