



DELIVERABLE

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

D4.1 – Requirements Infrastructure and Harvester

Revision: [1.2]

Authors:

Stefanie Ruehle (CERL)
José Borbinha (IST)
Gilberto Pedrosa (IST)

Project co-funded by the European Commission within the ICT Policy Support Programme

Dissemination Level

P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Rev.	Date	Author	Org.	Description
0.1	23/03/2011	Stefanie Ruehle	CERL	Initial draft
0.2	11/05/2011	Stefanie Ruehle	CERL	Revised draft
0.3	22/05/2011	José Borbinha	IST	Revision. First draft in formal template
0.4	23/05/2011	José Borbinha	IST	Revision (only typos and formatting)
0.5	25/05/2011	José Borbinha	IST	Revision (after comments from Valentine Charles, Robina Clayphan and Stefanie Ruehle)
0.6	27/05/2011	José Borbinha	IST	Revision after comments from the Aggregation Steering Committee
0.6.1	28/05/2011	José Borbinha	IST	Only fonts' adjustments...
0.9	08/6/2011	José Borbinha	IST	Extended revision after meeting in Den Haag
1.0	13/6/2011	José Borbinha	IST	Version for delivery
1.1	14/6/2011	José Borbinha	IST	Revision (clarified "full-text" as part of "data")
1.2	12/7/2011	José Borbinha	IST	Revision after internal review by Martin Moyle, Jan Molendijk and Marcin Werla.

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

Index

1. INTRODUCTION.....	1
1.1 Context within the project.....	1
1.2 Scope and Audience.....	1
1.3 Further Dependencies.....	2
1.4 Structure of this document	2
2. ABOUT REQUIREMENTS AND SYSTEM CONTEXT	3
2.1 Methodological Definitions.....	3
2.2 Addressed Facets	5
3. GENERIC GOALS AND DEFINITIONS.....	7
3.1 General Definitions	8
3.2 Open Issues.....	9
4. ARCHITECTURAL ASSUMPTIONS OF THE TEL AGGREGATOR.....	10
5. ACTORS OF THE TEL AGGREGATOR	11
6. THE TEL AGGREGATOR USE CASES.....	13
Use Case 1:Harvest Data Ingest.....	14
Scenario 1: Harvest Data Ingest	14
Use Case 2:Harvest Data Export.....	14
Scenario 2: Harvest Data Export.....	14
Use Case 3:Normalization and Enrichment	15
Use Case 4:Technical Reference	15
Scenario 3: Technical Reference.....	15
Use Case 5:Manage Data Provider	15
Scenario 4: Editing a Data Provider Record	15
Use Case 6:Manage Data Schema	15
Scenario 5: Editing a Data Schema	16
Scenario 6: Editing a Data Transformation	16
Scenario 7: Test and Asses Transformations	16
Use Case 7:Manage Data Ingests	16
Scenario 8: Manage a Data Ingest Task	16
Scenario 9: Audit a Data Ingest Task	17
Scenario 10: Validate Data Harvest Task	17
Use Case 8:Manage Data Exports.....	17
Scenario 11: Manage Data Exports.....	17
Use Case 9:Service Provider Report	17
Scenario 12: Service Provider Report	17
Use Case 10: Configuration and Maintenance	17
Scenario 13: Configuration and Maintenance	18
7. OTHER REQUIREMENTS AND ISSUES.....	19
8. FURTHER STEPS WITHIN WP4	20
9. REFERENCES	21

1. Introduction

This document presents the fundamental concepts and requirements for the European Library Aggregation Infrastructure. The goal of this infrastructure is to support the harvesting of digital content from individual libraries, transform it according to the needs, and made it available to Service Providers, namely to Europeana. For that, this infrastructure will consist in a system made of a set of computing services, operated by professionals with specific roles under the scope of The European Library, and interacting with external computing services operated by professionals of libraries and of Europeana and others.

1.1 Context within the project

In a broad perspective, this work is a consequence of the motivation for the Europeana Libraries project, stated in the DoW as:

- Bring to Europeana the digital collections of some of Europe’s leading research libraries from 11 countries. (...)
- Be the first project to offer digital collections where the text will be fully searchable in Europeana, making it possible to search inside books and other materials. Europeana Libraries will also devote time and expertise to enhancing full-text searching capabilities and features.
- Establish systems and processes capable of ingesting and indexing significant quantities of digitized material, including text, images, moving images and sound clips. The outcome will be an efficient and effective library-domain aggregator service for Europeana. Once the library aggregation model has been established over the two-year life of Europeana Libraries, the service will be fully capable of extension to other libraries across Europe...

In that scope, this work is part of the WP4 – “Extending The European Library aggregation infrastructure”, defined with the objectives of:

- Extend The European Library’s existing aggregation infrastructure to enable the aggregation of digital content from libraries in Europe for Europeana, including full-text content.
- Promote and build consensus on the adoption of standards related to the aggregation of digital content.
- Represent the European library-domain within the Europeana network and beyond.

In detail, this deliverable is the outcome of the Task 4.1, conceived with the purpose of:

“Develop the functional requirements for the library-domain aggregation infrastructure. This work will be built on The European Library’s existing business processes and will also consider the functionality currently providing in existing aggregation tools. (...)”

