European Commission
Innovation and Networks Executive Agency (INEA)
CEF Telecom

Europeana Common Culture
MS5 Report on Data Quality Improvement | January 2021

Chardas Hungarian Cuisine, New York
Magyar Kereskedelmi és Vendéglátóipari Múzeum, Budapest | CC BY-NC-ND

Information on the Action	
Grant Agreement Nº	INEA/CEF/ICT/A2018/1633581
Action Title (Art. 1 of G.A.)	Europeana Common Culture
Action Nº (Art. 1 of G.A.)	2018-EU-IA-0015

Milestone contributors	
Author(s)	Dimitra Atsidis, NISV; Jonathan Blok, NISV
Contributor(s)	Nicole Emmenegger, NISV; Fiona Mowat, EF; Common Culture partners
Reviewer(s)	Henning Scholz, EF; Milena Popova, EF; Valentine Charles, EF

Table of Contents

List of abbreviations

Abbreviation	**Description / explanation**
Activity	Activity in the context of the Generic Services projects is a group of related tasks within a project.
CHI(s)	Cultural Heritage Institution(s), includes key actors in the cultural heritage sector, such as libraries, museums, archives or galleries.
DSI	Digital Service Infrastructure
DQP	Data Quality Plan
EDM	Europeana Data Model
EF	Europeana Foundation
EMR	Enrichment Mapping Rules
EPF	Europeana Publishing Framework
ESE	Europeana Semantic Elements
LOD	Linked Open Data
NA(s)	National Aggregator(s), an entity that works with cultural heritage institutions to gather authentic and trustworthy data and make it accessible through Europeana or other dissemination channels. National aggregators define their scope by specific country and they work with contributors situated within that country.
NISV	Netherlands Institute for Sound and Vision

Introduction

This Milestone is a report on the Europeana Common Culture project data quality improvement activity. The goal of this report is to provide a comprehensive picture of the data quality improvements applied to datasets from the National Aggregators (NAs) participating in Tasks 3.1 and 3.2 of Activity 3. This report also addresses the assessment of the semantic enrichment applied to selected datasets as part of Task 3.3. It describes the processes, lessons learned and results of the data quality improvements and the semantic enrichment. It also includes insights to help content holders enhance their data quality by advising them how to establish priorities in their future data aggregation and enrichment activities.

The overall results of the Activity show that the full engagement of cultural heritage institutions is vital for NAs and the Europeana Foundation to continue to improve data quality and achieve not only the project goals but all future quality targets. These points must be part of a broader discussion and policy recommendations on national aggregation policies and national digital transformation strategies across Europe.

Furthermore, this activity succeeded in setting out clear, measurable and sustainable guidelines for aggregators when providing new or improved datasets for the Europeana collection. As such, the development and implementation of the use of Data Quality Plans provides a good methodology for future projects when working with aggregators towards shared project goals.

Activity 3: Improving data and metadata quality

Activity 3 of the Europeana Common Culture project focuses on improving the data and metadata quality of new and existing data on Europeana Collections. Measurable data improvement is understood here as raising existing content and metadata tiers of digital cultural heritage objects, as specified by the Europeana Publishing Framework (EPF)[1]. Besides this, task 3.3 of the activity has a specific focus of performing and testing different semantic enrichment processes on a small selection of datasets from different partners to see the effect on discoverability and usefulness.

The main objectives of Activity 3 as stated in the Common Culture grant agreement can be summarised as follows:

[1] https://pro.europeana.eu/post/publishing-framework

- Raise 4 million of existing content on Europeana Collections to at least Tier 2;
- c. 1.7 million new records are delivered, complying with Tier 3 or 4 specifications;
- Complete the transition to Europeana Data Model (EDM) of the metadata structure that are still using Europeana Semantic Elements (ESE);
- Enhance multilingual features by widening the adoption of multilingual vocabularies;
- Enrich the metadata of at least 8 datasets by connecting to semantic resources available as linked data.

The work under Activity 3 is organised in three tasks. Task 3.1 is focused on developing plans for improving data quality of participating National Aggregators; Task 3.2 implements the plans developed in 3.1 and ensures that the National Aggregators reach their targets; finally, Task 3.3 focuses on semantic enrichment processes and assesses a number of semantic enrichment tools.

The first chapter of this report will focus on Task 3.1 and the data plans that were developed. It will dive deeper into the methodology, looking into why and how the data plans were developed and what tier objectives were set for the participating NAs.

