DELIVERABLE - Extension

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

Extended Revised Version

Revision: [2.5]

Authors:

Stefanie Ruehle (CERL)

José Borbinha (IST)

Gilberto Pedrosa (IST)

Andreas Juffinger (TEL)

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
2.0	27/11/2011	José Borbinha	IST	Extended version considering the revised architecture
2.1	28/11/2011	José Borbinha	IST	Changes in the document style
2.3	01/12/2011	Stefanie Rühle	CERL	Minor Changes in typing and wording
2.4	05/11/2011	José Borbinha	IST	Minor revision
2.5	05/12/2011	Andreas Juffinger	IST	Minor revision

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

This document presents the fundamental concepts and requirements for the European Library Aggregation Infrastructure. It must be considered an extension to the deliverable D4.1, with an update of the core information.

1.1 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 actors of TEL Aggregation Infrastructure.
- Chapter 5 describes the use cases of the TEL Aggregation Infrastructure, and related requirements
- Chapter 6 presents the architectural assumptions of the TEL Aggregation Infrastructure.
- Chapter 7consolidates the architectural assumptions with the use cases.
- Chapter 8 presents remaining requirements (mainly non-functional)
- Finally, References lists the most relevant external references, where we stress those that must be understood as companion documents to this one.

1.2 Methodological Definitions

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.26] 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 technological systems 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.

[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.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] **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.

[Def.6] **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**.

[Def.7] **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.

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

2. 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.24] and **Full-Text Data** [Def.22] 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-a7a2-7718388eec1d&groupId=10128

2.1 General Definitions

[Def.8] **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 others 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 named simply of "Catalogue" for the same purpose).

[Def.9] **Collection:** An intentionally defined set of **Content,** compiled under a specific policy. This is a common concept in the library domain, so it is here used with the same meaning as in that domain.

[Def.10] **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.11] **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.12] **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.22].

[Def.13] **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.14] **Data Collection:** The **Data** corresponding to a specific **Collection**.

[Def.15] **Data Export Task**: This is a task of harvesting a Data collection from the TEL Aggregator by a Service Provider.

[Def.16] Data Ingest Task: A task of harvesting a Data Collection from a Data Provider.

[Def.17] **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.18] **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.

[Def.19] **Data Schema**: A description of the structure of a specific Data.

[Def.20] Enriched Data: Data that has been subject to a process of Enrichment, Normalization or Transformation.

[Def.21] **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 to assign to the attributes can consist in 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 **Normalization.**

[Def.22] **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 [Goal 3]and thus not covered (and so not to be confused) by the concepts of **Contextual Resources** ([Def.11]) or **Metadata**.

[Def.23] **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.24] **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.25] **Normalization**: 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 normalization of values (e.g. the normalization of date values to ISO 8601 compliant strings).

[Def.26] **TEL Aggregator**: The **Data** aggregator system 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.27] **Thumbnail**: Smaller and/or lower resolution version of still image Content. ([Ref.4])

[Def.28] **Transformation**: A kind of **Enrichment** by applying a set of **Mapping** rules to **Raw Data** in order to produce new **Enriched Data** structured according to a target **Data Schema**. It is important to stress that a **Transformation** only uses the Raw Data "as it is", which might imply the need of **Normalization** to assure that the **Enriched Data** is fully conformant with the target **Data Schema.**

[Def.29] **URI**: Uniform Resource Identifier, URLs (Uniform Resource Locators) are URIs. ([Ref.4])

[Def.30] Raw Data: The Data the TEL Aggregator collects from the Data Providers.

[Def.31] **Record**: The unit of **Metadata** concerning a single **Content** object.

2.2 Open Issues

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

[Open Issue 1] As a consequence of the [Def.22], 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 hole 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 then to Content.

3. Actors of the TEL Aggregator

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

[Actor 1] **Administrator**: This 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.