1.2 Scope and Audience

Requirements specify the properties a system needs to fulfil according to its objectives and scopes. Therefore, requirements must result from the defined goals of the system and of the related analysis.

To identify the requirements of the European Library Aggregation Infrastructure we therefore looked at the objectives of the Europeana Libraries project and examined processes and use cases of the current aggregation infrastructure of CERL and The European Library.

The main audience of this document is the Aggregation Steering Committee and the representatives of the partners libraries, which main concerns are here addressed. To address other complementary concerns, this

document will be complemented by a technical report focused on the design of the European Library Aggregation Infrastructure.

1.3 Further Dependencies

Further dependencies of this documents concern WP2 and WP3:

- In Task 2.1 is expected to develop "a range of scenarios for the aggregation of digital content into Europeana". These scenarios may imply a revision of this document.
- Planning the content ingestion in Task 3.2 conducts a survey. Results of that survey will improve the depth of knowledge about the content itself and may add further requirements concerning the handling of content.
- WP5 will define transformation rules from Europeana Semantic Elements (ESE) to the Europeana Data Model (EDM) that may extend the requirements for the transformation process.

Therefore, while the content of this document will support the implementation of the aggregation infrastructure, it might have to be revised depending on the results of WP2, WP3 and WP5.

1.4 Structure of this document

The document is organized as follows:

- Chapter 2 defines the main concepts relevant for this document.
- Chapter 3 describes the main goals of the Europeana Libraries project concerning the aggregation infrastructure and the directly associated requirements.
- Chapter 4 describes the Europeana Libraries Aggregation Infrastructure actors and related requirements.
- Chapter 5 describes the use cases
- Chapter 6 records other requirements and issues
- Chapter 7 describes further steps within WP 4
- Finally, References lists the most relevant external references, where we stress those that must be understood as companion documents to this one.

2. About Requirements and System Context

The key concepts on requirement engineering used in this document are inspired by [Ref.9], and the process aligned with the framework represented in the Figure 1. According to this framework, we understand this document as the result of a core activity reporting the results of requirements elicitation and negotiation. “The goal of the elicitation activity is to improve the understanding of the requirements, i.e., to achieve progress in the content dimension”, while the goal of the negotiation activity is to identify and solve the conflicts between the viewpoints of the different stakeholders [Ref.9].

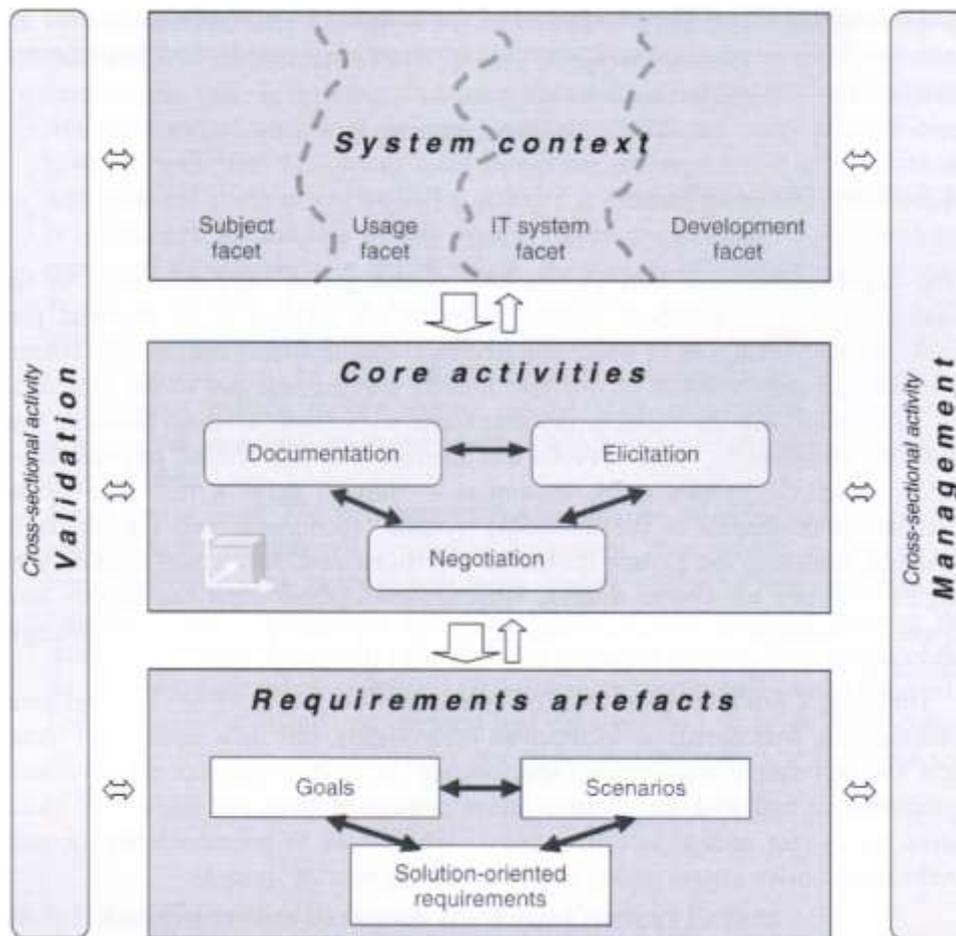


Figure 1: A requirement engineering framework, according to [Ref.9].