The second chapter will focus on Task 3.2 and how the data plans were implemented and how the objectives were met. It will showcase data quality improvements, as well as make an assessment of the challenges and the outcomes of this task.

The third chapter will focus on Task 3.3, introducing the participants, and the development of the plan for semantic enrichment processes, and zooming in on the lessons learned and outcomes of these enrichment activities.

Finally, the fourth chapter is the conclusion of this report outlining the completion of objectives and lessons learned for future data aggregation and enrichment activities to enhance data quality.

Activity 3 is led by the Netherlands Institute for Sound and Vision (NISV), with the Europeana Foundation (EF) in charge of task 3.1 and NISV for tasks 3.2 and 3.3. NISV coordinated the collection and review of the Data Quality Plans (DQPs) of the Common Culture partners.

The partners in tasks 3.1 and 3.2:

Short name	**Full name**	**Country**
UMA	UMA Information Technology	AT
PSRL	‘Pencho Slaveykov’ Regional Library	BG

DDB	German National Library - German Digital Library	DE
EIE	National Hellenic Research Foundation – National Documentation Centre	GR
DIGIPHIL	Petőfi Literary Museum	HU
ICCU-MIBACT	Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information	IT
MM-NLL	Martynas Mažvydas National Library of Lithuania	LT
NISV	Nederlands Instituut voor Beeld en Geluid	NL
PSNC	Poznan Supercomputing and Networking Center	PL
BNP	Biblioteca Nacional de Portugal	PT
NUK	National and University Library, Slovenia	SL
RAA	Riksantikvarieämbetet – Swedish National Heritage Board	SE
NLS	National Library of Serbia	RS
MOEC	Cyprus Ministry of Education and Culture	CY
CELN	Consortium of Estonian Libraries Network	EE
MECD	Ministerio de Educacion, Cultura y Deporte	ES
NLL	National Library of Latvia, LT	LT
TCD	Trinity College Dublin	IE
UH (FINNA)	National Library of Finland	FI

The partners in task 3.3:

Short name	**Full name**	**Country**
UMA	UMA Information Technology	AT
DDB	German National Library - German Digital Library	DE
EIE	National Hellenic Research Foundation – National Documentation Centre	GR
UH (FINNA)	National Library of Finland	FI

Task 3.1 Content coordination

Background

With the Europeana Publishing Framework (EPF), the Europeana Initiative has developed a clear strategy for data partners to deliver their digital cultural heritage objects with high quality content and metadata in order to be found, viewed, shared, used and reused by audiences in the best possible way. Several tiers have been devised for sharing collections with EF based on what data partners provide and what can be expected in return (from how the items will be showcased on the Europeana website to use in third-party applications). There are tiers for content[2] and, more recently,for metadata[3] and an accompanying practical Publishing Guide explains the requirements of each tier and what needs to be included in the data[4].

A Data Quality Plan (DQP) can be used as a pragmatic tool to translate the strategic aim for data quality and high level content and metadata tiers into actionable objectives. EF has worked in previous projects with DQPs, most notably within the Europeana Digital Service Infrastructure (DSI) projects. The DQP offered an actionable and measurable way to achieve the overall Activity 3 aims. The DQPs set up the methodology for T3.2 to implement the plans and monitor the progress in the quality improvements of provided datasets.

Before the start of the project, EF and Activity 3 leader NISV completed a DQP template[5] and a practical guide for NAs to create their own plan[6]. These were shared via the project Basecamp at the start of the project, and presented during the kick-off meeting in Riga, Latvia in mid January 2019, so that all partners were fully informed early on in the project In the weeks after the project kick-off the DQPs and the goals for the content and metadata tier objectives were agreed individually with each NA participating in the task.

[2]https://pro.europeana.eu/files/Europeana_Professional/Publications/Publishing_Framework/Europeana_publishing_framework_content.pdf

[3]https://pro.europeana.eu/files/Europeana_Professional/Publications/Publishing_Framework/Europeana_publishing_framework_metadata_v-0-8.pdf

[4] https://pro.europeana.eu/post/publication-policy

[5] Data Quality plan template Common Culture 2019/2020 - https://docs.google.com/document/d/1ZEYHr6lAgZ5mkwi-oHxfm_NzkC9l3704Z6MhXCUQFwU

[6] Data Quality Plan guide for Common Culture partners - https://docs.google.com/document/d/1XZieScfiREyki84DRH_uGCbbM_eNJMIVMpXgYewLuUE

