ECP-2007-DILI-517009

Europeana Local

Deliverable 3.3: Ingestion process and live content

Deliverable number	*D-3.3*
Dissemination level	*PP; Restricted to other programme participants (incl. Commission services and project reviewers)*
Delivery date	*31 May 2010*
Status	*Final*
Author(s)	*Lizzy Komen (Europeana), Valentine Charles (Europeana)*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

Contents

1 Introduction

EuropeanaLocal has the task of making available digital content from local and regional institutions through Europeana. To make this process manageable this involves the establishment of a harvestable network of OAI-PMH compliant metadata repositories to aggregate this content.

In summer 2010 the operational service of Europeana will be released, with the aim of ten million digital items accessible online. EuropeanaLocal plays an important role in reaching this aim by providing a large amount of content via OAI-PMH and testing data in the Europeana Content Checker, a test environment provided by the Europeana Foundation.

EuropeanaLocal was the first of the Europeana Group of projects to make use of this test environment, as an integral part of the delivery of data to Europeana. The *Central parallel test execution environment,* as referred to in the original title for this deliverable, is now known as the Europeana Content Checker.

2 Scope

This report relates to both *Deliverable 3.1 Central parallel test execution environment: demonstrator for first review* and *Deliverable 3.2: Technical report on initial harvesting results*. Where D3.1 described the EuropeanaLocal test environment set up, D3.2 allowed further improvements of the test environment based on initial harvesting results.

Deliverable 3.3: EuropeanaLocal –ingestion process and live content describes the procedure for establishing the Europeana Content Checker environment. It also shows the use of the Europeana Content Checker by EuropeanaLocal partners and how it plays an important part in the EuropeanaLocal ingestion workflow. This report also shows the successful ingestion of EuropeanaLocal data in Europeana.

At the beginning of the EuropeanaLocal project it was foreseen that the *Central parallel test execution environment* would be the area to test and store the data from the EuropeanaLocal project partners. Only by the end of the project all the data in the test environment would become available via Europeana at once. While time and technology have moved forward, this is no longer the case. Instead, an efficient EuropeanaLocal ingestion workflow has been established, in which the Europeana Content Checker plays an important role. This ingestion workflow allows continuous data integration into Europeana, rather than publishing all EuropeanaLocal data by the end of the project. This means that the 'significant indexed content' in the test environment as originally planned for D3.3, has been taken over by significantly ingested content in the live Europeana portal. Therefore, the original title of D3.3 has been modified to better reflect the current situation.

Because the integration workflow has been improved over time it was also no longer necessary to establish a registry of quality assured local/regional OAI-PMH repositories (D3.2). The harvesting of the OAI-PMH repositories takes place directly between Europeana and the existing EuropeanaLocal repositories, which have been set up within the WP4 Implementation phase.

The technical and operational requirements for joining Europeana and delivering content are described in more detail in *D5.4 – Template of conditions for agreement to participate in Europeana*.

3 Europeana Content Checker

3.1 *Europeana Content Checker infrastructure*

The Content Checker is a test and validation environment for content only. Its infrastructure is a Europeana's product. It comprises of:

- A Java Web-application
- A cross-platform
- An open source database

The Content Checker's infrastructure is flexible and works as a mirror reflecting the evolution of the production portal. New functionalities should be added both in production and test portal, this will become obvious especially after Rhine release.

3.2 *Use of the Content Checker by EuropeanaLocal partners*

In order to make content available to Europeana, Europeana Local partners make use of the Content Checker environment.

The Content Checker is a web-based tool allowing the Europeana Local aggregators to test and validate the content they delivered to Europeana. This tool provides two interfaces: a dashboard called "Ingestor"[1] where partners are importing their data and a "Portal"[2] interface, a copy of the original Europeana website (see Image 2).

Partners have first to upload their ESE XML files into the dashboard part (see Image 1), where files are checked and validated against the Europeana Semantic Elements Schema (ESE)[3]. If a file is not valid, they have to do the appropriate changes. After validation, data are indexed into a database in order to be searchable in the test portal, and images are cached.

[1] http://contentchecker.isti.cnr.it:8080/ingestor/

[2] http://contentchecker.isti.cnr.it:8080/portal/

[3] Europeana Semantic Elements specifications V3.2.2: http://version1.europeana.eu/web/guest/technical-requirements/

Images and data are stored into a database which is cleaned up every month. This tool is able to handle large xml files and multiple sets of data.