2.1 Methodological Definitions

Inspired by the mentioned framework, the requirements engineering methods and main concepts used in this document must be understood according to the definitions that follow.

[Def.1] **Actor:** An actor is a special class of stakeholder that "models a type of role played by an entity that interacts with the subject (e.g., by exchanging signals and data), but which is external to the subject (i.e., in the sense that an instance of an actor is not a part of the instance of its corresponding subject)." ¹ An actor may be therefore a person, software, hardware etc. and represents a role rather than a physical instance. E.g. a library may act as a content provider or as a service provider or appear in both roles. In the context of this project, we will understand the TEL Aggregator [Def.15] as the intended system, so the libraries, represented by their professionals and systems, will be understood as actors of that system.

[Def.2] **Goal:** "A goal is an intention with regard to the objectives, properties, or use of a system" [Ref.9]. A goal is therefore a business objective, mainly related with the "what" and "why" do we engage in an undertaking. When we use technological systems to support business activities, these systems are expected to be conceived, designed and implemented in order to clearly support the reaching of the business goals. The extension at how that support is really effective and efficient is commonly referred as alignment. In that sense the clear definition of the business goals is fundamental for the assessment of the alignment of the system with the business, making it relevant to stress the concept of goal.

[Def.3] **Open Issue:** An open issue is a business or technical issue detected in the context of the work reported in this document and requiring further clarification. This can occur because it'll depend from future results (for example, we must note that this document is being written before the terminus and final evaluation of the results of the projects Europeana V0.1 and EuropeanaConnect, which are expected to confirm or redefine Europeana requirements, especially the [Ref.7] and the [Ref.8]) or were identified by the first time and will require future investigation and consensus.

[Def.4] **Requirement:** Requirements specify the properties a system needs to fulfil according to its objectives and scopes and intended to be taken in consideration during the design and implementation of that system. Most of the time requirements are expressed as solution-oriented requirements, i.e., as an imposition on the structure or the behaviour of a system. In this sense, a requirement is a constraint, imposed or perceived from the interests of a stakeholder (identified from scenarios, or which might comprise for example legal requirements imposed by external entities), and especially from the goals. In a broader perspective, goals and scenarios also might be understood as expressions of requirements [Ref.9].

[Def.5] **Scenario:** "A scenario describes a concrete example of satisfying or failing to satisfy a goal (or a set of goals)" [Ref.9]. A scenario is a brief narrative, or story, that describes a hypothetical use of a system. A scenario must be clear in the actor using the system and what is that Actor trying to accomplish. In the context of this document the concept of scenario is used to describe one possible execution of a use case.

¹ OMG Unified Modeling Language™ (OMG UML), Superstructure Version 2.4, January 2011 p. 604 f. (<http://www.omg.org/spec/UML/2.4/Superstructure/Beta2/PDF>)

[Def.6] **Stakeholder:** A stakeholder is any identified entity with an expressed or perceived potential interest in a system. In our actual scenario, the main stockholders are: The European Library, the intended owner of the system to be built; the libraries partners in the consortium, as Data Providers; and Europeana. The major stakeholder is the business owner, who is the ultimate responsible by the expression of the business goals.

[Def.7] **Use Case:** A use case is a technique to describe a specific and clearly identified interaction between an actor and a system. Each use case provides one or more **scenarios** that convey how the system should interact with actors. Use case diagrams have been proving to be a very useful technique to express the functionality of the system as it is intended to be perceived by its actors, with no need to detail on technical or implementation details. Use cases also can be used "to capture the requirements of a system. The required behaviour of the subject is specified by one or more use cases, which are defined according to the needs of actors."² Therefore, there can be a tight interrelation between use cases and requirements. The use case technique is recognized as a potentially valuable supporting technique in the phases of requirements elicitation (the processes we run to identify, formulate, discuss and agree on requirements), as also a relevant technique to latter on support the modelling and design of that same system.

2.2 Addressed Facets

We also must be clear on defining the scope of this document concerning all the possible contextual concerns around the system to be developed. Considering the period during which this document was developed, relatively to the overall life span of the project, and the main intended audience (the business decision makers) we make it clear that:

According the reference framework, the scope of this document must be identified with the **subject facet** ("...the objects and events in the system's context that are relevant for the system") and the **usage facet** ("... comprise all aspects concerning the system usage by people and other systems") [Ref.9].

This implies the work reported in this document needs to be complemented to also comprise the IT system facet ("... comprises all aspects of the operational and technical environment, including policies and strategies defining restrictions or guidelines for the use of any type of technology or operational environment") and the development facet ("...process guidelines and constraints, development tools, quality assurance methods, maturity models, quality certifications, and other means or techniques for ensuring the quality...").

We also understand these facets as fundamentals, but which details concern to a more restrictive audience in the scope of the project, namely to the technical teams. Therefore, to address these facets it was decided to create a technical working group to address them, in a coordinated action complementary to this one and mainly related with the Task 4.2.

