Grant Agreement 297292

EUROPEANA INSIDE

Content Re-Ingestion Report

Deliverable number	*D4.4*
Dissemination level	*Public*
Delivery date	*May 2014*
Status	*Final*
Author(s)	*Nathalie Poot (KMKG)*

This project is funded under the
ICT Policy Support Programme part of the
Competitiveness and Innovation Framework Programme.

Revision History

Revision	Date	Author	Organisation	Description
v0.1	2014-05-14	Nathalie Poot	KMKG	Draft
v1.0	2014-05-28	Nathalie Poot	KMKG	Final version – review all partners

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

Contents

1 Introduction

1.1 Background and role of the deliverable in the project

This report is part of Work Package 4 (WP4). This Work Package is dedicated to the **coordination of content** to Europeana: more than 960,000 records will be delivered using the Europeana Connection Kit (ECK).

This deliverable reports on the outcome of *Task 4.3 Content Re-Ingestion Pilot*: a number of content providers were **to evaluate the potential for the dynamic re-ingestion of enriched metadata back into their systems.**

The testing of content re-ingestion is part of the evaluation of iteration 3 ECK prototype. This report - *D4.4 Content Re-Ingestion Report* (M26 – May 2014) - focusses only on the test results from content re-ingestion. The results from testing iteration 3 ECK are reported on in *D4.3(v1) Export Evaluation Report* (M26 – May 2014).

This deliverable represents the point of view from the content providers (CPs). It gives an insight into their experiences with the software the technical partners (TPs) developed and released for iteration 3. It should be seen in close relation to *D4.6 Technical Specification* presented by K-INT in M25 (April 2014).

WP4 is further **dependent on the outputs of WP2, WP3 and WP5** for its deliverables. Testing content re-ingestion is part of iteration 3 ECK prototype. This prototype was developed and released under WP5 (production). The previous iterations (iteration 1 in M12 and iteration 2 in M18) were developed and released as part of WP3 (development).

WP4 is further dependent on the **outputs of WP2, WP3 and WP5** for its deliverables. Iteration 3 ECK prototype was developed and released under WP5 (production). The previous iterations (iteration 1 in M12 and iteration 2 in M18) were developed and released as part of WP3 (development). The deliverables for WP4 also build on the previous reports within the work package.

The development of the ECK and the evaluation of iteration 3 are based on:

- *D2.1 Requirement Analysis*: explanation of all ECK requirements, based on a survey among the project partners.
- *D2.2 Use Cases:* three use case scenarios.
- *D2.3 Recommendations for Technical Standards:* research on best practice and quality instruments already in place within the Europeana project family.
- *D2.4 Functional Requirement:* there are three kind of requirements: high level requirements, workflow requirements and non-functional requirements. The workflow requirements are identified as: manage, select, prepare, validate, supply, data acceptance and enrich and return.
- *D2.5 Technical Specification:* describes the overall architecture of the ECK.
- *D3.5 Technical Integration Report:* progress report on the development of the ECK.
- *D4.2 Content Export Schedule:* presents the schedule for content delivery. It specifies the order in which participating institutions carry out the export of their data using the ECK.
- *D4.1(v1) Control Export Evaluation Report:* report on the test results from iteration 1 ECK prototype.
- *D4.1(v2) Control Export Evaluation Report:* report on the test results from iteration 2 ECK prototype.
- *D4.3 (v1) Export Evaluation Report:* report on the test results from iteration 3 ECK prototype.
- *D4.6 (v5) Technical Specification*: report on the technical specifications of the ECK.

- *D4.5 (v1) Summative Evaluation Report:* a summative evaluation of the content delivery process to Europeana using the ECK.

D4.4 Content Re-ingestion evaluates the tools that have been developed as part of the iteration 3 for testing of content re-ingestion. **This deliverable reports on the process of content re-ingestion and highlights any issues and recommendations.**

The results presented will be used for:

- *D4.3 (v2) Export Evaluation Report:* reports on the results of testing iteration 4 ECK, the production version.
- *D4.5 (v2) Summative Evaluation Report*: evaluates the outcomes of all export and re-ingestion activity and highlighting key issues for the final technical implementation.
- WP5: their object is to use the lessons learned in WP2, WP3 and WP4 to develop and launch a full production version of the ECK with accompanying support and documentation materials.