Methodology - Development of Data Quality Plans

The main aim of Activity 3 is to improve data quality. But what is data quality for the Europeana Initiative and why was the work in T3.1 and T3.2 structured around the development of DQPs? Higher quality and more meaningful content and metadata means better discoverability, viewing, sharing, use and reuse of cultural heritage collections wherever and whenever possible. Good data quality in the cultural heritage collections provided to Europeana means that the data includes direct and working links to digital objects, high quality previews and digital objects, accurate rights statements, context (time, place, type, subject information) and language attributes, multilingual (LOD) vocabularies, depth of description (including meaningful titles) and at least all mandatory elements as described in the EDM documentation. These points are described in depth in the frameworks and documentation available to data partners that want to submit data[7]. A DQP can then be used as a mechanism to translate the strategic aim for data quality and the high level content and metadata tiers into actionable objectives.

We proposed to structure and operationalize the DQPs based on the tiers for content and metadata as described by the EPF, while also keeping to the Publishing Guide and EDM guidelines. Although the focus of Activity 3 was on the content tiers in accordance with the grant agreement, NISV and EF in devising the DQPs did include objectives for the metadata tiers in a bid to improve data quality along all lines. The intention was that each Common Culture NA partner should get a better understanding about the quality of their own datasets and agree on goals they need to work on as part of Activity 3.

The main objectives of Activity 3 were refined and extended into more specific content and metadata tier objectives and actionable goals. As a result, in the DQP template, six generic objectives were established for all NA's to specify further for the datasets they selected to work on. Datasets that meet these objectives are included in the final numbers provided in the Outcomes section.

- All updated and new datasets need to be delivered in valid EDM External[8].
- All datasets not compliant to the EPF content need to be either updated to at least EPF Content tier 1 or depublished.
- All updated datasets need to be at least EPF Content tier 2 to count towards the project targets.
- All new datasets need to be at least EPF Content tier 3 to count towards the project targets.

[7] https://pro.europeana.eu/share-your-data/process
[8] At the start of the project some 10 NA's still delivered the deprecated metadata format ESE

- All updated datasets need to be at least EPF Metadata tier A to count towards the metadata targets that were planned to be established in the DSI context.
- A certain percent of datasets need to reach EPF Metadata tier B[9].

At the start of the project, when we provided the template with the general objectives, and guidelines for creating a DQP, NAs were asked to create their own DQP and specify the information and the goals for the datasets they each selected to work on in T3.2. NAs created their own DQPs because they have the best knowledge of their datasets, the data quality status and the improvements that are feasible to make within the timespan of the project. The first step was asking each NA to create an inventory of their datasets indicating each set's content and metadata tier. This then provided input for each of their individual DQP objectives. Content tier inventories existed at Europeana for some NA's, other NA's made the content tier inventory themselves following the provided guidelines. The metadata tier inventories were provided by EF. Between M3 and M6 once all the plans were received, EF and NISV jointly gathered, reviewed and revised the DQPs of each partner.

Data Quality Plans

The results of task 3.1 are the individual data quality plans of each Activity partner. The plans became an internal living document to monitor and assess the data quality improvements. They were adjusted with the ongoing work in T3.2 when more details on the datasets became available.

The process of the partners creating the DQPs, analyzing their datasets, and submitting drafts on the one hand, and EF and NISV gathering, reviewing and refining all DQPs, on the other hand, was an intensive task that took several months to finalize. Much more detail was needed than specified in the objectives of the project proposal. The time spent and the details gathered were, however, needed to have a good understanding of the feasibility of executing each DQP, which in turn would lead to higher content and metadata quality. Partners worked on the improvement of their data while refining and updating the DQP's, making T3.1 and T3.2 a parallel process with a close working relationship between EF and NISV.

[9] For the metadata objectives the partners had the flexibility of deciding which metadata fields of the EPF Metadata component to work on for each dataset, since some fields might be more relevant and enriching for some records/datasets compared to others.

Task 3.2 Improving each participant's data quality according to data plans

Context and process

The objective of task 3.2 is to improve each NA's data quality according to the content and metadata tier objectives in the DQPs as established in task 3.1. The DQPs are used to internally monitor the progress in quality improvements that apply to the goals and datasets identified in each partner's plan.

NISV is the coordinator of this task, implementing the DQPs and monitoring data quality improvement progress. EF is responsible for the ingestion of the data and providing feedback on the data quality in a timely manner.

The DQPs are implemented in this task and the proposed quality improvements carried out by the NAs. Improved and new datasets for Common Culture follow the normal Europeana route for data ingestion and publication. In this task the process was monitored and assessed in an internal midterm review and through the creation of a statistical overview of tier divisions per dataset and partner.