² OMG Unified Modeling Language™ (OMG UML), Superstructure Version 2.4, January 2011, p. 603

3. Generic Goals and Definitions

According to the reference definition (“A goal is an intention with regard to the objectives, properties, or use of a system” [Ref.9]), from the internal perspective of the scope of this work we can express the main goals of the TEL Aggregator as follow:

[Goal 1] Establish systems and processes capable of ingesting and indexing significant quantities of digitized material, including text, images, moving images and sound clips.

[Goal 2] “Extend The European Library’s existing aggregation infrastructure to enable the aggregation of digital content from libraries in Europe for Europeana, including full-text content”. (from the DoW)

[Goal 3] Offer, to Europeana in particular, but also to any other potentially interested service provider, **Metadata** [Def.14] and **Full-Text Data** [Def.13] where the text will be fully searchable, making it possible to search inside books and other materials.

The analysis of these goals motivates in the following main requirement:

[Req.1] The TEL Aggregator must take in consideration the most recent versions of the reference requirements defined by Europeana, namely the Technical Requirements as defined in [Ref.1] and [Ref.2], and the Business Requirements as defined in [Ref.3].

The references mentioned in [Req.1] are taken from the following sources:

- [Ref.1] Providing Content - Technical Requirements - <http://version1.europeana.eu/web/guest/technical-requirements/>
- [Ref.2] V1.0 Project - Technical Requirements - <http://version1.europeana.eu/web/europeana-project/technicaldocuments/>
- [Ref.3] V1.0 Project - Business Requirements - <http://version1.europeana.eu/web/europeana-project/documents/>

At this moment we also stress these specific documents, listed and linked from those indices, as especially relevant for this deliverable:

- [Ref.4] Europeana Data Provider & Aggregator Agreements (07/04/2010) - http://version1.europeana.eu/c/document_library/get_file?uuid=1c287538-d3c9-4843-9992-c4539f16aec0&groupId=10602
- [Ref.5] Metadata Mapping & Normalisation Guidelines for the Europeana Semantic Elements (Version 2.1 – 31/03/2011) - http://version1.europeana.eu/c/document_library/get_file?uuid=b3cfcf47-da0a-4c6b-b1d7-9b08e162643e&groupId=10128
- [Ref.6] Europeana Semantic Elements specifications v3.4 (Version 3.4 – 31/03/2011) - http://version1.europeana.eu/c/document_library/get_file?uuid=77376831-67cf-4cff-a7a2-7718388eecl1d&groupId=10128

3.1 General Definitions

Considering these specific goals we can now define:

[Def.8] **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. ([Ref.4])

[Def.9] **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 like linked data, ontologies, vocabularies, thesauri, classifications, taxonomies, etc. (definition taken from [Ref.4], where it is referred ad "Context Resources")

[Def.10] **Data:** Catch-all term including Metadata, Thumbnails, Audio and Moving image previews. ([Ref.4]). **In the scope of this document, this concept also includes, by default, Full-text Data** [Def.13].

[Def.11] **Data Aggregator:** Organisation that collects, formats and manages Data from Data Providers before make that available to Service Providers (such as Europeana). ([Ref.4])

[Def.12] **Data Provider:** Organisation that makes **Data** available to a Data Aggregator (such as the TEL Aggregator) or a Service Provider (such as Europeana). ([Ref.4])

[Def.13] **Full-text Data:** Data ([Def.10]) in the form of text representing literal transcriptions of written or spoken words from the Content ([Def.8]). This is a new class of Data to be considered in Europeana, related to the [Goal 3] and thus not covered (and so not to be confused) by the concepts of Context Resources ([Def.9]) or Metadata ([Def.14]).

[Def.14] **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. ([Ref.4])

[Def.15] **TEL Aggregator:** The Data Aggregator, realized by the European Library Aggregation Infrastructure under the responsibility of The European Library (and sometimes also mentioned in the DoW as the European Library Aggregation Infrastructure).

[Def.16] **Thumbnail:** Smaller and/or lower resolution version of still image Content. ([Ref.4])

[Def.17] **URI:** Uniform Resource Identifier, URLs (Uniform Resource Locators) are URIs. ([Ref.4])

3.2 Open Issues

A group analysis of the previously expressed goals and definition raises an issue:

[Open Issue 1] As a consequence of the [Def.13], we can assume for the short-term that the ESE element <unstored> can be used as requirement to support the [Goal 1] but it might not be sufficient to support with excellence the [Goal 3] . This raises an issue that should be presented to Europeana, for further discussion and eventual revision of ESE. Anyway, this issue is not a major constraint to the work to be developed in the initial phases of this project, so it is not considered impeditive.

[Open Issue 2] Even if this project is focussed on a specific community of Data Providers, the Europeana Libraries, which are expected to hold mainly bibliographic Content, in this moment it is not clear how homogeneous will be Content provided by those entities and how appropriate is the actual [Def.8]. For example, Europeana considers finding aids as an acceptable Content from archives, which from the perspective of a library is a kind of Data closer to Metadata than to Content.

4. Architectural Assumptions of the TEL Aggregator

This document is expected to be used as a basis for the design of the TEL Aggregator, for which some fundamental assumptions also need to be considered. These assumptions make the following **non-functional** requirements:

[Req.2] The TEL Aggregator must consider in its architecture the existence of a CRM (Customer Relationship Management) component or service, which in this moment is expected to be supported by Sugar CRM technology³.