1.2 Approach

In preparation of **testing content re-ingestion as part of iteration 3 ECK,** the following approach was used:

1. **Informing content partners** on the content re-ingestion process:

    - **CPs were informed** on the process of testing content re-ingestion and on the new enrich and return functionalities (Appendix I).
    - At the 3rd Networking Event in Athens (M25 – 9th and 10th of April) **meetings in small groups were held** with TPs and CPs from the testing groups on Basecamp.
    - A presentation on the semantic enrichments made by Europeana was given at the Networking Event in Athens by a member from the Europeana task force on **"**Multilingual and Semantic Enrichment Strategy".

2. It was defined which **type of enrichments** would be tested and how the **re-ingestion workflow** would look like.

3. A **test plan** for testing iteration 3 content re-ingestion and **two evaluation forms** were provided to all partners via Basecamp in M24 (March 2014). The deadline to complete the evaluation forms was the 30th of April 2014.

1.3 Structure of the deliverable

This deliverable reports on the **outcome of testing content re-ingestion as part of testing iteration 3 ECK prototype.** The deliverable is structured in the following way:

- An overview of the preparation before testing the iteration 3 ECK prototype
- The results of testing the iteration 3 ECK prototype – content re-ingestion
- Conclusions and next steps
- APPENDIX I: Preparing testing iteration 3 ECK – content re-ingestion: informing content partners.
- APPENDIX II: Content providers survey – iteration 3 on content re-ingestion.

2 Preparing testing iteration 3 ECK: content re-ingestion

2.1 Development of the ECK in 4 iterative phases

The ECK is released in 4 iterative phases. Each of the 4 iterations include specific functionalities as described in *D2.4 Functional requirement* and *D4.6 (v5) Technical Specification*.

This **iterative approach** replaces the more traditional waterfall approach that was originally described in the DoW. One of the main advantages is that new functionality can be given to users sooner, allowing them to find flaws while there is still time to correct them in later iterations.

While the TPs develop and implement the ECK, feedback is needed on the functionalities, bugs, usability and recommendations can be given for improvements. It is the responsibility of the CPs **to test and provide feedback on these different ECK releases.**

Iteration 1 ECK prototype considered all requirements from *D2.4: Functional Requirements* that have been designated as 'Must' haves with the exception of the actual data push and harvest interfaces onto Europeana and other aggregators. This iteration was mainly concerned with **selecting** and **preparing data**. Some other requirements (functional requirements marked as 'Should' or 'Could', High Level Requirements and non-functional requirements) have also been taken into account.

- The results of testing iteration 1 ECK are part of *D4.1(v1) Control Export Evaluation Report* (M16 - July 2013).

Iteration 2 ECK prototype focused on **management overview of status** and **data publication**. The testing was on the functional requirements that have been designated as 'Must' haves and that belong to all workflow steps. This iteration also included requirements that were planned, but not yet operational in iteration 1.

- The results of testing iteration 2 ECK are part of *D4.1(v2) Control Export Evaluation Report* (M20 – November 2013).

Iteration 3 ECK prototype is a refinement of the functionalities tested in the previous iterations and includes two new functionalities: **push or pull** and the **enrich and the return process** from the Europeana portal (content re-ingestion).

- This report focusses on the results of testing content re-ingestion as part of iteration 3 ECK. The overall results of testing iteration 3 ECK are part of *D4.3(v1) Export Evaluation Report* (M26 – May 2014).

2.2 Informing content partners

All CPs were informed on what needed to be tested for content re-ingestion: which type of enrichments will flow back into their CMS? Which functionalities need to be developed? CPs were asked to consult their TP beforehand to discuss how content re-ingestion would look like in their own system. CPs that would not be able to participate in testing content re-ingestion were asked to inform WP4 lead (Appendix I).

In preparation of testing iteration 3 content re-ingestion **meetings in small groups with the TP-CP testing groups** from Basecamp were held at the **3rd Networking Event in Athens** (M25 - April 2014). TPs presented their test plan for iteration 3 and CPs had the opportunity to ask questions on testing content re-ingestion.

