DELIVERABLE

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

D4.2 – Validating the library-domain aggregation infrastructure

Authors:

Stefanie Ruehle (CERL)

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
1.0	09/03/2012	Stefanie Rühle	CERL	Initial draft of feature list
1.1	02-03/2012	Stefanie Rühle	CERL	Validation
1.2	06/2012	Stefanie Rühle	CERL	Validation
1.3	26.10.2012	Stefanie Rühle	CERL	Restructuring of Requirements
1.4	11.2012	Stefanie Rühle	CERL	Validation
1.5	04.12.2012	Stefanie Rühle, Andreas Juffinger, Chiara Latronico	CERL, TEL	Second draft of the report
1.6	13.12.2012	Stefanie Rühle	CERL	Open issues closed
1.7	25.01.2012	Sefanie Rühle	CERL	Final version after reviewing

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 validation results of the Europeana Libraries Aggregation Infrastructure as described in Task 4.3.1. Basis for this validation are the requirements defined in [1]. The validation process ran parallel to the extension and revision of the aggregation infrastructure.

1.1 Structure of this document

The document lists the requirements defined in [1] in reference to the services that run the infrastructure. The services are:

- **REPOX:** This service is responsible for the metadata harvesting, the mapping of formats not supported by UIM to one of the supported formats and the validation of the provided metadata according to compliance with the formats supported by UIM.
- **UIM:** This service is responsible for the transformation of metadata. It supports the mapping, normalization and enrichment of data and the validation of the transformation results.
- **Sugar CRM:** This service is responsible for the configuration of REPOX and UIM, the management of the data providers and governs the ingestion and transformation tasks.
- **TEL API:** This service provides a platform where the transformed data will be published.

Beside the requirements that are service dependent, [1] defines requirements that are non-functional and requirements that apply to all services (cross-service requirements).

Not all of the requirements defined in [1] are mandatory for running the aggregation infrastructure. Therefore for the validation process the requirements were prioritized by using

- **1** for the lowest level of necessity means that the requirement is **nice-to-have**
- **2** for the medial level of necessity means that the requirement **should have** been fulfilled
- **3** for the highest level of necessity means that this requirement **must have** been fulfilled.

On the other side the results of the tests are also expressed in numbers that give a fast overview to what degree the requirements are fulfilled. We use the following numbers:

- **0** is the lowest level of fulfilment and means that the requirement is still **open**
- **1** is the next level and is used when the requirement is **done in parts**
- **2** is the highest level and used when the requirement is **done**.

1.2 Definitions

The terms used in this document correspond to the definitions introduced in [1].

2. Sugar CRM

The tasks of the Sugar CRM are:

- The configuration of all services embedded in the infrastructure
- The management and maintenance of data about data providers
- The management and maintenance of data ingest tasks
- The management and maintenance of data export tasks

2.1 Configuration

Req. from [1]	Description of the requirement from [1]	Priority	Done
Req.54	The configuration service must be a combined configuration service for all systems, to ensure a minimum configuration effort by the operators	2	**2**
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	3	**2**
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	3	**2**
Req.68	Sugar CRM should compile human-readable versions out of the UIM logs and make these available to the Aggregation Team	2	**2**

2.2 Manage Data Provider

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
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	3	**1**	Implemented is the creation of the record by the Aggregation Team.
Req.28 a)	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.	3	**2**	
Req.28 b)	If mandatory information is missing, the user must get an error report	2	**2**	
Req.29	The Aggregation Team must get an alert if a new Data Provider Record applies	1	**0**	Because the creation of Data Provider Records is done by the Aggregation

				Team, (see Req. 27) an alert system is not necessary.
Req.30	The system must support the editing of a Data Provider Record (including its removal) by the Aggregation Team	3	**2**	
Req.31	The Aggregation Team must get an alert if a Data Provider Record was changed.	1	**0**	Because the edition of Data Provider Records is done by the Aggregation Team (see Req. 27) an alert system is not necessary.
Req.66	Sugar CRM must distinguish between the description of the Content Provider (Provider Management) and the description of the collections (Collection Management). It must be possible to interlink a Content Provider record with numerous collection records.	2	**2**	
Req.67	Sugar CRM should provide a user interface where the state of registration is visible.	1	**0**	As soon as the registration process starts, the Aggregation Team sends a case number to the provider, who may get all information concerning the status of his collection sending a request to the Aggregation Team referencing to his case number.