[Req.3] The TEL Aggregator must consider in its architecture the existence of a data harvester and aggregator supported by the REPOX technology⁴.

[Req.4] The TEL Aggregator must consider in its architecture the existence of a technical reference service, to support the sharing of fundamental reference information

[Req.5] The TEL Aggregator must consider in its architecture the existence of a service for data profiling, normalization and enrichment..

³ <http://www.sugarcrm.com>

⁴ <http://rebox.ist.utl.pt/>

5. Actors of the TEL Aggregator

Recalling the [Def.1], we can present the following actors for the TEL Aggregator.

[Actor 1] **Data Provider**: Data Provider is defined in 0. Specifically in the context of this project, the Data Provider is a person or organisation using the **TEL Aggregator** to supply content to a Service Provider. Typically a **Data Provider** in this project is a national or research library, but considering the long-term implications of the [Goal 2] the results of this project cannot limit the participation of any library as future **Data Provider**.

[Actor 2] **Data Provider Service**: Data Provider Service is the computational service under the control of a **Data Provider**. It is necessary to identify this actor besides the [Actor 1] because the **TEL Aggregator** will have different specific use cases for each of them.

[Actor 3] **Service Provider**: The Service Provider is a person or organisation interested in the harvesting of **Data** from the **TEL Aggregator**. The main Service Provider of the **TEL Aggregator** is **Europeana**, but the **TEL Aggregator** may also be used by services of other aggregators or libraries collecting others or their own **Data**.

[Actor 4] **Service Provider Harvester**: The Service Provider Harvester is a remote service behaving under the control of a **Service Provider**.

[Actor 5] **Aggregation Team**: The Aggregation Team is a person, organisation etc. scheduling and monitoring the harvest, transformation, validation and provision of content in the **TEL Aggregator** and validating the results of these processes. The Aggregation Team behaves under the responsibility of The European Library.

[Actor 6] **Administrator**: The system administrator is a person, organisation etc. managing, monitoring and maintaining the system of the infrastructure concerning reliability and security. He is responsible for the performance of and the traffic on the systems, he schedules the processes and overviews the volume of the content, the usage of the systems and the partners involved. The **Administrator** is a generalization of the **Aggregation Team**, in the sense that it also comprises all the associated roles.

The actual definition of [Actor 3] brings a new open issue:

[Open Issue 3] The terms and conditions for use of the service by other Service Providers than Europeana must be clarified (this issue is not clarified by the actual DoW, but is in the scope of the WP2).

Consequently, the analysis of the [Open Issue 3] recommends for now, by precaution, the assumption to recognize Europeana as its unique source of requirements for Service Provider (but this assumption can be revised, depending of future clarifications to the [Open Issue 3])

The actual definition of [Actor 5] also might need to be revisited, as it is recognized as an eventually too generic definition, considering that:

[Open Issue 4] At this moment the Aggregation Team is understood as a unique generic entity, but this conceptualization might have to be specialized in the future, as it is expected that specialized roles might need to be conceptualized, as for example a role to process Data Provider

applications, other only focused in the ingesting, other focused on the Data quality, and other focused on the Data transformations.

Finally, these global requirements should apply to all the operations:

[Req.6] All the actions performed in the **TEL Aggregator** by the **Administrator** or by **Aggregation Team** are definitive, with no need of any other further confirmation

[Req.7] Each action performed in the **TEL Aggregator** by a **Data Provider** or by a **Service Provider** must be analysed and decided if it is definitive or if it needs a further confirmation of the **Aggregator Team**.

6. The TEL Aggregator Use Cases

Considering the non-functional requirements expressed in [Req.2], [Req.3], [Req.4] and [Req.5], we can assume from this perspective the contexts represented in the Figure 2 and in the .

Figure 2 shows the Data Aggregator use cases, Figure 3 shows the use cases for the Reference, TEL CRM and Normalizer, which are individually detailed in the following sections.