[Actor 2] **Aggregation Team**: This 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 3] **Data Provider**: This concept is defined in [Ref.1]. Specifically in the context of this project, the Data Provider is an organisation or person using the **TEL Aggregator** to supply content to a Service Provider. Specifically, 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 4] **Data Provider Service**: This 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 5] **Service Provider**: This 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 6] **Service Provider Harvester**: This is a remote service behaving under the control of a **Service Provider**.

The actual definition of [Actor 5] 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 2] 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.

4. The TEL Aggregator Use Cases

The fundamental use cases of the TEL Aggregator are represented in the Figure 1.

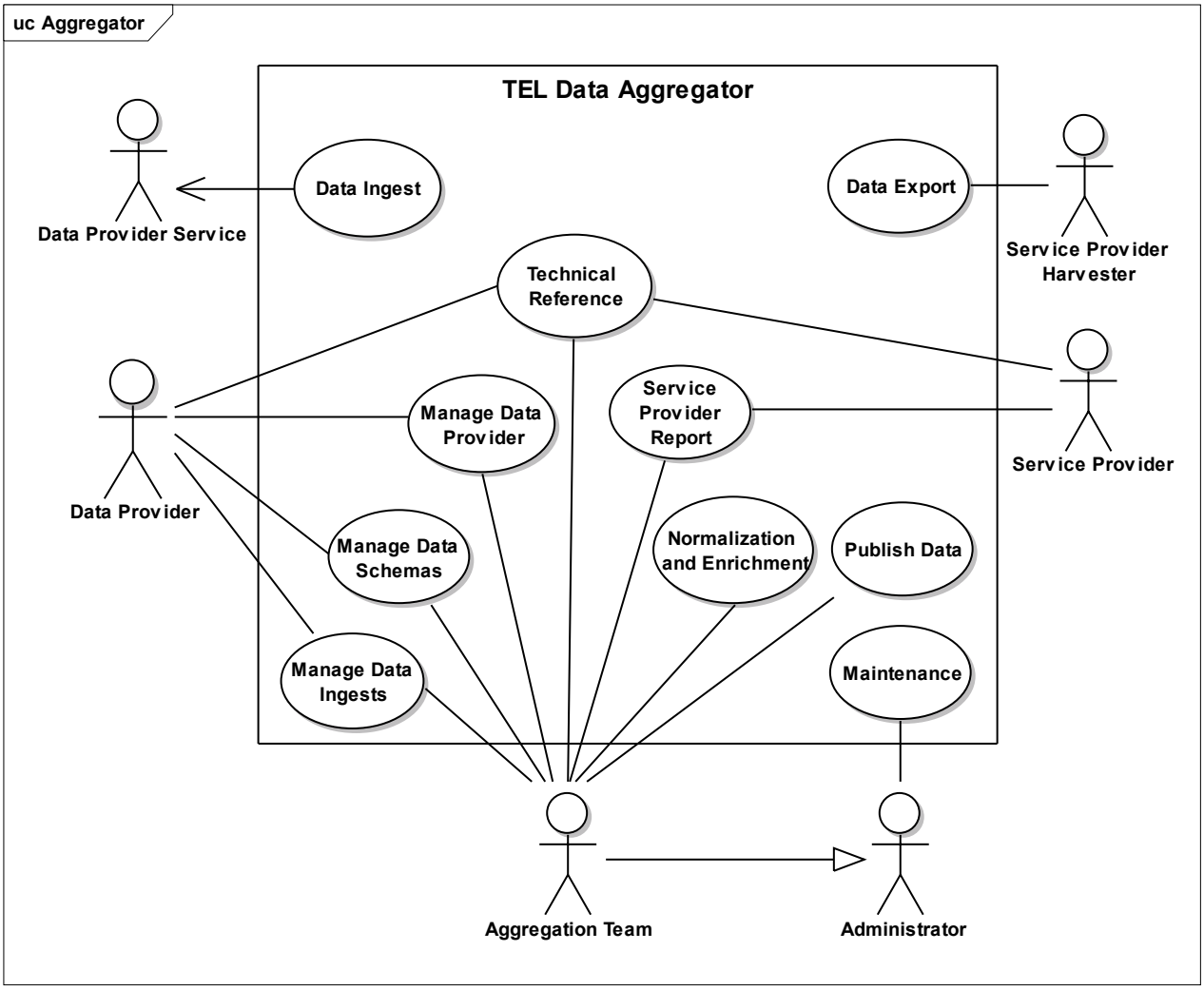

Figure 1: TEL Aggregator Use Cases

These global requirements should apply:

[Req.2] All the actions performed in the TEL Aggregator by the Administrator or by the Aggregation Team are definitive, with no need of any other further confirmation by any other actor.

[Req.3] The action possible to be performed in the **TEL Aggregator** by a **Data Provider** or by a **Service Provider** must be able to be defined by the **Aggregation Team**, at any moment, to be subject to a confirmation by the **Aggregation Team**, or to be executed with no need of that confirmation.

[Req.4] Each subsystem of the **TEL Aggregator** system must provide evidence, to be accessible to the **Aggregation Team**, that it executed any operation as configured (this requirement is intended to stress to the design and development teams the need of the necessary ***record keeping*** features for this purpose).

[Req.5] The **TEL Aggregator** system must provide evidence to the **Aggregation Team** that it executed any operation as configured (this requirement is intended to stress to the development team the need of the necessary ***record keeping*** features for this purpose).

[Req.6] The **TEL Aggregation Team** must be able to monitor, in real time, all the actions being executed by the system and change their course of action (ultimately, any action in progress must be able to be aborted with no effect to the state of the system verified before that action was started).