In order for CPs to fully comprehend the type of enrichments made by Europeana a presentation on the subject was given at the **3rd Networking Event** (M25 - April 2014) by a member from the Europeana task force on **"**Multilingual and Semantic Enrichment Strategy". Roxanne Wyns (LIBIS – KU Leuven, BE) presented the goal of the task force and the final conclusions from the report[1].

2.3 Content re-ingestion workflow and participants

There are four fields enriched on the Europeana portal. Europeana uses four different vocabularies. Testing content re-ingestion focusses on evaluating the workflow back of those enrichments into the systems of the CPs. It does not entail user-generated-content.

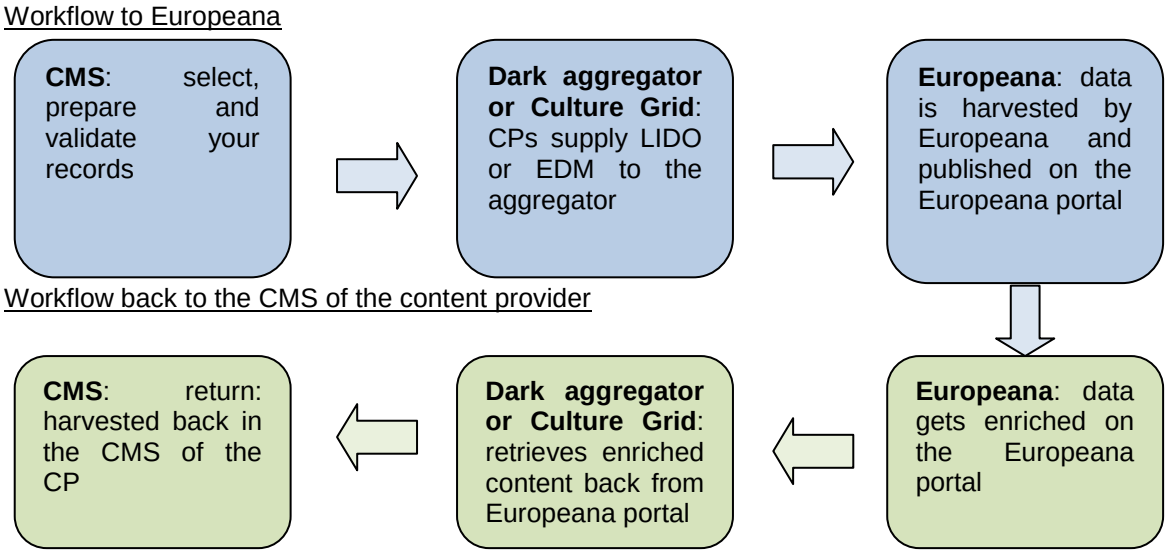
Enriched fields	Used vocabulary by Europeana
Agents (persons) (Creator, Contributor) (dc:creator and dc:contributor)	DBpedia
Places (Geographic data, Coverage) (dcterms:spatial and dc:coverage)	Geonames
Time periods (date, date of creation, time period) Edm_timespan (dc:date, dc:coverage, dc:temporal, edm:year)	Semium Time
Concepts (topics) (Subject) (SKOS_concept (dc:subject and dc:type)	GEMET and DBpedia

Table 1: Overview enriched fields and used vocabularies by Europeana

It was required that the content re-ingestion pilot involved at least 5 content providers and 2 aggregators (DoW). In M22 (January 2014) CPs were asked to consult their technical partner and to inform WP4 lead if they could not participate in testing content re-ingestion. Only two content partners responded that they would not be able to participate.

The two aggregators that make content re-ingestion process possible are the **Inside Dark Aggregator** and **Culture Grid**. The aggregator retrieves the published records from Europeana and generates an enrichment record that can be requested by the CMS[2].

Illustration of the workflow from Europeana to the CMS of the CPs:



[1] The task force ran from October 2013 until March 2014. The final report was published the 7th of April 2014: http://pro.europeana.eu/documents/468623/8b75b054-712e-432b-a0f7-761898e6f60e.
[2] *D4.6 (v5) Technical Specification* (M25 – April 2014): p. 33.