Implementation of the Data Quality Plans

Before the project began, there were early estimations of the target tier division for each NA. This estimated target was the basis for each DQP. Based on these estimations, the NAs then specified which improvements they aimed to make for each of the content and metadata tier objectives and what datasets would be worked on. It wasn't until July 2019 when the technical implementation of the Publishing Framework was completed that the tier metrics could be calculated at ingestion time and stored in the metadata. This allowed a much more precise monitoring of the progress towards the targets set forth in the DQPs. At the end of the project the final status of the distribution of tiers was again measured. In the Outcomes section these statistics are compared to measure the success of the data quality improvements.

In this task, every NA has its own tools and processes in place to work on their data, still there are some commonalities across all the NAs that can be mentioned here, as described in the report on Activity 2 *Landscape of National Aggregation in Europe*.

A good and close collaboration between each NA and their local Cultural Heritage Institutions (CHIs) is key to improving data quality. NAs must be able to provide general support to their CHIs on issues such as metadata and mapping, aggregation workflow process and IPR guidelines.

NAs are often flexible in accepting any metadata schema from their CHIs and then transform the source metadata to valid EDM. Making correct metadata mappings from the source data to EDM is an important step for NAs to ensure data quality. NAs often process the data using metadata enrichment, normalisation and validation processes, these can be used as a set of quality rules for the CHIs and guarantee a level of quality. Validation might check presence of mandatory fields or the correctness of the data format and structure. Quality can be further improved by applying normalization, like cleaning metadata and using value lists. Enrichment takes the information that is given in the metadata fields, and adds more comprehensive information either manually or (semi-)automatically, e.g. adding links to LOD vocabularies.

However, technical processes are not the only relevant aspect for improving data quality. CHIs often need clear incentives, for example, to apply a license that allows for wider reuse or to provide content in the highest quality available, and the related negotiation with the NAs is extensive and time consuming. Incentives for CHIs can be the opportunity to be part of a Europeana blog, gallery, exhibition, thematic collection, editorial, social media post or to the reuse of their data for education, research, and in third party applications and services. This ties in with the EPF motto that the more you give in terms of data quality, the more you get, like audience reach and overall visibility.

A few NAs share below how they improved their data quality and some of the outcomes of that work :

Kulturpool (UMA) managed to negotiate with their CHIs better content quality and therefore an increase in Tier 4 material to 60,300 records in total, while they also provided their first IIIF dataset. They also consolidated datasets from two inactive Europeana aggregators, decreasing the amount of Tier 0 material on Europeana Collections.

To improve the quality of the datasets, the National Library of Serbia (NLS) upgraded the in-house metadata conversion tools, as well as the conversion via the Aggregator for Europeana. The existing datasets on Europeana were reviewed. The licences, content and metadata tiers were checked and updated where needed and possible. In order to polish the metadata and boost the user experience, the following is now included in the metadata wherever possible: edm:Place and edm:TimeSpan, skos:note with the script in which the names of creators and/or contributors are written (Cyrillic or Latin); birth and death years for the creators and contributors; Wikidata links. The presence and adequacy of language tags was also checked.

FINNA provided updates of several datasets by improving the metadata and fixing broken links. The rights statements of these datasets were also checked and re-defined during the project. Their largest dataset from the National Library of Finland with more than 500.000 records with Tier 0 and 1, was improved for 99% to Tier 4. What was important for FINNA and the CHIs they work with was that Europeana started to support the multilingual YSO vocabulary, which is the General Finnish ontology. This development supports multilingual access to Finnish content in Europeana, and can be an added incentive for Finnish CHI's.

SOCH continuously worked throughout the project to check incorrect licenses on objects. There was ongoing communication directly with the data partners about how they can increase data quality, and most datasets were reharvested with updated content and metadata. Through a major communication effort in early summer 2019, they contacted over 100 Swedish cultural heritage institutions to investigate whether they want to become a partner to SOCH and Europeana. This has resulted in 21 new data partners during the project which contribute more than 130,000 new objects in Tier 3-4. The total number of new and existing objects in Tier 3-4 has increased with more than 339.000 objects during the project. The EDM mapping for Europeaana was amended to include the xml tag for language (Swedish).