[Req.7] All the reports provided by the TEL Aggregator must be able to be consulted on-line, downloaded as structured information, or scheduled to be sent by pull techniques such as email.

Use Case 1: Data Ingest

This use case comprises the following scenarios:

Scenario 1: Metadata Ingest

[Req.8] The system must be able to harvest **Metadata** at least via **OAI-PMH, FTP, HTTP and Z39.50**.

[Req.9] The system must be tolerant and robust to not conformant **OAI-PMH** features from the side of the **Data Providers**. Any verified system's limitation to this requirement must have an acceptable technical explanation.

[Req.10] The system must be able to harvest **Metadata** as **LoD**[3].

[Open Issue 5] It remains an open issue what best techniques must be used to harvest **Metadata** as **LoD**. The research of the possible options ant techniques must be aligned with the similar efforts being carried out in the wider scope of the Europeana initiative. This implies that the [Req.10] will be pending of these decisions.

[Req.11] The system must ingest the **Metadata** constantly and automatically during the time slots the **Content Provider** has announced and therefore as it must had been defined for the **Data Ingest Task**.

Scenario 2: Thumbnail Ingest

[Req.12] The system must be able to harvest the **Thumbnails** when those are referenced in the **Metadata**.

[Req.13] The system must harvest the **Thumbnails** constantly and automatically during the time slots the **Content Provider** has announced and therefore as it must had been defined for the **Data Ingest Task**.

Scenario 3: Full-Text Ingest

[Req.14] The system must be able to harvest **Full-text** when that is referenced in the **Metadata**.

[Open Issue 6] It remains an open issue if it will exist a unique way 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".

[Req.15] The system must harvest the **Full-text** constantly and automatically during the time slots the **Content Provider** has announced and therefore as it must had been defined for the **Data Ingest Task**.

[3] http://linkeddata.org/

Use Case 2: Data Export

Dependency: Any analysis and decision on this use case also must take in consideration the stated for the "Use Case 9: Publish Data".

[Req.16] It must be possible for a **Service Provider** to execute a **Data Export Task** at any moment and for any set of **Data** available to it.

[Open Issue 7] It remains an open issue the possible IPR models to be considered by the **TEL Aggregator**. When these models are defined, the [Req.16] must be revisited.