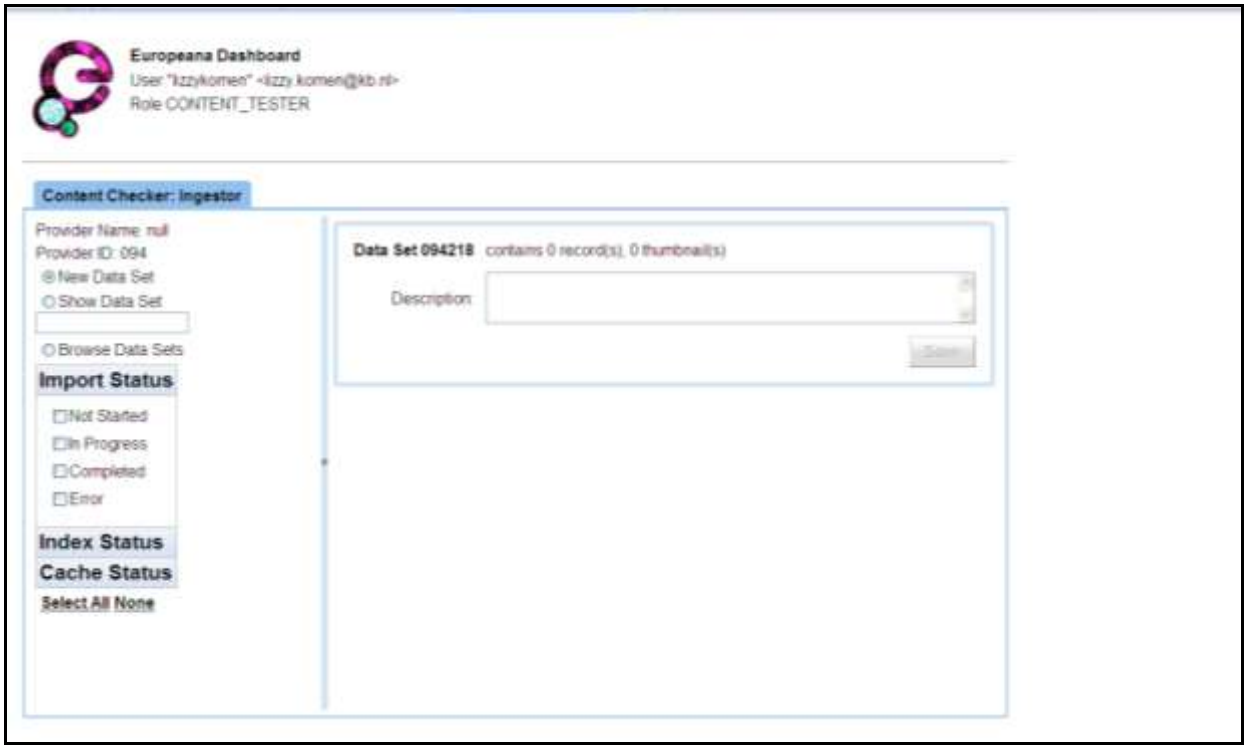

Image 1: Europeana Dashboard

Then partners have to go in the "Portal" part to search for their items and check their records in the Europeana test environment. What they see in this interface is what they will get once data will be uploaded in the production environment. At this stage partners can checked their data according to the quality checks list which is available within the Content Checker Guide[1]. It's important to check such aspects like the display of thumbnails, links to the objects and character encoding issues.

[1] Content Checker Guide : http://version1.europeana.eu/web/europeana-project/technicaldocuments/