3 Test results

3.1 Evaluation forms

Testing content re-ingestion is part of iteration 3 ECK (released in M24, March 2014, testing and feedback in M25, April 2014). The **test process for testing iteration 3** was similar to testing iterations 1 and 2. An **overall test plan** was provided to all partners via Basecamp in M24 (March 2014). It was stressed that **good communication** and **co-operation** are crucial to make the testing and evaluation process run smoothly.

All CPs were asked to provide **feedback on content re-ingestion** by completing two evaluation forms:

1) Acceptance and Usability test form iteration 3

The evaluation forms lists all required functionalities[3]. CPs needed to indicate 1) whether the functionality is present and working and add remarks and 2) rate the functional requirement: how easy was it to perform the functionality (very easy, easy, difficult, very difficult) and explain why.

The goal was to gather feedback on the content re-ingestion process: how does the enriched data flow back into the system of the CPs and are they happy with it?

2) Content providers survey iteration 3: content-re-ingestion:

Not only the content re-ingestion process was to be evaluated, also the **quality** of the enriched metadata. In close collaboration with Europeana a survey to evaluate the enrichments was made (which fields are enriched, are they satisfied with the enrichments, what is the main advantage of the enrichments,….).

The questions in the survey:

1. Indicate which fields were enriched?
   - Agents (persons) (creator, contributor)
   - Places (geographic data, coverage)
   - Time periods (date, date of creation, time period)
   - Concepts (topics) subject
2. Take a look at the enriched data and provide feedback on the quality of the enrichments. Are they accurate? Explain why (not)?
3. Which from the enriched fields do you consider to be the most useful?
4. How do you plan to re-use the enriched data?
5. What do you consider to be the main advantage of the enriched content?

Both forms needed to be completed and send back before Wednesday the 30th of April 2014.

3.2 Testing content re-ingestion

As described in *D4.6 (v5) Technical Specification* the development for content re-ingestion was completed for iteration 3 by the technical partners. However the content re-ingestion process could NOT be tested, since a change on Europeana's API was needed. Without those changes the enrichment return process does not work.

[3] *D4.6 (v5) Technical Specification* (K-INT): Appendix 2

CPs were therefore not able test the **enrich and return functionalities** within this iteration. They will be tested as part of **iteration 4** (release in M27 – June 2014, testing and feedback in M28 – July 2014).

CPs that had already published content on Europeana within the EUInside project were however asked to complete the survey on content re-ingestion by evaluating the quality of the enrichments on the Europeana portal. This gave CPs a chance partners to learn more about the enrichments before they flow back into their system.

The survey was completed by 4 CPs: National Gallery-Alexandros Soutzos Museum - NAG (GR), Municipio do Seixal - SEI (PT), Stiftelsen Lansmuseet Vasternorrland - SLV (SE) and Royal Museums of Art and History - KMKG (BE).

Overview of the fields that were enriched and the amount of records with enriched fields:

Enriched fields	Used vocabulary	NAG (2859 published records)	SEI (12,149 published records-	SLV (34,143 published records)	KMKG (14,617 published records)
Agents	**DBpedia**	Yes; 5 records	Yes; 1 record	Yes; 34,143 records (enrichments by SLV)	Yes; 48 records
Places	**Geonames**	No enrichments	Yes; 4,598 records	Yes; 34,143 records (enrichments by SLV)	Yes; 6,459 records
Time periods	**Semium Time**	No enrichments	Yes; 3,455 records	No enrichments	Yes; 2,970 records
Concepts	**GEMET and DBpedia**	No enrichments	Yes; 6,787 records	Yes; 34,143 records	Yes; 3,053 records

Table 2: Overview enriched fields

CPs were **overall positive** about the enrichments of **Agents (DBpedia)** and **Places (Geonames)**. The enrichments were accurate by DBpedia and it was mentioned that from the verified records with Geonames, the enrichments entailed added values that were important to the record context[4].

Enrichments in **Time Periods (Semium)** were not always correct. For example, the enrichment took in account dates that are not the most important: on a record that as the following dates: 1950-1970; 1950-2000; 2001; being the first one the creation date, the enrichment only considered the last date presented.