Most of EKTs updated sets were initially rated as Tier 0 due to broken links. In many of these cases the repositories were functioning but the OAI-PMH was not. For those EKT made extensive mappings and web crawling procedures and managed to reharvest the content directly from the CHIs websites, including functional URLs to digital files and thumbnails. To be able to do so, EKT first had to extend their own Harvester tool in order to be able to collect metadata from item pages (scraping) gathered via web crawling techniques. EKT has enriched over half a million records by 62 providers. During the Common Culture project they created a bilingual vocabulary (greek and english) of subjects based on the UNESCO thesaurus and conducted thematic enrichments to the entirety of their content. They also conducted enrichments on Historical Periods using an EKT vocabulary linked to DPedia. After those enrichments the majority of their metadata is tiered B and C.

Monitoring Progress and Supporting Partners

In mid 2019, NISV, with input from EF, completed an in-depth review of each partners' status against their original data quality plan targets, with an eye for identifying risks or points for clarification. Between June and September 2019, NISV conducted individual calls with each partner and a representative from the EF Data Publishing Services team to discuss, amend and plan for harvesting all new and updated datasets before then end of the project.

Based on these check-ins, NISV developed a detailed publication schedule for the EF Data Publishing Services team to plan and prepare for the incoming sets. Over the remaining course of the project we worked with the partners to maintain and update the publication schedule, especially in light of the project extension and taking into account an overall slow down of activity due to the Covid-19 pandemic lockdown measures across Europe.

Furthermore, in late 2019 NISV implemented a peer to peer knowledge sharing model for the Activity 3 partners to foster direct conversations between the NAs and encourage them to offer advice and guidance to each other. Six groups of 3-4 partners each were set up and initiated a first joint conversation. The groups consisted of a mix between established and emerging aggregators and/or those with similar infrastructure or experiences. For example, the DDB, German National Library, was connected with the Estonian aggregator, CELN, since they are using a version of the DDB software for their mappings. Conversations with the partners in Cyprus and Hungary with Poznan in Poland were facilitated to support them using the Repox software.

Request for extension

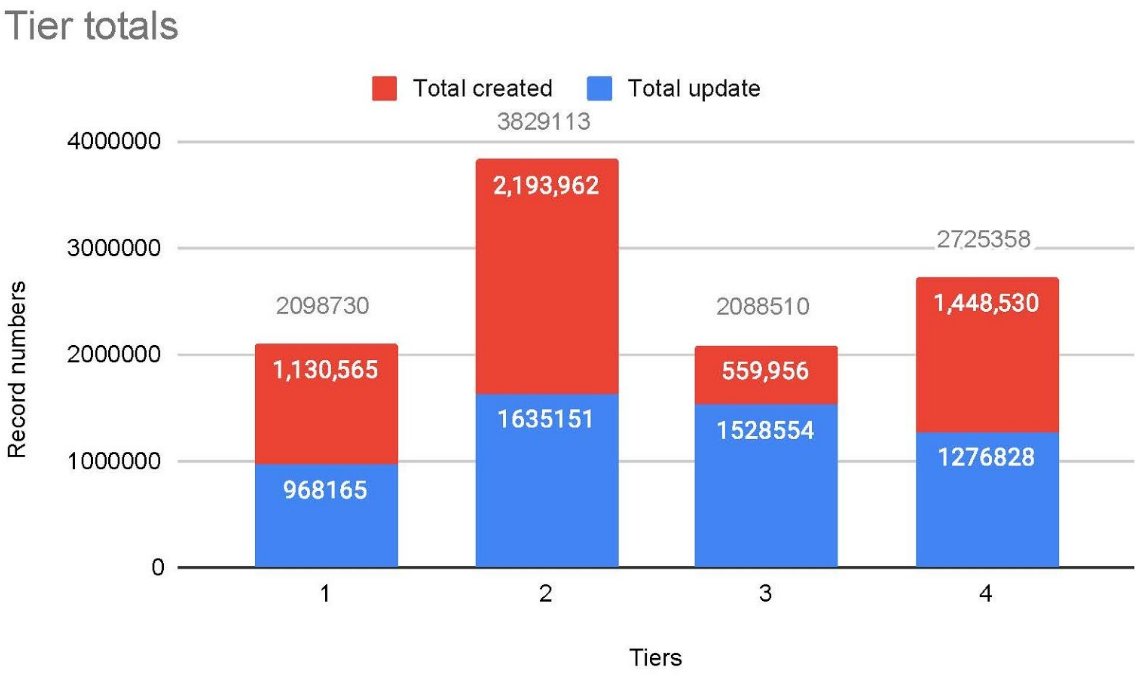
Despite our efforts to connect with and support the partners by early 2020 it was becoming clear we would not reach our Activity targets by end of June 2020 and would need a six month extension until end of December 2020. At that time, the project had only reached 15% of the target 4 million improved records (600,000 records) and 13% of the target 1.7 million new records (220,000 records). This was mainly due to the fact that many of the project partners faced various challenges, such as staff changes, technical problems, and, most importantly, a strong dependency on a few CHIs to submit data. Our conservative estimate at the time was that by mid-2020 we would reach 50% of the data targets eg. 2 million improved records from tier 0 and 1 to tiers 2+ and above and not more than 800,000 new records in tiers 3/4.