Smaller service deficits found during the test

- To enter the ISO 639-3 it is possible to search for the code. But a search with the name doesn't work (e.g. search form German had no result) - **Fixed**
- A dataset/collection can only have one language, but it may be possible that metadata are provided in more than one language. – Currently not necessary in the moment, there is no use case for that.
- The "TEL Discipline" list is not searchable and it is not possible to assign more than one discipline. - **Fixed**
- The ID of the collection, the data provider etc. has to be entered manually. A problem in this context is that Sugar CRM allows the use of the same identifier for more than one collection. Sugar CRM should give a warning if this happens, because collections, data providers etc. have to be identified in the aggregation infrastructure by a unique identifier (E.g REPOX ignores another collection with the same identifier).- **Fixed**

2.3 Manage Data Ingest

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
Req.35	The system must support the creation of a new Data Ingest Task by the Aggregation Team concerning a specific Data Provider	3	**2**	
Req.36 and Req.38	The system must support the editing (including removal) of a Data Ingest Task by the Aggregation Team, concerning a specific Data Provider.	3	**2**	
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.	3	**2**	
Req.40	The results of an execution of a Data Ingest Task must be classified “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	3	**2**	
Req.43	The classification of “OK” for a Data Ingest Task must be restricted to the Aggregation Team.	3	**2**	
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).	2	**2**	
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.).	2	**2**	

2.4 Data Transformation

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
Req.42	It must be 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 to it.	2	**2**	Report of data changes happens via UIM

3. REPOX

The tasks of REPOX are:

- The harvesting of bibliographic data, authority files and fulltext
- The provision of data transformed, enriched and normalized by Europeana Libraries via OAI-PMH
- The transformation of non-compliant metadata schemes into metadata schemes supported by the UIM

3.1 Metadata Ingest

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
Req.51	The Harvesting service must be responsible by bringing the data on site, harvesting it from the Data Providers.	3	2	
Req.8	The system must be able to harvest Metadata at least via OAI-PMH, FTP, HTTP and Z39.50	3	2	OAI-PMH done
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	2	2	The data harvested by REPOX must be well formed but not valid.
Req.10	The system must be able to harvest Metadata as LoD	2	2	The system is able to harvest compliant EDM Data. Currentliy there is no use case for other data models.
Req.11	The system must harvest the Metadata constantly and automatically during the time slots the Aggregation Team has defined according to the Content Providers announcements.	3	2	
Req.14	The system must be able to harvest Full-text referenced in the Metadata	3	2	The harvesting of Full-text is technically possible. But because of open copyright issues the publishing of the data is not possible in the moment.
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 by the Data Ingest Task.	3	2	
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 "Scenario19: Validate Data Ingest Task".	2	2	
Req.55	REPOX must overwrite updated Metadata	3	2	

	records and Full-text automatically			
Req.59	Aggregation Team needs access to Raw Data repository to store, copy, overwrite and delete content (on collection level or on record level).	3	**2**	
Req.75	The Raw Data Repository, as being the reference data, needs to be incrementally backup on a daily basis.	3	**2**	

Smaller service deficits found during the test:

- The data about the Data Provider in REPOX are not identical with the data entered in Sugar CRM. For Country and Homepage default values are used. If these fields are needed at this service than the data from the Sugar CRM should be provided. - **Fixed**
- The information about the original metadata format and the output format of a collection should be transferred from Sugar CRM to REPOX automatically. - **Fixed**

4. UIM

The tasks of UIM are:

- Transformation (i.e. mapping, normalization and enrichment) of metadata
- Validation of provided metadata according to its compliance with formats supported by UIM
- Harvesting of thumbnails referenced in the metadata
- Validation of links to digital objects desribed by the provided metadata
- Validation of metadata transformed by UIM