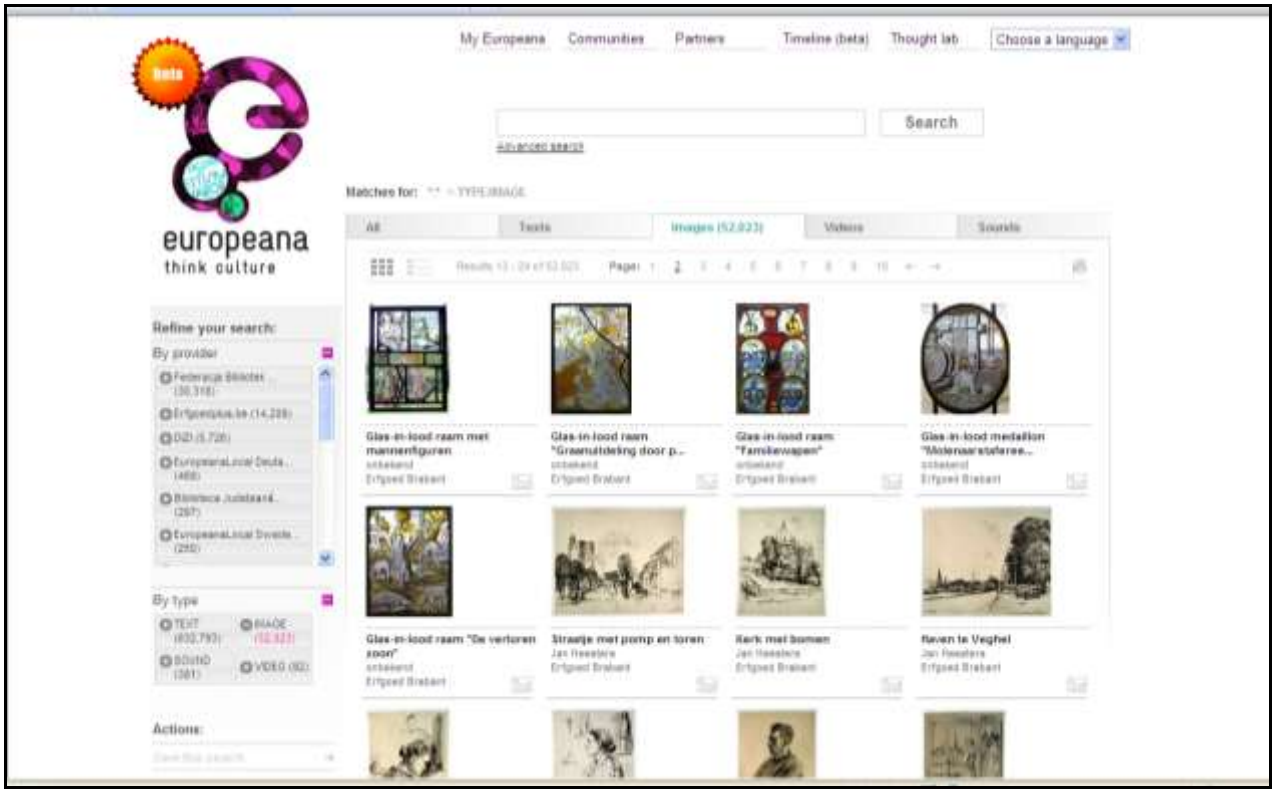

Image 2: Content Checker Portal, with uploaded test content

3.3 *The Content Checker in the ingestion workflow*

The Content Checker is an important step in the ingestion workflow especially in an aggregation model. It provides to aggregators and partners a working environment to test and approve data. Any mistakes in mapping or in the display are visible and aggregator can easily go back to their partners. It also improves the visibility of the aggregators work, as data in process are also visible by the Europeana office. The Content Checker is a first point of communication between Europeana Local and Europeana office as aggregators need to request first an access to the tool and then sometimes require advises related to the quality of the data (Annexe C).

4 EuropeanaLocal content in Europeana

The ingestion workflow as set up between the EuropeanaLocal partners and Europeana has been proven to be successful and work efficiently. See also Annexe B to this document, with a Case Study of the Polish contributor. The EuropeanaLocal partners have already provided an impressive amount of data to Europeana by July 2010, of which the results are shown in Annexe A. The project's contribution to the Europeana Rhine release in summer 2010 is therefore considerable, with some 3,252,044 items available through Europeana.eu. The remaining content from the EuropeanaLocal partners will be ingested according to the ingestion planning for both Rhine and Danube releases. Already, a better representation of local and regional content is currently available via the portal as a result of the EuropeanaLocal project.

ANNEXE A - Ingested EuropeanaLocal content in Europeana by July 2010

July 2010

Collections ID	Institutions	Indexed per collection	Indexed for the project	Country
9401	Bibliotheca Judeteana Octavian Goga Cluj	0		Romania
9402	dLib.si - Digital Library of Slovenia	35,150		Slovenia
9402	Kamra	1,648		Slovenia
9402	Spletna galerija, Semantika d.o.o.	1,517		Slovenia
9403	Angewandte Informationstechnik Forschungsgese (EuropeanaLocal Austria)	6,844		Austria
9404	Federacja Bibliotek Cyfrowych Poland	356,314		Poland
9405	Collections Trust (CultureGrid)	582,026		UK
9406	City of Helsinki	97		Finland
9407	Ministry of Culture	1,234,252		Spain
9408	Bekes Megyei Tudashaz es Konyvtar	1,011		Hungary
9409	Fundacao Museu Nacional Ferroviario	0		Portugal
9410	ABM Resurs/Stiftelsen Länsmuseet Västernorrland (EuropeanaLocal Sweden)	81,752		Sweden