Scenario 4: Data Export by OAI-PMH

[Req.17] The system must be able to export **Metadata** through **OAI-PMH**.

Scenario 5: Data Export by SRU

[Req.18] The system must be able to export **Metadata** through **SRU**[4].

Scenario 6: Data Export as LoD

[Req.19] The system must be able to export **Metadata** as **LoD**.

[Open Issue 8] (see [Open Issue 5] …)

Use Case 3: Normalization and Enrichment

The purpose of this use case is to monitor, normalize and enrich ingested Data. It is expected to also cover the assessment of the quality and completeness of the results of the Data Ingest Tasks.

[Open Issue 9] It remains an open issue the formal entity definition of the new Data produced by the Normalization and Enrichment. It will be not anymore 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.

[Req.20] It must be possible to validate the **Enriched Data** concerning compliance to the expected format/schema and must be produced non-conformity reports accessible to the **Aggregation Team**.

[Req.21] When any **Enrichment** task fails because of ambiguous **Metadata** content it must be possible for the **Aggregation Team** to intervene and at least solve the ambiguity manually.

Scenario 7: Data Transformation

[Req.22] The system must be able to **Transform** all the ingested **Metadata** (namely OAI-DC, MARC21, UNIMARC, MODS, ESE) for the target **Data Schemas** internally defined by TEL, namely to the TEL Application Profile

[Open Issue 10] Data transformations to other schemas to be required in the future by Europeana also should be possible, but that must be subject to a specific analysis. This must comprise.

[4] http://www.loc.gov/standards/sru/

Scenario 8: Data Normalization

[Req.23] The system must be able to assure that the **Enriched Data** is conformant with the **Data Schemas** internally defined by TEL (namely to the **TEL Application Profile).**

[Open Issue 11] Data Normalization to other schemas to be required in the future by Europeana also should be possible, but that must be subject to a specific analysis.

Scenario 9: Data Enrichment

[Req.24] The system must be able to apply techniques to produce **Enriched Data** making use of **Contextual Resources**.

[Req.25] The system must be able to apply techniques to produce **Enriched Data** making use of the harvested **Full-Text**.

Use Case 4: Technical Reference

Scenario 10: Technical Reference

[Req.26] The system must support a forum for dissemination of information, sharing of knowledge, and for controlled interaction by all its human actors.

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 the following scenarios:

Scenario 11: Create a Data Provider Record

[Req.27] The system must support the creation of a new **Data Provider Record** by the **TEL Aggregation Team** or by a **Data Provider** itself.

[Req.28] The record of a Data Provider should concern the owner of the collection and contact address, the name of the collection, used standards and schemas, which interfaces and timeslots may be used. If mandatory information is missing, the user must get an error report.

[Req.29] An alert to the Aggregation Team if a new Data Provider Record applies.

Scenario 12: Editing a Data Provider Record

[Req.30] The system must support the editing of a **Data Provider** Record (including its removal) by the **Aggregation Team**.

[Req.31] An alert to the Aggregation Team if a Data Provider Record was changed.

Use Case 6: Manage Data Schema

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

[Req.32] When Enrichment fails because of ambiguous Metadata content it must be possible for the Aggregation team to intervene and at least solve the ambiguity manually.

This use case supports the registration and management of all the information related to a Data Schema, for which the following scenarios are relevant:

Scenario 13: Create a Mapping

[Req.33] The system must support the creation of a new **Mapping** by the **Aggregation Team.**

Scenario 14: Editing a Mapping

[Req.34] The system must support the editing of a **Mapping** (including its removal) by the **Aggregation Team.**

Use Case 7: Manage Data Ingests

Scenario 15: Create a Data Ingest Task

[Req.35] The system must support the creation of a new **Data Ingest Task** by the **Aggregation Team**, concerning a specific **Data Provider.**

Scenario 16: Edit a Data Ingest Task

[Req.36] The system must support the editing (including removal) of a **Data Ingest Task** by the **Aggregation Team**, concerning a specific **Data Provider**.

