DELIVERABLE

Project Acronym:	**Europeana Cloud**
Grant Agreement number:	**325091**
Project Title:	**Europeana Cloud: Unlocking Europe's Research via The Cloud**

D2.7 - Migration and Upload of Metadata and Content

Authors:

Nuno Freire (EF)
Ola Nowak (PSNC)
Marcin Helinski (PSNC)

Revision: 1

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

Dissemination Level		
P	**Public**	**P**
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Version	*Status*	*Author*	*Partner*	*Date*	*Changes*
0.1	*First Version*	*Nuno Freire*	*EF*	*08/04/2016*	
0.2	*Additions*	*Aleksandra Nowak, Marcin Werla, Marcin Helinski*	*PSNC*	*11/04/2016*	*Addition of details on PCNS migration*
0.2	*Revision*	*Carlo Meghini*	*CNR-ISTI*		
1.	*Publish Version*			*13/04/2016*	

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

Abstract

This deliverable describes Europeana Cloud Services, the new suite of metadata and digital content services, developed by the Europeana Cloud project. It describes how the vision of the product was shaped, its scope (by the end of the project), user experience and some future plans.

Content

Introduction

This document describes the requirements and procedure for exporting of metadata records from Europeana (EF), The European Library's (TEL), and PSNC, to the Europeana Cloud Infrastructure.

Technology and Background

The export procedure for the first two partners is based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard. Both repositories are accessible through the OAI-PMH protocol, which allows for application of a common solution to both datasets. It is also expected that the software developed for this export, will be extended in the future into a full-fledged OAI-PMH harvesting service of Europeana Cloud.

For migrating data from PSNC, the REST API of Clepsydra, PSNC aggregator system, was used.

Access to the datasets

Europeana

The OAI-PMH service of Europeana is openly accessible and Europeana's complete dataset may be harvested using the following OAI-PMH configuration:

- **Base URL**: http://oai.europeana.eu/oaicat/OAIHandler
- **Metadata prefix**: edm
- **Set spec**:<empty>

Additional information about Europeana's OAI-PMH service is available on Europeana Labs[1] and on the documentation page[2] of the service. The service contains all the metadata records available on the Europeana Collections website, slightly above 50 million records in the time of writing.

The European Library

The OAI-PMH service of the European Library can only be accessed from authorized servers. For setting up the access for a OAI-PMH harvester, the team of the European Library should be contacted.

To harvest the complete dataset, the following OAI-PMH configuration and procedure should be applied. The OAI-PMH service of the European Library does not support the harvesting of the complete dataset with the specification of an empty setSpec OAI-PMH parameter. The approach to follow is then to send the initial *ListSets* request to obtain the list of the available sets, and then harvest each set separately.

- **Base URL**: http://ingest2.tel.ulcc.ac.uk:8181/oaipmh
- **Metadata prefix**: edm
- **Set spec**: <list of set should be obtained from a ListSets request>

The service exposes approximately 160 million metadata records.

PSNC

Digital Libraries Federation[3] (DLF), the aggregator ran by PSNC uses Clepsydra aggregation system to manage aggregating, storing and processing of data. For data migration procedure the REST API of storage component, Clepsydra Storage, was used. More information about Clepsydra as well as its REST API documentation[4] can be found on its documentation pages[5].

For every record available in the DLF portal there are multiple representations stored in Clepsydra: the source object downloaded from data providers, the intermediate representations produced during processing and the final representation used for presentation on the DLF portal. Also thumbnails of images are stored.

[1] http://labs.europeana.eu/api/oai-pmh-introduction
[2] http://oai.europeana.eu/oaicat/index.shtml
[3] http://fbc.pionier.net.pl/
[4] http://fbc.pionier.net.pl/pro/clepsydra/storage/clepsydra-storage/rest.html
[5] http://fbc.pionier.net.pl/pro/clepsydra/

During the migration, the following data representations were transferred to Europeana Cloud for each record:

- FBC - the richest format. Used for presentation on DLF portal
- METS - harvested from data providers, transferred when available
- th-scaled - thumbnail in a size compatible with DLF portal, when available

A migration tool using Clepsydra's REST API in specified criteria was developed. The tool also uses Europeana Cloud REST API to upload records into Europeana Cloud. Clepsydra records representing the same digital object were uploaded as subsequent representations of an Europeana Cloud record.

One time batch upload was performed, during which 1,857,032 records were uploaded. All of these records have an FBC representation. A subset of 1,842,123 records have also METS representation and a slightly smaller subset of 1,724,639 records have thumbnails.

An optimised upload procedure is being developed as well. It uses internal Clepsydra mechanism to react on data changes. When a Clepsydra record is inserted, updated or deleted the corresponding operation will be performed in parallel on the corresponding record representation in Europeana Cloud. For each Clepsydra record in FBC schema an XSLT transformation will be performed using Data Processing Service. Migration tool will use Data Processing Service REST API to execute transformation from FBC XML schema to EDM XML schema.

Clepsydra's REST API endpoint can only be accessed from an authorized network. For setting up the access Digital Libraries Federation team should be contacted.

To harvest the DLF dataset the following configuration should be applied.

- **Base URL**: http://met-storage.synat.pcss.pl:8080/clepsydra-storage/rest/records
- **schemaId**: fbc, mets, th-scaled
- **token** (optional, to retrieve following pages of results).