4.1 Metadata Ingest

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
Req.12	The system must be able to harvest the Thumbnails when those are referenced in the Metadata	3	**2**	The Thumbnails are harvested via API on the fly and stay in cache after that. But they can also be harvested during the ingestion process if that is technically necessary.
Req.13	The system must harvest the Thumbnails constantly and automatically during the time slots the Content Provider has announced and therefore as it has been defined for the Data Ingest Task	3	**0**	Not necessary because of Req.12

4.2 Manage Data Schema

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
Req.33	The system must support the creation of a new Mapping by the Aggregation Team.	2	**1**	XSLTs are created in a subversion system and maintained there, so that it is possible to retrace the editing history. XSLTs can easily been integrated, changed and removed. The UIM transformation logins refer to the used XLSTs.
Req.34	The system must support the editing of a Mapping (including its removal) by the Aggregation Team.	2	**2**	

4.3 Manage Data Ingest

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
Req.39	The results of a Data Ingest Task must be validated concerning compliance to standard schemas and it must keep record about any detected non-conformity.	2	**1**	Work in progress
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.	3	**2**	

4.4 Data Transformation

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
Req.22 and Req.47	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.	3	**2**	UIM does not support transformation of MODS, so MODS data are transformed to MARC before they are brought into the UIM.
Req.24	The system must be able to apply techniques to produce Enriched Data making use of Contextual Resources (authority files)	2	**2**	Currently the alignment to Geonames, Subject Headings in MADS format and Persons and Organisations from VIAF is realized.
Req.25	The system must be able to apply techniques to produce Enriched Data making use of the harvested Full-Text.	2	**0**	Because of the open copyright issues this requirement wouldn't be realized during the project.
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)	3	**2**	
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.	2	**1**	Work in progress
Req.21 and Req.32	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	1	**2**	
Req.56	UIM must overwrite updated Enriched Data records automatically.	3	**2**	

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).	3	**1**	The Aggregation Team can store, overwrite and delete content on collection level not on record level

5. TEL APIs

The task of the TEL APIs is:

- Provide a platform for the publishing of transformed metadata

5.1 Data Transformation

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
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	2	**2**	Content Provider may validate their data using the TEL API. It's possible to ingest a limited number of data for validation purposes.

5.2 Data Export

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
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.	3	**2**	
Req.17	The system must be able to export Metadata through OIA-PMH.	3	**2**	
Req.18	The system must be able to export Metadata through SRU	2	**1**	Instead of SRU the Aggregation Infrastructure provides Open Search
Req.19	The system must be able to export Metadata through LoD	3	**2**	The Aggregation Infrastructure can provide LoD Data in EDM. There is no use case for other data models.
Req.48	It must be possible to define controlled data exports for only specific Data Providers	2	**2**	The selection of specific data collections is possible via the Collection ID.
Req.53a	The publishing interfaces must make the Data available in all the required variants,	1	**2**	The internal data schema allows it to harvest the transformed and enriched data as

				well as the raw data on the basis of the ID of the data provider, the collection ID and the ID of the original data.
Req.53b	Including Metadata provision to Europeana	3	**2**	
Req.62	The APIs must allow the Service Providers to harvest the content automatically and constantly.	3	**2**	
Req.63	The APIs must allow selective harvesting in regard to: datestamps; metadata formats; collections; content provider.	3	**2**	
Req.64	The APIs must log harvesting statistics – how many datasets of what collection have been harvested by whom and whether harvesting failed.	2	**2**	
Req.65	APIs must make the harvesting statistics available to the KPIANALYSE database.	2	**2**	
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.)	2	**1**	Validation of export schemas is work in progress see Req.20

6. Cross-Service and Non-Functional Requirements

6.1 Global Requirements

Req. from [1]	Description of the requirement from [1]	Priority	Done	Note
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 [2] and [3] and the Legal Requirements as defined in [4]	3	**2**	The requirements of the EDM will be fulfilled till the end of 2012
Req.49	The TEL Aggregator must be designed as an infrastructure making the best use of the existing applications (which when necessary will be extended or complemented to support all the desired requirements)	2	**2**	
Req.50	Each subsystem of the TEL Aggregator system must provide detailed information, to be accessible to the Administrator, of any failure (this requirements is intended to stress to the design and development teams the need of the necessary record keeping features for this purpose).	3	**2**	Log data is captured in Repox and in UIM.
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.	3	**2**	
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.	2	**0**	Only the Aggregation Team is allowed to perform actions in the Aggregation Infrastructure
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 the design and the development teams the need of the necessary record keeping features for this purpose).	3	**2**	
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).	3	**2**	