[Req.37] The system must have a mechanism to declare and manage rules to prioritize the execution of Data Ingest Tasks, making it possible to automatize generic scenarios.

Scenario 17: Manage Data Ingest Tasks

[Req.38] The system must support the editing (including removal) of a **Data Ingest Task** by the **Aggregation Team**, concerning a specific **Data Provider**.

Scenario 18: Audit a Data Ingest Task

[Req.39] The results of a Data Ingest Task must be validated concerning compliance to standard schemas and it must be keep record about any detected non-conformity.

[Goal 5] In an ideal scenario, the assessment of a **Data Ingest Task** should be performed (self-assessed) by the **Data Provider**.

The lack of technical skills from those actors might imply the intervention of the Aggregation Team to assess the quality of a **Data Ingest Task**.

[Req.40] The results of an execution of a **Data Ingest Task** must be classified of "OK" or "Test", where "OK" means the results were considered conformant with all the requirements and "Test" means that conformance still has to be assessed and confirmed.

[Req.41] It must be provided a 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.42] It must be provided a report describing the changes in the **Data** resulted from each **Data Ingest Task** as a consequence of the application of each available **Enrichment** process applied to each **Data Ingest**.

Scenario 19: Validate Data Ingest Task

[Req.43] The classification of "OK" for a **Data Ingest Task** must be restricted to the **Aggregation Team**.

[Req.44] It must be possible, for a **Data Ingest Task** that had an execution that was considered "OK", to also automatically consider "OK" a future execution of the same **Data Ingest Task**. This rule can remain effective until explicitly changed (or the **Data Ingest Task** is changed itself).

[Req.45] It must be possible to validate the links to digital objects by a link checker and provide reports to the **Aggregation Team** if they are broken.

Use Case 8: Service Provider Report

Scenario 20: Service Provider Report

[Req.46] It must be provided a report of each execution of a Data Export Task performed by a Service Provider, 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 9: Publish Data

Dependency: Any analysis and decision on this use case also must take in consideration the stated for the "Use Case 2: Data Export".

Scenario 21: Creating a Data Export

[Req.47] Must be provided a mechanism to associate the automatic application of any possible **Enrichment** scenario, so **Data** can be published for **Service Providers** in schemas different of the schema it was harvested from the **Data Provider**.

Scenario 22: Manage Data Exports

[Req.48] It must be possible to define controlled data exports for only specific **Data Providers.**

Use Case 10: Maintenance

This use case must support all the tasks of configuration and maintenance of the Data Aggregator and of its associated services, as proved to be necessary (namely, the management of users, deployment of services, etc.).

[Open Issue 12] Due to its technical specificities and dependency of the final system architecture, the detailed elicitation of requirements and design of the "Use Case 10: Maintenance" will be developed in a further step.

5. 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 architectural **non-functional** requirements, as described below.

[Req.49] The **TEL Aggregator** must be designed as an infrastructure making the best use of the following existing applications (which when necessary will be extended or complemented to support all the desired requirements): **REPOX**[5], **UIM** and **SugarCRM**[6].

[Req.50] Each subsystem of the **TEL Aggregator** system must provide detailed information, to be accessible to the **Administrator**, of any failure (this requirement is intended to stress to the design and development teams the need of the necessary ***record keeping*** features for this purpose).

From a perspective of architecture of services, those assumed in the conceptual architecture of the TEL Aggregator are defined as follow (represented in the Figure 2).

[Req.51] The **Harvesting** service must be responsible by bringing the Data on site, harvesting it from the Data Providers.

[Open Issue 13] The REPOX system currently supports harvesting of Metadata by OAI-PMH, HTTTP, FTP and Z39.50, but it needs to be extended for efficient SRU harvesting, authority file handling, full text harvesting and EDM handling. Of all these, only the EDM handling is a relevant issues, as already pointed in the [Open Issue 5].

