
# EnrichEuropeana+ Content Aggregation - First Ingestion into Europeana

Version 1.0

Documentation Information

| | |
|---|---|
| Action Number | 2020-EU-IA-0075 |
| Project Website | https://pro.europeana.eu/project/enricheuropeana |
| Contractual Deadline | 30 November 2021 |
| Nature | Report |
| Author | Marcin Helinski |
| Contributors | Kathryn Cassidy, Padraic Stack, Sonja Galina, Rafał Raczyński |
| Reviewer | Sergiu Gordea |
| Version | 1.0 |
| Date | 30.11.2021 |

**Co-financed by the Connecting Europe Facility of the European Union**

# Contents



# Introduction

EnrichEuropeana+ (fully titled ‘Enriching Europeana through citizen science and artificial intelligence - unlocking the 19th century') aims to enhance Europeana Transcribe (www.transcribathon.eu) as a service for cultural heritage institutions.

# Scope

This document describes the progress in tasks 1.1 and 1.2

# Objectives

The main objectives of EnrichEuropeana+ are:

- To engage public users and professionals in enhancing the semantic and multilingual description of Cultural Heritage objects by continuing the development of Europeana Transcribe.
- To increase accessibility of manuscripts related to historical events and societal transformations in Europe within the 19th Century through a new Citizen Science crowdsourcing campaign to stimulate user engagement for transcribing, translating, and adding semantic enrichments.
- To transform Europeana Transcribe into a service used by Cultural Heritage Institutions to crowdsource the enrichment of cultural object descriptions and improve the multilingualism of metadata.

The main objectives of this Milestone are:

- Ingest first materials to Europeana Collections. Ingested materials should be selected from the list of materials planned for contribution in this project.
- Help content providers to gain knowledge and to experience successful ingestion into Europeana
- If possible ensure that the quality of the ingested materials is at a high level.

# Ingestion process

The aggregation of new materials into Europeana Collections is performed through the national aggregator infrastructures. There are two accredited national aggregators involved directly in the EnrichEuropeana+ project. These are Digital Repository of Ireland (DRI) from Ireland and

Federacja Bibliotek Cyfrowych (FBC) from Poland. Additionally, materials from State Archives Zagreb (SAZ) will be delivered to Europeana through the Croatian National Aggregator with whom the project partners have begun a cooperation to achieve the aggregation goals.

Figure 1 shows the data flow of the ingestion process.