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, and any action progress must be able to be aborted with no effect to the state of the system verified before that action was started).	3	**2**	
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	2	**1**	Reports are send by push technique via email
Req.26	The system must support a forum for dissemination of information, sharing of knowledge, and for controlled interaction by all its human actors.	1	**2**	The TEL Wiki is the platform where the Aggregation Team publishes information for the content provider in regard to schema requirements, interface requirements etc.

6.2 Non-Functional Requirements

Req. from [1]	**Description of the requirement from [1]**	**Priority**	**Done**	**Note**
Req.69	The system must scale up to • 1.000 providers the minimum • 10.000 collections the minimum • 500.000.000 records • Unlimited number of mapping rules	3	**1**	The limitation of storage base is now 250.000.000 records but can be fixed to 500.000.000 if needed
Req.70	The aggregation infrastructure needs to support reprocessing of all records (up to 500.000.000) within a month in the maximum.	2	**2**	The infrastructure supports the reprocessing of 13.8 million records per day and is able to reprocess about 414 million records within a month. Currently about 106 million records are ingested. So the reprocessing of all records will last about 7.5 days
Req.71	Major improvements in enrichment might make it necessary that all data in the Raw Data Repository needs to be reprocessed and	2	**2**	Minimum average is 180 record/second

	republished. This leads to a processing speed of harvesting, enrichment and normalization of a minimum average of 200 records/second			
Req.72	The services must be executed by only the authorized actors.	3	**2**	Access to the infrastructure is allowed only by user name and password. In addition changes concerning the XSLTs are traceable by person.
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.	3	**2**	In addition the infrastructure has access to two Repox systems so that if one is down the other may carry the harvesting.
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.	3	**2**	

7. Conclusion

Based on the requirements prepared in 2011 and described in D4.1 – Requirements Infrastructure and Harvester (Extended Revised Version) [1], the Aggregation Infrastructure was validated in 2012 in February/March, than in June and at last in November/December along with the enhancement of the Aggregation Infrastructure. Shortcomings and open issues that were found during the first validation were settled and proved in the next validation rounds. During a meeting in The Hague in December 2013 the last open issues were closed and this report was finalized. The result of this report is:

- Most of the requirements of level 3 (i.e. must have) are implemented, only three of them partly (see. Req. 27, 57 and 69), the rest totally. An exception is Req.13, this requirement became obsolete, because the thumbnails will be harvested via API on the flight (see Req.12).
- Most of the requirements of level 2 (i.e. should have) are implemented, six of them partly (see Req. 33, 39, 20, 18, 46 and 7), the rest totally. An exception is Req. 25, this requirement couldn't be implemented till the end of the project because of open copyright issues and Req.3, which became obsolete, because only the Aggregation Team is allowed to perform actions in the Infrastructure.
- Requirements of level 1 (i.e. nice to have) are implemented except the Req. 29, 31 and 67, this requirements became obsolete because the registration of data providers and their collections cannot be done by the data providers themselves but only by the Aggregation Team (see Req. 27).

Thus the requirements from D4.1 were implemented with only 6 exceptions which mainly occurred in cases were requirements became obsolete because they depended on other requirements that were implemented in another way than expected.

8. References

[1] D4.1 – Requirements Infrastructure and Harvester (Extended Revised Version)

[2] Europeana Professional – Technical Requirements - http://pro.europeana.eu/technical-requirements

[3] Europeana Professional – Europeana Data Model (EDM) Documentation - http://pro.europeana.eu/web/guest/edm-documentation

[4] Europeana Professional – Legal Requirements for Providing Data - http://pro.europeana.eu/web/guest/licensing