[Req.52] When the **Raw Data** is provided according to a Data Schema not recognized by the UIM system, a **Transformation** to a recognized **Data Schema** must be done by the **REPOX** system as part of the **Data Ingesting Task** and with a result auditable as part of the "Scenario 19: Validate Data Ingest Task".

[Open Issue 14] It is expected that UIM will handle the Data Schemas OAI-DC, ESE, TEL, EDM, MARC-21, UNIMARC and EAD, but the definitive list of Data Schemas supported by UIM is not definitely defined, so what must be supported by UIM or by REPOX must be constantly revised during the project lifetime.

[Req.53] The **publishing** interfaces must make the **Data** available in all the required variants, including Metadata provision to Europeana.

[Req.54] The **Configuration** service must be a combined configuration service for all systems, to ensure a minimum configuration effort by the operators.

Note: The combination of the configuration management for all relevant systems with the SugarCRM is considered highly beneficial, because that creates a single point of reference where all information about providers and the appropriate ingestion steps can be combined.

[Req.55] REPOX must overwrite updated Metadata records and Full-Text automatically.

[Req.56] UIM must overwrite updated Enriched Data records automatically.

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

[6] http://www.sugarcrm.com

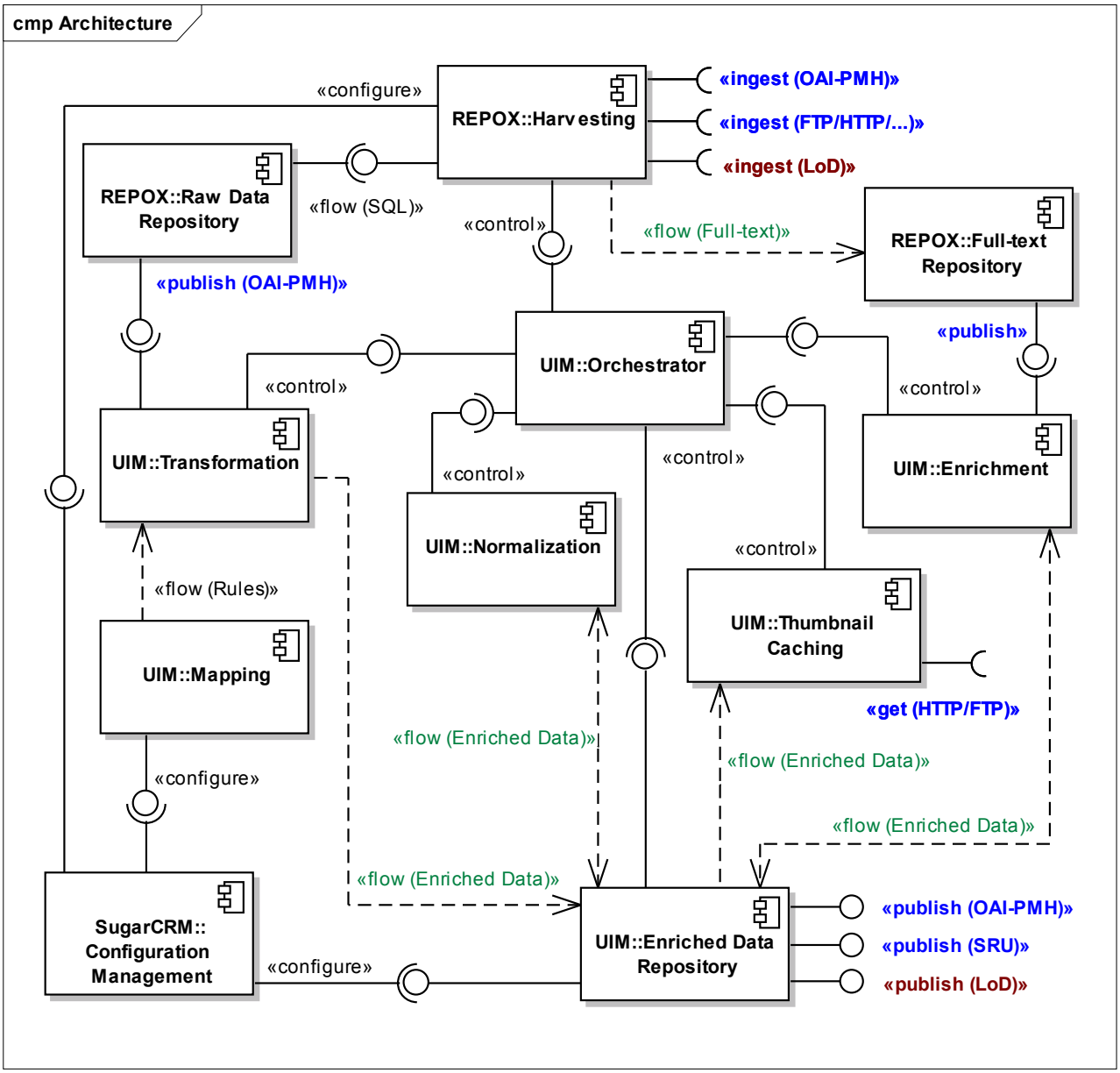

Figure 2: Conceptual architecture of core services of the TEL Aggregator infrastructure

[Req.57] The Aggregation Team needs access to the Enriched Data Repository and is allowed to store, overwrite and delete content (on collection level or on record level).

[Req.58] Content Providers need reading access to the Enriched Store for manual validation purposes. This validation should be embedded to a limited access frontend so that providers can validate the data in a controlled appropriate form.

[Req.59] **SugarCRM** needs access to the **Raw Data repository** to store, copy, overwrite and delete content (on collection level or on record level).

[Open Issue 15] The **Technical Reference** use case can be supported by a regular technological solution for purposes like this, namely a wiki platform.