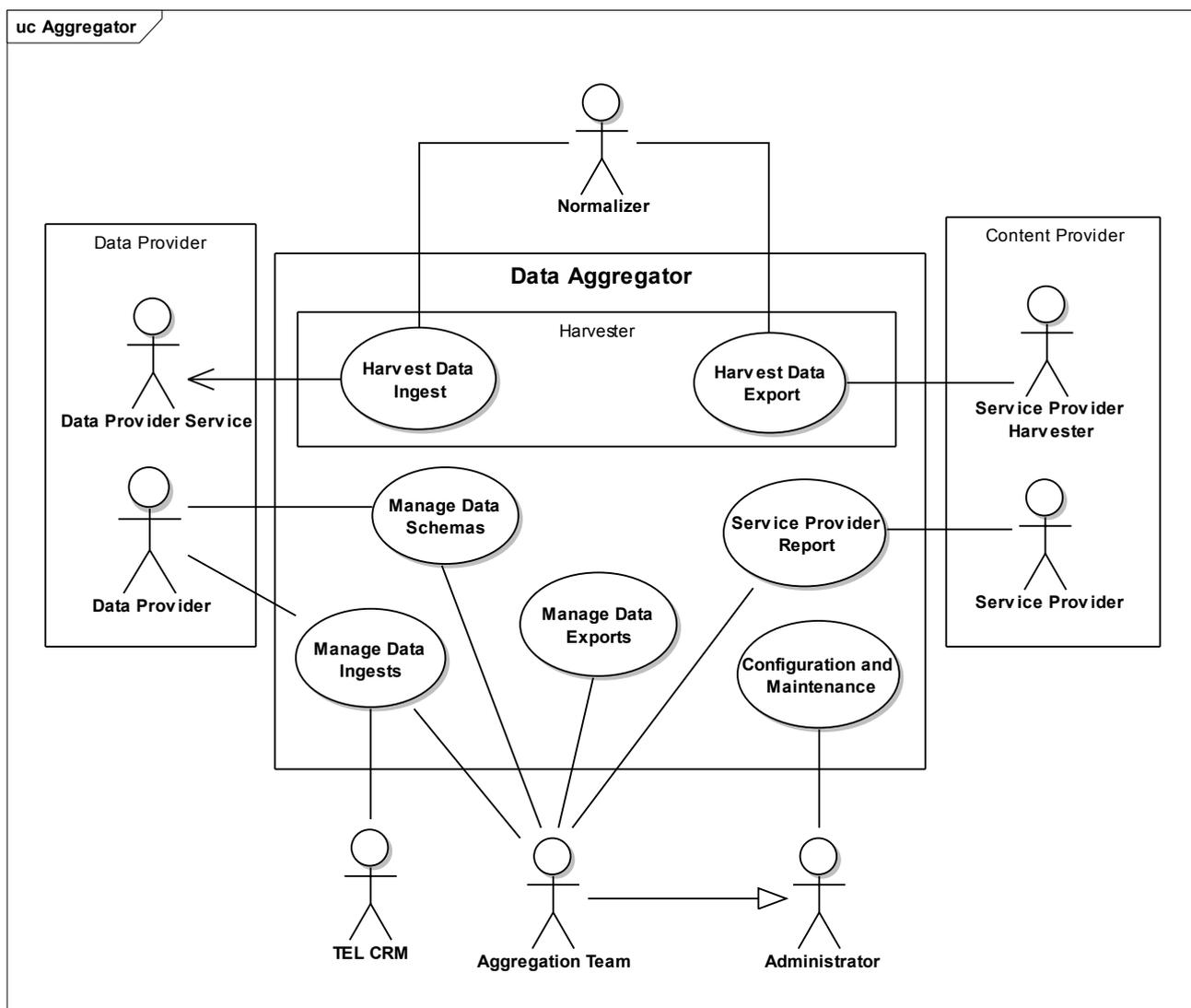


Figure 2: TEL Aggregator Use Cases

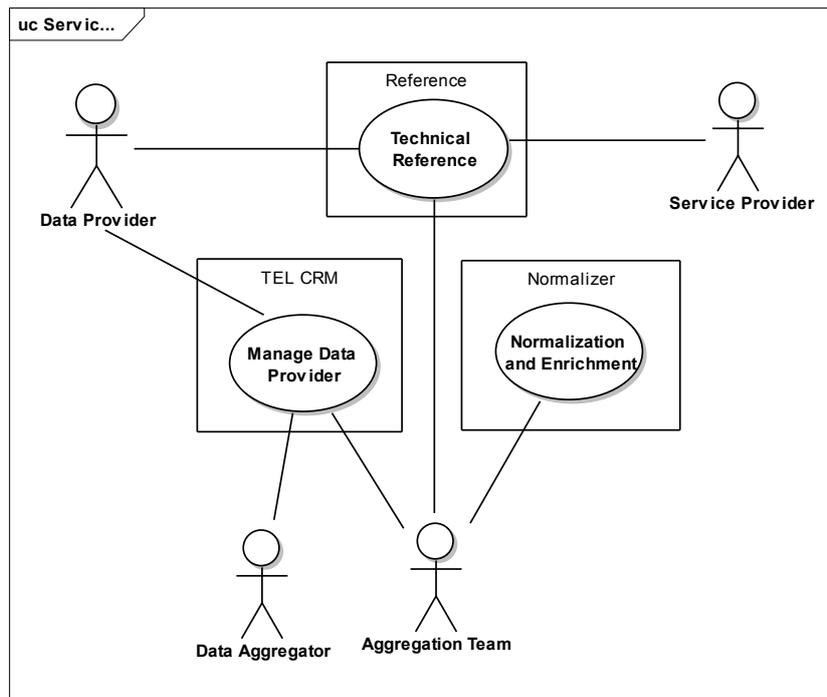


Figure 3: Use cases for the Reference, TEL CRM and Normalizer.

Use Case 1: Harvest Data Ingest

Scenario 1: Harvest Data Ingest

[Req.8] The system must execute the Data Ingest Tasks as configured.

[Req.9] The system must be able to harvest Metadata at least through OAI-PMH.