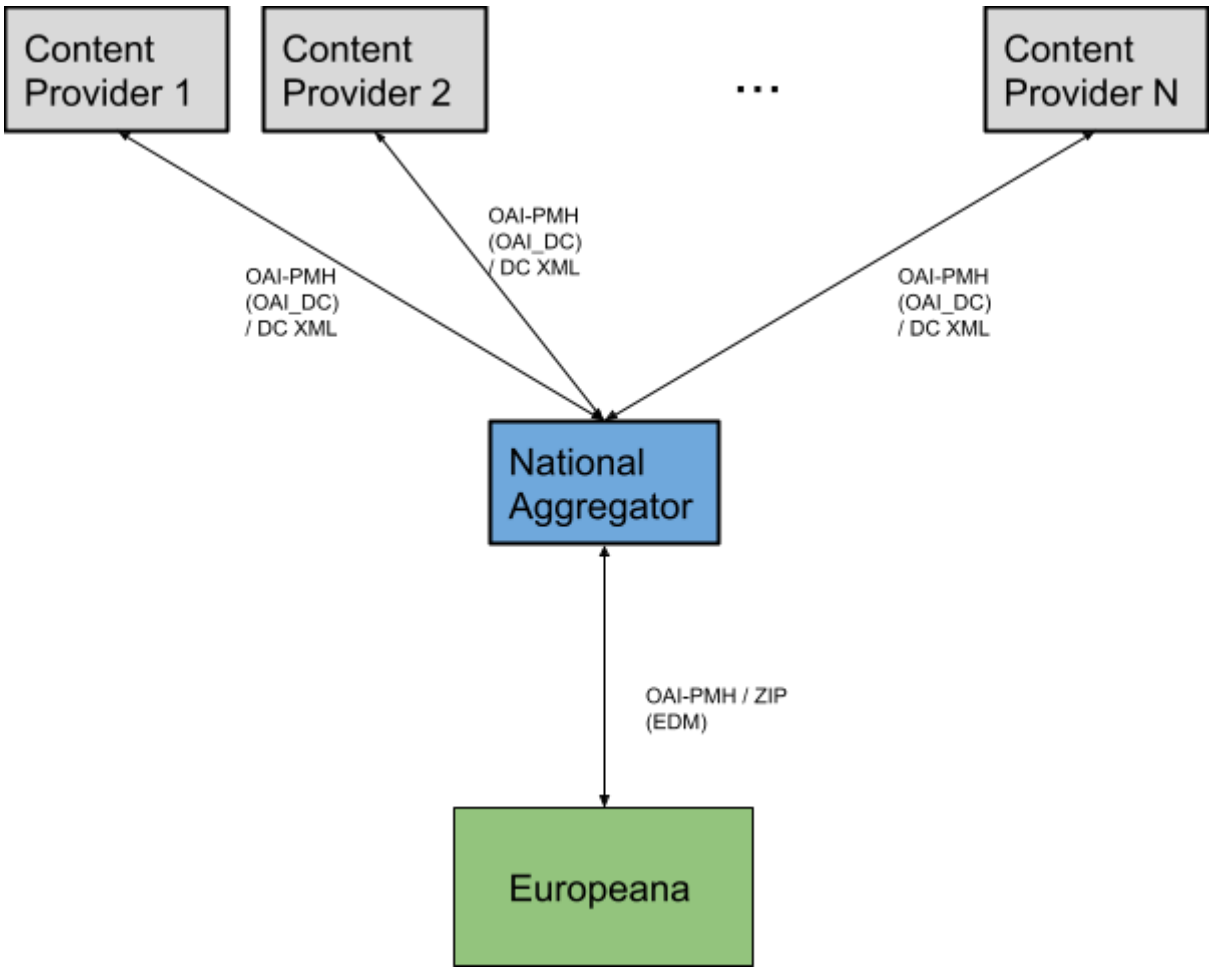


Figure 1. Ingestion process

Typically, records are harvested by the National Aggregator from the Content Providers through the OAI-PMH interface represented using the OAI_DC specifications. The National Aggregator performs aggregation of these records using the OAI-PMH protocol. Alternatively, content providers may send DC XML files directly to the National Aggregator (as is the case with the Digital Repository of Ireland). Once aggregated in the NA platform, the records are usually processed to the aggregator's internal schema. Additionally, if the content provider wants to be ingested to Europeana, the National Aggregator generates a representation according to the Europeana Data Model specifications. These records are delivered to the Europeana aggregation infrastructure by publication via OAI-PMH, or by generating zip archives of smaller data sets containing records using the EDM-XML representation.

On the Europeana side, the processing of the records starts by converting the data into the internal EDM schema. While processing, additional information is attached to individual records and the evaluation of the technical quality of materials is performed through the computation of the content and metadata tiers.

# Quality improvement

As mentioned in the previous section, the materials collected from the data providers are transformed by the aggregators into the EDM schema before ingestion to Europeana. The quality of the original metadata has a great influence on the final metadata tier achievable in Europeana. Therefore, whenever it is possible, national aggregators offer support and advice to data providers in order to help them improve their metadata and content quality. In some cases, certain improvements are also possible on the national aggregator level.

As the result of cooperation between FBC and University of Wrocław metadata of materials from UWr planned for ingestion in EE+ project have been prepared in a way that potentially allows reaching the highest content and metadata tiers in Europeana. On the UWr side metadata has been enriched with links to LOD Vocabularies supported by Europeana (i.e. VIAF, Geonames, etc.). On the side of FBC, to achieve this, the processing workflow has been improved to include the support of IIIF profile in EDM and to generate the embedded representation of contextual entities which contains all mandatory metadata elements.

The Digital Repository of Ireland also provides support for the IIIF profile in EDM, and has created IIIF representations of the content from Dublin City Council. Furthermore, DRI is working with the Library of Trinity College Dublin to ensure that their content and metadata reaches a high Tier level.

# Digitization of new materials

A subset of the materials planned to be contributed by the EnrichEuropeana+ project are newly digitized. The project partners UWr, DCC and SAZ have begun the digitization of new data collections.

On behalf of UWr, Wroclaw University Library has digitized and prepared complete metadata for five collections consisting of 2034 handwritten letters with correspondence of professors: J.D.F. Neigebaur, Siegfried Brie, Alfred Hillebrandt, Otto Lummer and Friedrich Haase and 29 handwritten maps of cities and battlefields dating from the time of the French Revolution.

The State Archive in Wroclaw has delivered 103 scanned objects (accounting books of the city of Wroclaw) with metadata originally exported from ICA-AtoM. The Archive of University of Wroclaw has delivered 178 scanned objects (various handwritten documents and protocols from pre-war University) with metadata also originally exported from ICA-AtoM. UWr has converted these exported metadata to Dublin Core format and enriched them with links to LOD vocabularies.