[Open Issue 16] The **Orchestrator** represents the service responsible for all the coordinated workflow actions in the system.

[Req.60] The **Orchestrator** must assure the task of Harvesting, Enrichment and transfers are executed constantly and automatically during the time slots the **Aggregation Team** has scheduled.

[Req.61] The **Aggregation Team** must be able to fully control the **Orchestrator**, including schedule, cancel and prioritize any step in the workflows it controls.

[Open Issue 17] The actual amount of data currently managed by TEL already makes it necessary to deploy the ingestion services across different distributed physical hardware. An advantage here is that collections are per definition independent from each other and are not influenced by other collections. But it could become a problem concerning contextual resources like authority files. The content in authority files does influence all collections, which refer to the authority file in question. E.g. if an author name is corrected in an authority record, all collections referring to this record need to be updated or marked.

5.1 Interoperability

[Req.62] The APIs must allow the Service Providers to harvest the content automatically and constantly.

[Req.63] The APIs must allow selective harvesting in regard to: datestamps; metadata formats; collections; content provider

[Req.64] The APIs must log harvesting statistics - how many datasets of what collection have been harvested by whom and when harvesting failed.

[Req.65] APIs must make the harvesting statistics available to the KPIANALYSE database.

5.2 Provider Management and Collection Management

[Req.66] SugarCRM must distinguish between the description of the Content Providers (Provider Management) and the description of the collections (Collection Management). It must be possible to interlink a Content Provider record with numerous collection records.

[Req.67] SugarCRM should provide a user interface where the state of registration is visible.

5.2.1 Configuration Management

[Req.68] SugarCRM should compile human-readable versions out of the UIM logs and make these available to Aggregation Team.

6. Consolidated architectural view

The Table 1 shows a consolidated view of the use cases of the TEL Aggregator and the respective supporting services and subsystems and its components.