[Open Issue 5] It remains an open issue if it will exist a unique way (e.g., OAI-PMH) to harvest of Full-text Data or if the TEL Aggregator will not have to support multiple solutions (the expected scenario, considering the coexisting of solutions based on OAI-PMH but also other solutions based on HTTP-GET, FTP, etc.). This issue must be clarified after the first lessons from the execution of the content ingestion plan (defined in the WP3) and while the making of the “D4.3 – Report on how the full-text content will be made available to Europeana”.

Use Case 2: Harvest Data Export

[Def.18] **Data Export Task:** A Data Export Task is a task of harvesting of a Data collection from the TEL Aggregator by a Service Provider.

Scenario 2: Harvest Data Export

[Req.10] It must be possible for a Service Provider to execute a Data Export Task at any moment and for any available collection of Data.

[Open Issue 6] It remains an open issue the possible IPR models to be considered by the TEL Aggregator. If these models are defined, the [Req.10] must be redefined.

Use Case 3: Normalization and Enrichment

This use case must support functions to analyse, profile, normalize and enrich ingested Data, and make the results available for Data Export. These services must support processes to assess the quality and completeness of the results of the Data Ingest Tasks.

[Open Issue 7] It remains an open issue the characterization of the new Data produced by these means. It will not be the original data harvested from the data provider, but its reference to the original data sets also cannot be lost. This is an open issue to be discussed with the Service Providers, namely with Europeana.

Use Case 4: Technical Reference

This use case must support a forum for publishing and sharing of technical reference documents, for dissemination of information, and for interaction by all the human actors.

Scenario 3: Technical Reference

[Open Issue 8] The detailed elicitation of requirements and design of the “Use Case 4: Technical Reference” will be developed in a further step. It will be important to consider here different roles for, for example, members of the Aggregation Team that will have special roles to control content, and/or the roles of the other users considering the access to information versus the submission of contributions or annotations.

Use Case 5: Manage Data Provider

This use case supports the registration and management of all the information related to a Data Provider, such as contact data, data collections and harvesting processes. This use case comprises several scenarios, for which the following definitions are relevant:

[Def.19] 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 also the information about all the Data that the Data Provider is willing to provide for Data Harvest Tasks.

Scenario 4: Editing a Data Provider Record

[Req.11] It must be possible to edit a Data Provider Record.

The actor edits a Data Provider Record:

- if the action is for a new Data Provider, a new record is created and edited;
- if the action is for an already registered Data Provider, that existing record is edited.

Use Case 6: Manage Data Schema

[Def.20] Data Schema: A Data Schema is a description of the structure of a collection of Data.

[Goal 4] Offer to the Data Providers a support to map their Data schemas to the schemas the TEL Aggregator wants to make available to Data Providers, namely that or those required by Europeana.

Scenario 5: Editing a Data Schema

[Req.12] Must be provided a mechanism to register and manage Data schemas intended to be recognized in the Use Case 1: Harvest Data Ingest and/or in the Use Case 2: Harvest Data Ingest

The actor edits a Data Schema:

- if the action is for a new schema, a new schema is created and edited;
- if the action is for an already registered task, that existing schema is edited (which can imply simply to annotate it).

Scenario 6: Editing a Data Transformation

[Req.13] Must be provided a mechanism to associate the automatic application of transformations, so Data can be harvested by Service Providers in schemas different of the schemas it was harvested from the Data Providers.

[Req.14] It must be possible to edit a Data schema transformation.

The actor edits a Data Transformation:

- if the action is for a new transformation, a new transformation is created and edited;
- if the action is for an already registered transformation, that existing transformation is edited (it must be considered to make it possible to reuse existing transformations, to support this action more efficiently and effectively).

Scenario 7: Test and Asses Transformations

[Req.15] Must be provided a mechanism to test Data transformations using Data existing in the system and providing as feedback audit reports.

[Open Issue 9] The definitive non-functional requirements and design decisions of how to register in the TEL Aggregator the description of the schemas and of the mappings must be subject to a more detailed technical analysis, ideally involving the technical teams of The European Library and Europeana.

Use Case 7: Manage Data Ingests

[Def.21] **Data Ingest Task:** A Data Ingest Task is a task of harvesting by the TEL Aggregator of a Data collection from a Data Provider.

Scenario 8: Manage a Data Ingest Task

[Req.16] It must be possible to edit a Data Ingest Task.

The actor edits a Data Harvest Ingest Tasks:

- if the action is for a new task, a new task is created and edited;
- if the action is for an already registered task, that existing task is edited.

[Req.17] The TEL Aggregator must have a mechanism to declare and manage rules to prioritize the execution of Data Harvest Tasks, making it possible to automatize generic scenarios.

[Req.18] It must be provided a console for the Aggregation Team to monitor all the Data Ingest Processes, either those being executed in real-time or those scheduled for a future time.

Scenario 9: Audit a Data Ingest Task

Ideally, the assessment of a Data Harvest Task should be performed (self-assessed) by the Data Provider, but the lack of technical skills from that actor might imply an intervention of the Aggregation Team.

[Req.19] The results of an execution of a Data Harvest Task must be classified of “Test” or “OK”, where “OK” means the results were considered conformant with the European requirements and “Test” means that conformance still has to be assessed and confirmed.

[Req.20] It must be provided an audit report of the results of each Data Ingest Task, comprising the concerns of quantity (number of records and attributes in the records, etc.) and of quality (consistency of the values of the attributes, conformance with the schema, etc.).