In summary UWr has prepared 2344 completed digital objects for first ingestion. This accounts for 81.6% of the originally declared number of objects. For the second ingestion, UWr is preparing 500 additional data records.

Dublin City Council (DCC) has delivered 503 items to Europeana as part of the EnrichEuropeana+ project already and a further 600 are present in the DRI (national aggregator) awaiting review and publication before being submitted to Europeana.

Dublin City Council is in the process of digitising and preparing metadata for two additional large collections. These are the Dublin City Council Manuscript Minutes 1841 – 1880, which contain about 16,600 pages in 32 bound volumes of manuscripts and the Minute Books; and Jury Books of the Wide Streets Commission which consist of approximately 21737 pages in 63 bound volumes of manuscripts. Test volumes have successfully been ingested to the DRI (National

aggregator) and it is expected to have the complete collections present in the DRI (National aggregator) and ready for ingestion into Europeana by the end of December 2021.

The Library of Trinity College Dublin has delivered five new collections consisting of 2,480 objects to the DRI. Some small metadata enrichments have been performed on these by the National Aggregator and they are now ready for aggregation to Europeana. These will be submitted in early December.

State Archives in Zagreb is currently digitizing three of the designated collections: Collection of Ivan Ulčnik, Collection of Dragutin Hirc and Records of the city government meetings.

The digitization and documentation for the Collection of Ivan Ulčnik is work in progress. The whole collection contains 1650 items, and a subset of 491 items has been submitted to the Croatian National Aggregator. The Collection of Dragutin Hirc contains more than 1500 records, of which 944 have been already digitized. The metadata associated with the digitized materials is created in parallel. These two collections are planned to be contributed first for aggregation into Europeana.

The third collection, containing the Records of the city government meetings will subsequently be digitized and it will be contributed to be aggregated into Europeana in the next months.

The technical infrastructure of the Croatian National Aggregator is currently under development. This development is carried out within the scope of the project named e-Culture - Digitizing the cultural heritage (e-Kultura - Digitalizacija kulturne baštine)[1]. The Croatian National Aggregator is currently running the first tests to validate the process of aggregating small data collections into Europeana. The first version of the system is expected to be available in the first months of 2022.

## Selection of materials

In order to identify the materials added or updated in Europeana as the result of this project, the decision to tag them has been made. The tag will have the value “EnrichEuropeana” and will be placed in the *dcterms:isPartOf* field of *edm:ProvidedCHO* section of the EDM records. This will allow tracking the ingestion of all contributed content.

Materials intended for ingestion to Europeana have been selected from the list that was provided in the project proposal, as well as a number of relevant collections identified since then. This list contains materials that are ready for aggregation by the national aggregators as well as some new content that will be aggregated in the near future. For the first ingestion the focus has been on materials that are present in the national aggregator services. Those include the following collections:

Table 1: Overview of the data collections aggregated into National Aggregators and Europeana

| **Collection Name** | **Data Provider** | **National Aggregator** | **#records NA** | **#records Europeana** |
|---|---|---|---|---|
| Manuscripts, letters (1780-1920) from Elbląska | Elbląska Biblioteka Cyfrowa | FBC | 25 | 25 |

[1] https://min-kulture.gov.hr/vijesti-8/predstavljen-projekt-e-kultura-digitalizacija-kulturne-bastine-u-muzeju-mimara/19230

| | | | | |
|---|---|---|---|---|
| Biblioteka Cyfrowa | | | | |
| Inventory of correspondence of Johann Daniel Ferdinand Neigebaur | UWr- University of Wroclaw (UoW) | FBC | 254 | 254 |
| Handwritten letters of Siegfried Brie (1853-1931) professor of law, rector of UoW | UWr- University of Wroclaw (UoW) | FBC | 211 | 211 |
| Handwritten letters of Alfred Hillebrandt (1853-1927), researcher of the Sanskrit | UWr- University of Wroclaw (UoW) | FBC | 175 | 175 |
| Handwritten letters of Friedrich Haase (1808-1867), classical philologist, prof of UoW | UWr- University of Wroclaw (UoW) | FBC | 1026 | 1026 |
| Handwritten letters of Otto Lummer (1860-1925), physicist, prof. of UoW | UWr- University of Wroclaw (UoW) | FBC | 368 | 368 |
| Maps collection from Wroclaw University Library | UWr- University of Wroclaw (UoW) | FBC | 28 | 28 |
| Handwritten documents from the Archive of the University of Wroclaw | UWr- University of Wroclaw (UoW) | FBC | 178 | 178 |
| Manuscripts, files of the City of Wroclaw from the State Archive in Wroclaw | UWr- University of Wroclaw (UoW) | FBC | 103 | 103 |
| Akta urzędnika | Regionalia Ziemi Łódzkiej | FBC | 24 | 24 |
| Collection of Ivan Ulčnik, | State Archives in Zagreb | Croatian National | 491 | 0 |