Enrichments in **Concepts (GEMET and DBpedia)** gave mixed results. For several records, when DBpedia was used, the enrichment was correct (e.g 'Tapestry' was correctly enriched)[5] Enrichments by GEMET weren't always accurate (e.g 'Architectural plan' was wrongly enriched by GEMET with "A scheme of action, a method of proceeding thought out in advance"[6]).

[4] Example of an accurate enrichment by DBpedia: http://www.europeana.eu/portal/record/2032006/RMAH_147977_FR.html?start=1&query=europeana_collectionName%3A+2032006*+AND+edm_agent%3A*&startPage=1&rows=24.

[5] http://www.europeana.eu/portal/record/2032006/RMAH_172508_FR.html?start=30&query=europeana_collectionName%3A+2032006*+AND+edm_agent%3A*&startPage=25&rows=24;

[6] http://www.europeana.eu/portal/record/2032006/RMAH_147094_FR.html?start=1&query=europeana_collectionName%3A+2032006*+AND+edm_place%3A*&startPage=1&rows=24

Most of the enrichments from Stiftelsen Lansmuseet Vasternorrland (SLV) were not done by Europeana. They delivered enriched content themselves.

CPs concluded that the added value of the enrichments is the overcoming of language barriers. The use of enrichment gives the possibility to search subjects in other languages thus enhancing the visibility the records.

Conclusions and next steps

Unfortunately the **content re-ingestion workflow** could not be tested as part of iteration 3 ECK. However, since the development was completed by the technical partners, content partners will be able to test the **enrich and return functionalities** as part of **iteration 4** (release in M27 – June 2014, testing and feedback in M28 – July 2014).

Work on preparing for testing content re-ingestion was completed. CPs were fully informed on the enrichments made by Europeana and how they will flow back into their system:

- Separate TP-CP meetings were held at the 3rd Networking Event in Athens.
- A presentation was given on the semantic enrichments made by Europeana.
- CPs that had already published content for EUInside on Europeana were asked to take a detailed look at the enrichments on the Europeana portal and to evaluate the quality of the data.

APPENDIX I – Preparing testing iteration 3 ECK - content re-ingestion: informing content partners

Dear content partners,

For the upcoming testing phase in **April 2014** WP4 will - as part as *Task 4.3 Content Re-Ingestion Pilot* - test and evaluate the potential for the re-ingestion of enriched metadata. The ability of the Europeana Connection Kit (ECK) to re-ingest enriched metadata from Europeana back into the (dark) aggregator and ultimately back into your system will be assessed.

This re-ingestion process is part of **iteration 3** (to be released in March 2014 (WP5) and tested and evaluated in April 2014 (WP4).

Within the course of the EUInside project, the possibility of the re-ingestion of enriched data was discussed and described in previous deliverables (WP2) (you can consult the reports attached or on the EUInside website):

- D2.1 Requirement Analysis (p. 36-37)
- D2.2 Use cases (p. 17)
- D2.4 Functional requirement (p. 18)

Why content re-ingestion?

The idea of enriched content re-ingestion is to give you, as content partner, the possibility to incorporate enriched data from Europeana either inside or outside your own CMS for the purpose of re-use (e.g. to publish on your own website).

Basic workflow re-ingestion

1. Select a sample package of submitted records for re-ingestion to Europeana and make them available to the (dark) aggregator [content partner].
2. The records are published on Europeana via the (dark) aggregator and automatically enriched by Europeana [Europeana].
3. The enriched records come back into the (dark) aggregator [technical partner].
4. From the (dark) aggregator the data will flow back into the system of the content partner [technical partner to prepare the system to re-integrate the enhanced records].

After testing this pilot re-ingestion an evaluation report will highlight issues and will formulate recommendations (*D4.4 Content re-ingestion report* - May 2014).

Which requirements must be present?

10 requirements were defined (*D2.4 Functional requirement*). The content re-ingestion process will be evaluated by the content partners upon these requirements.