9411	Veria Central Public Library	91,654		Greece
9412	Erfgoed Brabant	111,664		Netherlands
9414	ABM-Utvikling	636,068		Norway
9417	An Chomhairle Leabharlanna	0		Ireland
9418	Conseil Général de la Gironde	22,679		France
9419	DIZI UAB	7,731		Lithuania
9420	Roskilde Kommune	0		Denmark
9421	Provincie Limburg, BE	0		Belgium
9422	Province Oost-Vlandern	0		Belgium
9424	National Library of Latvia	0		Latvia
9425	Eesti Rahva Muuseum (Estonian National Museum),	0		Estonia
9426	AcrossLimits Technologies Ltd	0		Malta
9427	Cyprus Research and Educational Foundation	0		Cyprus
9428	Zentral- und Landesbibliothek Berlin (EuropeanaLocal Germany)	47,449		Germany
9429	Slovenské národné múzeum (Slovak National Museum	21,748		Slovakia
9430	Regione Marche	0		Italy
9431	Public Library 'Pencho Slaveykov'	12,016		Bulgaria
9432	Moravská zemská knihovna	424		Czech Republic

9433	Moravské zemské muzeum/ClTeM	0		Czech Republic
9434	Moravská galerie v Brně	0		Czech Republic
		3,252,044	3,252,044	
		Indexed per collection	Indexed for the project	

ANNEXE B - Case study: Polish Digital Libraries Federation

Marcin Werla, Poznań Supercomputing and Networking Center (PSNC)

PSNC, in which Marcin is the leader of the Digital Libraries Team http://dl.psnc.pl/, acts as the hub of the Polish Digital Libraries Federation, aggregating metadata from Poland's regional and institutional digital libraries. They also act as the national co-ordinator for Europeana Local, and late last year became the first Europeana Local co-ordinator to deliver metadata to Europeana's central index.

Marcin describes the steps towards integrating the Polish content.

'The first step towards Europeana was the analysis of the metadata that we aggregate to see how consistent it was with the Europeana Semantic Elements [ESE] and its mapping guidelines.

'Our metadata was Dublin Core simple which mapped to ESE with some normalization (e.g. we used ISO 639-2 for the DC language element).
Another issue was fields specific in the ESE like 'Europeana type', referring to the nature of the content - text, image, video and audio.

'It was important to establish the standard that would map precisely to ESE across all Polish institutions that contribute data to the Digital Libraries Federation. We gave several presentations to our providers to advise them how to clean and augment their metadata. A group was set up to develop the new metadata schema, and some of our 45 contributing digital libraries had to do some minor modifications to their Dublin Core records.

'Cleaning up the metadata was complicated, but once done, allowed for automated transfer. We manually prepared the mappings for 'type' rather than expect our providers to do it. The remaining mappings from Dublin Core plus extras to ESE were prepared and we ran the modified data on the system each night to update the information from our content providers.

We then ran the data on the Europeana Content Checker. We exported XML files from our OAI-PMH interface and uploaded these to the Content Checker. Throughout the process we worked closely with Lizzy Komen, Europeana Local's liaison officer in the Europeana office.

The Content Checker shows us how records are displaying in a test version of the Europeana interface. We were able to share this display and get feedback from our providers. Some were concerned about how their multilingual records were showing; others wanted to check the display of information about the rights in the objects; yet others wanted to be sure that their name would be clearly visible in the metadata.

The naming issue is important in an aggregation hierarchy. The Polish Digital Libraries Federation aggregates content from around 50 regional and institutional digital libraries, representing in all some 300 individual content holders. It's important that the original content holders as well as the data providers are correctly identified.

We gave ourselves a month for checking. We had 2 people who spent 2 weeks uploading and modifying the code, then 2 weeks getting feedback from all our source libraries. If the display wasn't right we modified the code then repeated the process several times until the data display suited all parties.

The next step was for the Europeana office to test our OAI-PMH interface so they could harvest the records. On the 24th November we completed our work in the Content Checker; on 27 November we had confirmation from Europeana that they were ready to start downloading data from the interface.

Then it was out of our hands: Europeana harvested the 257,000 records and completed the internal processing. This involved normalising the records and indexing them (caching the thumbnails was not possible at this time in our case). Then the Europeana office let us know as soon as they were ready to go live, which they did on 11 December 2009.

Annexe C - Ingestion Workflow