[Req.21] It must be provided an audit report crossing the result of each Data Ingest Task from a Data Provider with each possible mapping to a Data schema registered in the TEL Aggregator (when the schema is one requested by Europeana, this audit will make it possible for each Data Provider to self-audit the delivery of its Data to Europeana).

Scenario 10: Validate Data Harvest Task

[Req.22] The classification of “OK” for a Data Harvest Task must be restricted to the Aggregation Team.

[Req.23] It must be possible, for a Data Harvest Task that had an execution that was considered “OK”, to record that future executions also can be automatically considered OK. This rule can remain effective until it is explicitly changed or the Data Harvest Task is changed itself.

Use Case 8: Manage Data Exports

Scenario 11: Manage Data Exports

[Req.24] It must be provided a console for the Aggregation Team to monitor all the Data Export Tasks being executed in real-time.

Use Case 9: Service Provider Report

Scenario 12: Service Provider Report

[Req.25] It must be provided a report of each execution of a Data Harvest Task performed by a Service Provider.

[Req.26] It must be provided an audit report of the results of each Data Harvest Task, comprising the concerns of quantity (number of records and attributes in the records, etc.) and of quality (consistency of the values of the attributes, conformance with the exporting schema, etc.).

Use Case 10: Configuration and Maintenance

This use case must support all the scenarios of configuration and maintenance of the Data Aggregation and of its associated services.

Scenario 13: Configuration and Maintenance

[Req.27] The Administrator must be able to register users for the role of the Aggregation Team

[Open Issue 10] The detailed elicitation of requirements and design of the “Use Case 10: Configuration and Maintenance” will be developed in a further step.

7. Other Requirements and Issues

[Req.28] All the reports provided by the TEL Aggregator must be able to be consulted on-line, downloaded with the information structured, or scheduled to be sent by pull techniques such as email.

[Open Issue 11] Other non-functional requirements and design decisions that can have implications on performance must be subject to specific analysis involving the technical teams of The European Library and Europeana.

8. Further steps within WP4

Requirements due to the use of full-text with EDM will be published in an extra report in month 12 (D4.3 - Report on how the full-text content will be made available to Europeana). In that sense the next steps in WP 4 are:

- Based on the requirements of this document, the business processes of The European Library will be redesigned and implemented between month 6 and month 12. At the same time requirements for the integration of the full-text harvester will be better defined and reported in the D4.3 (during this period, the full-text harvest will re-use the pragmatic experience from the TELplus project).
- Month 6 to 12 is the first phase of the extension of the The European Library ingestion infrastructure to the Europeana Libraries Aggregation infrastructure. At the beginning of this phase it could become necessary to modify the requirements according to the results of the surveys of WP 2, WP 3 and WP 5. Besides that, WP4 also will analyse how the Europeana Data Model may be used with full-text. Results of this analysis also will be reported in the D4.3.
- In month 13 to 18 the processes and technical support by the infrastructure (including the handling of full-text content) will be validated. Systems of the infrastructure will be proved according to the requirements and will be modified according to validation results.
- Finally in month 18 to 24 the robustness, scalability and performance of the infrastructure and its systems will be validated and optimized.

9. References

Europeana publishes its reference document online. For the purpose of this document the most relevant references were identified from the following indices:

- [Ref.1] Providing Content - Technical Requirements - <http://version1.europeana.eu/web/guest/technical-requirements/>
- [Ref.2] V1.0 Project - Technical Requirements - <http://version1.europeana.eu/web/europeana-project/technicaldocuments/>
- [Ref.3] V1.0 Project - Business Requirements - <http://version1.europeana.eu/web/europeana-project/documents/>

More specifically, this document uses and makes specific reference to the following references from the indices above:

- [Ref.4] Europeana Data Provider & Aggregator Agreements (07/04/2010) - http://version1.europeana.eu/c/document_library/get_file?uuid=1c287538-d3c9-4843-9992-c4539f16aec0&groupId=10602
- [Ref.5] Metadata Mapping & Normalisation Guidelines for the Europeana Semantic Elements (Version 2.1 – 31/03/2011) - http://version1.europeana.eu/c/document_library/get_file?uuid=b3cfcf47-da0a-4c6b-b1d7-9b08e162643e&groupId=10128
- [Ref.6] Europeana Semantic Elements specifications v3.4 (Version 3.4 – 31/03/2011) - http://version1.europeana.eu/c/document_library/get_file?uuid=77376831-67cf-4cff-a7a2-7718388eec1d&groupId=10128
- [Ref.7] Functional Requirements: SpecificationsDanubeRequirementsContentInToolsIngestion tools - <http://europeanalabs.eu/wiki/SpecificationsDanubeRequirementsContentInTools>
- [Ref.8] Functional and Technical Specifications for requirement #1322 United Ingestion Toolset - <http://europeanalabs.eu/wiki/DanubeFunctionalTechnicalSpecificationContentInTools>
- [Ref.9] Klaus Pohl: Requirements Engineering - Fundamentals, Principles, and Techniques. Springer 2010: I-XVII, 1-813. ISBN 978-3-642-12577-5