Europeana Cloud Data Model

The metadata records are uploaded to Europeana Cloud through the Europeana Cloud API.This section defines how the data model of Europeana Cloud is applied to the metadata records of Europeana and the European Library.

Data Provider

The datasets should be uploaded to Europeana Cloud under specific DataProvider entries:

- The DataProvider identifier for the Europeana dataset should be defined as follows:

Data Provider Identifier: Europeana_Foundation
Organisation name: Europeana Foundation
Official Address: Prins Willem Alexanderhof 5, 2595 BE,The Hague,Netherlands
Organisation website URL: http://pro.europeana.eu
Digital Library URL: http://europeana.eu
Contact person: Cecile Devarenne <cecile.devarenne@europeana.eu>

- The DataProvider identifier for the TEL dataset should be the same that is to be used for the migration of the Europeana Newspapers Images, which is defined as follows:

Data Provider Identifier: The_European_Library
Organisation name: The European Library
Official Address: Prins Willem Alexanderhof 5, 2595 BE,The Hague,Netherlands
Organisation website URL: http://www.theeuropeanlibrary.org/tel4/aboutus
Digital Library URL: http://www.theeuropeanlibrary.org
Contact person: Nuno Freire <nuno.freire@theeuropeanlibrary.org>

- The DataProvider identifier for the PSNC dataset should be defined as follows:

Data Provider Identifier: psnc_clepsydra_provider
Organisation name: PSNC
Organisation website URL: http://www.man.poznan.pl/
Digital Library URL: http://fbc.pionier.net.pl
Contact person: fbc@man.poznan.pl

Record

Two local identifiers will be associated with the Europeana Cloud Record Identifier.

- OAI-PMH identifier of the metadata record
- The TEL Internal Record Identifier (for the TEL dataset only)

Representation

Each record should contain one EDM representation. The EDM metadata records are stored in XML representation, following the EDM XML schema. This representation is named *edm*.

Version

For the purpose of this export of the datasets, the export will be executed only once. For this reason, each representations will contain only one version. This version is be marked as persistent.

Files

Each *edm* representation will contain the file *metadata.edm.xml*. The Content-Type of all files should according to the RFC7303, which allows two possible values: “”application/xml” or “text/xml”.

Dataset

The data of the European Library will be organised into the following Europeana Cloud datasets:

- TEL Core Dataset - comprises the totality of the data uploaded at this time.
- TEL CC0 Dataset - contains the data that is available under a CC0 license.
- TEL Open Dataset - contains the data that is available under an open Creative Commons license,
- TEL Newspapers - contains the data from Europeana Newspapers

The TEL datasets overlap in terms of the metadata records they contain.

The Europeana dataset will not be organized into subsets. Figure 1 shows the structure and some field values of a sample Europeana Cloud record.

Figure 1: Sample Europeana Cloud Record

Content Migration: Europeana Newspapers

The project Europeana Newspapers stores its digital content (scanned newspaper pages) on a remote location server. The migration tool was implemented to upload the content from the remote server to Europeana Cloud. Altogether, the dataset consisted of more than 5 million JPEG2000 files, summing up to almost 10TB of data.

The migration process was performed in a way that local record identifiers used in Europeana Newspapers could also be used in Europeana Cloud. In this way, all the records added during the migration can be identified both by local identifier and by the cloud identifier. Additionally, all the created records were assigned to the same data provider named *TheEuropeanLibrary*. The

representation used for the images was named *presentation_images*. Each representation version consists of several image files. In order to determine the record identifier for them a specially prepared mapping file was used. That file associates local record identifier with paths to image files on the source storage. File name used for the file in Europeana Cloud contains the whole path which was one of the requirements for the Europeana Newspapers migration.

For example filename can look like this:

node-3/image/NLE/Edasi/1932/11/16/1/19321116_1-0001.jp2.

Migrated image files can also be accessed using the local identifier without knowing the cloud identifier. For example:

https://cloud.europeana.eu/api/data-providers/TheEuropeanLibrary/records/3000118376944/representations/presentation_images/node-3/image/NLE/Edasi/1932/11/16/1/19321116_1-0001.jp2

Future work

The described above work, done under the project Europeana Cloud, can be continued in the future under new projects in the following ways:

- Europeana Cloud
  - Extension of the export procedure into a full-fledge harvesting service for OAI-PMH services
- The European Library
  - Uploading of the original metadata record from the data provider
  - Uploading full-text representation of Europeana Newspaper data
  - Upload of Representation for Linked Data, according to the TEL Linked Data Model
- Europeana
  - Upload of the original metadata record from the data provider
  - Upload of record previews and digital objects
  - Upload of intermediate versions of metadata records
- PSNC
  - Thumbnails presented on FBC portal should be retrieved directly from Europeana Cloud.
  - Integrate all Clepsydra aggregation system components to work with Europeana Cloud. Harvesting data components should be adopted to store their records directly in Europeana Cloud. Data Processing Service should be used as a tool for records transformation.

Summary

This deliverable specified the export of metadata and content from the data repositories of PSNC, Europeana and the European Library to Europeana Cloud. The specification covered the source of the data and the way the data is structured in Europeana Cloud after the export.