1. *Requirement*: **Available enriched content alert** (WFR.07.01)
Explanation: The system reports on available enriched content
Priority: Could

2.*Requirement*: **Acceptance or declining of enrichments on record level** (WFR.07.02)

Explanation: The system allows CP to accept or decline the enriched data (entire records)
Priority: Should

3.*Requirement*: **Automatic ingest of enriched data** (WFR.07.03)
Explanation: Enriched data is ingested automatically in the CPs system after approval by the CP
Priority: Could

4. *Requirement*: **Separate enriched data** (WFR.07.04)
Explanation: The system allows separation based on the origin of the metadata (e.g. original, enrichment, human, machine, user, expert)
Priority: Could

5. *Requirement*: **Enriched IPR identification** (WFR.07.05)
Explanation: The system provides insight in the additional IPR and, for user-generated content, privacy issues regarding the data from external origin
Priority: Could

6. *Requirement*: **Choose target ingest** (WFR.07.06)
Explanation: The system allows return data to be ingested in the system of choice by the CP
Priority: Could

7. *Requirement*: **Acceptance or declining of enrichments on field level** (WFR.07.07)
Explanation: The CP can either accept or decline the enriched data (on field level)
Priority: Could

8. *Requirement*: **Persistent ID's enrichment** (WFR.07.08)
Explanation: The URIs or PIDs enhanced by the system are sent back to the content provider (ref.: *WFR.03.26. Apply PIDs*)
Priority: Should

9. *Requirement*: **Pull option** (WFR.07.09)
Explanation: The ECK contains a pull option, at the request of the data provider:

- Immediate, delayed or according to a preset schedule;
- Full or filtered: e.g. related to a specific object or group of objects.

Priority: Could

10. *Requirement*: **Enriched data management** (WFR.07.10)
Explanation: The system provides management information on which returned enriched data sets are ingested in the CPs system
Priority: Could

Definition priorities:

- *Should*: A high-priority that should be included in the system if it is possible. If not, some explanation is required.
- *Could*: Considered desirable but not necessary. This will be included if time and resources permit.

What type of enrichments?

The records will automatically be enriched with:

- Location/coverage (geonames)

- Subject (topic) (Gemet Thesaurus)
- Period
- Agent

!Action required:

- The content re-ingestion module is at the moment being development by the technical team. **Please consult your technical partner to discuss the content re-ingestion process** (e.g. questions: How will the distinction be made between the original content in your CMS and the enriched content? Will the enriched content flow back into your CMS or into another system? Will you have the possibility to browse into your enriched metadata without having to implement them into your own CMS?,…).
- Testing content re-ingestion will give you the unique possibility to evaluate the **process and value of content re-ingestion** (e.g. Are you satisfied with the process and the enriched content? How do you plan to re-use the enriched data?,…). If you are not able to participate in testing re-ingestion please provide us an argumentation **before Monday the 17th of February** (n.poot@kmkg-mrah.be).

APPENDIX II – Content providers survey iteration 3 on content re-ingestion

Evaluation content re-ingestion: quality of the enrichments (iteration 3)

Name content provider:	
Author name:	

Please provide feedback on the enriched content: which fields were enriched and are you satisfied with the quality of the enrichments? Where do you feel is room for improvement?

6. **Indicate in the right column which fields were enriched?**

Enriched fields	**Used vocabulary**	**Yes OR No and How many records were enriched?**
Agents (persons) (Creator, Contributor) 'dc:creator' and 'dc:contributor'	DBpedia	
Places (Geographic data, Coverage) Enriched fields are 'dcterms:spatial' and 'dc:coverage'	Geonames	
Time periods (date, date of creation, time period) Edm_timespan Enriched fields: dc:date, dc:coverage, dc:temporal, edm:year	Semium Time	
Concepts (topics) (Subject) SKOS_concept (enriched fields are 'dc:subject' and 'dc:type')	GEMET and DBpedia	*(e.g. Yes, 875 records were enriched)*

7. **Take a look at the enriched data and provide feedback on the quality of the enrichments. Are they accurate? Explain why (not)?**

Enriched fields	**Used vocabulary**	**Noteworthy success** **Why?**	**Noteworthy failure** **Why?**
Agents (persons)	DBpedia		

Places	Geonames		*(e.g. the indicated place refers to a country while a city is intended)*
Time periods	Semium Time		
Concepts (topics) (Subject)	GEMET		

8. **Which from the enriched fields do you consider to be the most useful?**

9. **How do you plan to re-use the enriched data?**

10. **What do you consider to be the main advantage of the enriched content?**