| | | | | |
|---|---|---|---|---|
| 1848./1930. | | Aggregator | | |
| Patrick English collection | Dublin City Council | DRI | 65 | 65 |
| Dublin Reconstruction (Emergency Provisions) Act 1916 | Dublin City Council | DRI | 40 | 40 |
| Dublin Castle Tracts | Oireachtas Library | DRI | 1567 | 1567 |
| Wide Streets Commission (1681-1851) | Dublin City Council | DRI | 276 | 276 |
| Dublin City Surveyors Maps 1695-1827 | Dublin City Council | DRI | 131 | 131 |
| Papers of Major Richard William George Hingston | Trinity College Dublin | DRI | 763 | 0 |
| Papers of Michael Davitt | Trinity College Dublin | DRI | 556 | 0 |
| Papers of John Millington Synge | Trinity College Dublin | DRI | 130 | 0 |
| Oscar Wilde Collection | Trinity College Dublin | DRI | 116 | 0 |
| J.D. White Collection | Trinity College Dublin | DRI | 915 | 0 |
| Parish registers and chronicles from Małopolska Biblioteka Cyfrowa | Małopolska Biblioteka Cyfrowa | FBC | 12 | 0 |
| Łódzka Regionalna Biblioteka Cyfrowa ( medical books) | Łódzka Regionalna Biblioteka Cyfrowa | FBC | 85 | 0 |
| Documents and letters from and to Bolesław Prus from Biblioteka Cyfrowa Wojewódzkiej Biblioteki Publicznej im. Hieronima | Biblioteka Cyfrowa Wojewódzkiej Biblioteki Publicznej im. Hieronima Łopacińskiego w Lublinie | FBC | 184 | 0 |

| | | | | |
|---|---|---|---|---|
| Łopacińskiego w Lublinie | | | | |
| Wide Streets Commission General Maps | Dublin City Council | DRI | 600 | 0 |
| **TOTAL** | **25 Collections** | | **8314** | **4462** |

Annex A to this document contains the summary of the ingested datasets together with a quality breakdown[2].

The table below shows the overall quality breakdown of the records forming part of the first ingestion to Europeana.

Table 2: Overview of the content and metadata quality of the records aggregated into Europeana

| **Content Quality** | **#Tier 4** | **#Tier 3** | **#Tier 2** | **#Tier 1** | **#Tier 0** |
|---|---|---|---|---|---|
| | 3430 | 460 | 41 | 516 | 15 |
| **Metadata Quality** | **#Tier C** | **#Tier B** | **#Tier A** | | |
| | 1339 | 3030 | 93 | | |

Most of the records ingested within the scope of this milestone reached the highest content tier (i.e. ~ 77%). A large part of these records are already accessible in IIIF format. Lower content tiers are usually related to content with limited access or improper referenciation of content which affects the media processing and the content tier computation in Europeana Aggregation infrastructure. The latter problem will be addressed and probably solved through the next ingestion.

The Metadata tier of most of the ingested records is B which is the minimum required tier for this project. As a result of using a variety of LOD vocabularies over 30% of the ingested records reached the highest metadata tier C. There is still a small number of records (i.e. 93) for which the metadata quality was evaluated to tier A. These records are expected to be improved within the metadata enrichment through the Transcribathon events.

# Conclusions

This document presents the work carried out for achieving milestone 3 of EnrichEuropeana+ Action. It describes the ingestion process and the work that has been done to improve the quality of records metadata.

---

2 https://docs.google.com/spreadsheets/d/16FuM46XVq88t2i7EWvjKkyFHU_mr-ZZQTE-u2WQLwC0/edit?usp=sharing

This document also provides an overview of materials that were ingested into Europeana in the first batch. There were 8314 records delivered to the National Aggregator infrastructures and 4462 were already published in the Europeana Portal. The largest part of the materials planned to be aggregated by the action belong to the collections contributed by SAZ, which will be aggregated through the Croatian National Aggregator infrastructure which was still not released to a production environment yet. These materials and other collections contributed by the partners will be aggregated into Europeana and the detailed report will be included in the final technical report of the action.