Use Case and Scenarios	Supporting Subsystem	Subsystem Components
Use Case 1: Data Ingest		
Scenario 1: Metadata Ingest	REPOX	OAI-PMH Client
Scenario 2: Thumbnail Ingest	UIM	Thumbnail Caching
Scenario 3: Full-Text Ingest	REPOX	Full-Text Ingest Client
Use Case 2: Data Export		
Scenario 4: Data Export	UIM	Enriched Data Repository
Scenario 5: Data Export by SRU		
Scenario 6: Data Export as LoD		
Use Case 3: Normalization and Enrichment		
Scenario 7: Data Transformation	UIM	Transformation
Scenario 8: Data Normalization	UIM	Normalization
Scenario 9: Data Enrichment	UIM	Enrichment
Use Case 4: Technical Reference		
Scenario 10: Technical Reference	*(…wiki?, …? …)*	*Reference*
Use Case 5: Manage Data Provider		
Scenario 11: Create a Data Provider Record	SugarCRM	Configuration Management
Scenario 12: Editing a Data Provider Record		
Use Case 6:Manage Data Schema		
Scenario 13: Create a Mapping	UIM	Mapping
Scenario 14: Editing a Mapping		
Use Case 7: Manage Data Ingests		
Scenario 15: Create a Data Ingest Task	SugarCRM *(SugarCRM will control REPOX for this use case, directly supports the Same scenarios)*	Configuration Management
Scenario 16: Edit a Data Ingest Task		
Scenario 17: Manage Data Ingest Tasks		
Scenario 18: Audit a Data Ingest Task		
Scenario 19: Validate Data Ingest Task		
Use Case 8: Service Provider Report		
Scenario 20: Service Provider Report	TEL API	
Use Case 9:Publish Data		
Scenario 21: Creating a Data Export	SugarCRM	Configuration Management
Scenario 22: Manage Data Exports		
Use Case 10: Maintenance		
…maintenance	SugarCRM	Configuration Management

Table 1: Consolidated view of the Use Cases and the respective supporting services and subsystems.

7. Other Non-Functional Requirements

The aim of the project is to build an efficient and sustainable aggregation infrastructure. Therefore it is necessary to define qualitative requirements to specify what efficiency means.

7.1 System scalability

[Req.69] The system must scale up to the requirements expressed in the Table 2.

Entity	Magnitude	Description
Data Provider	1.000 providers the minimum	Nearly 50 National Libraries and 500 Research Libraries and Projects are expected (WP2)
Data Sets	10.000 collections the minimum (equivalent to 1.000 catalogues)	5 to 10 collections per library can be assumed (1 to 5 catalogues per library). The factor for Research Libraries is unknown.
Records	500.000.000	Currently TEL has 60 million records on site and an estimate ranges to a level up to 300 million.
Mapping Rules	unlimited	

Table 2: System scalability requirements.

7.2 Performance Requirements

[Req.70] The aggregation infrastructure needs to support reprocessing of all records (up to 500.000.000) within a month in the maximum.

[Req.71] Major improvements in enrichment might make it necessary that all data in the Raw Data Repository needs to be reprocessed and republished. This leads to a processing speed of harvesting, enrichment and normalization of a minimum average of 200 records/second.

7.3 Security Requirements

[Req.72] The services must be executed by only the authorized actors.

7.4 Reliability Requirements

[Req.73] Under stated conditions the aggregation infrastructure needs to run the processes constantly without any breakdowns. Upgrades and maintenance of the components of the infrastructure must be possible without downtime of the whole infrastructure.

[Req.74] Backups must be able to run automatically and regularly, to ensure that after breakdown content will be restored with no loss of information and at a speed limited only by the hardware performance. Worst-case is the loss of Raw Data or Enriched Data of the last week (worst-case scenario is therefore that the processing work of one week needs to be rescheduled). Besides computational resources such a process would only involve minimal human resources – all the mapping/normalization specifications must be backup on a daily basis.

[Req.75] The Raw Data Repository, as being the reference data, needs to be incrementally backup on a daily basis.

8. 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