As can be seen in the outcomes below, the additional time was well invested and resulted in us collectively being able to successfully reach our target objectives. This in spite of the additional burden of a slow down in activity across Europe due to the Covid-19 pandemic.

Outcomes

At the time of writing (17 January 2020), the participating NAs improved 4,440,533 records to tier 2+ (111.01% of the target) and provided 2,008,486 new records in tier 3+ (118.15% of the target). A particular highlight is that 27% of the newly created records (1,448,515 records) and ca. 24% of the updated records (1,276,843 records) are in tier 4.

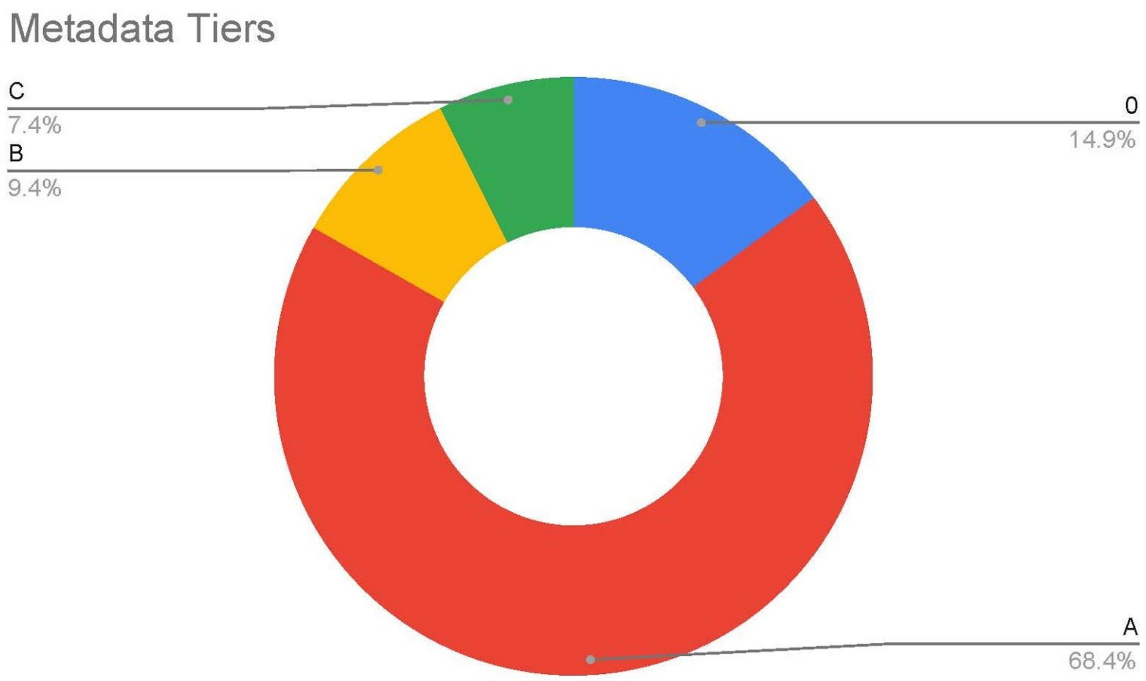
The total distribution of all tiers (including both updated and newly created records can be seen in the chart below:



Based on the final stats the quality improvement objectives for Activity 3 have been reached successfully. The objective was to raise 4 million existing content on Europeana Collections to at least Tier 2, we managed to exceed the target and have raised 4,440,533 records into Tier 2+. Up to 1.7 million new records needed to be delivered as Tier 3+, and we exceeded this target as well by providing 2,008,486 new content in Tier 3+ with many more datasets already delivered by project partners before the end of the project that are still in the publication pipeline.

Another objective was to complete the transition to EDM of the metadata structures that are still using ESE. This objective was successfully met since all updated and new datasets were provided in EDM, the only format formally accepted by EF.

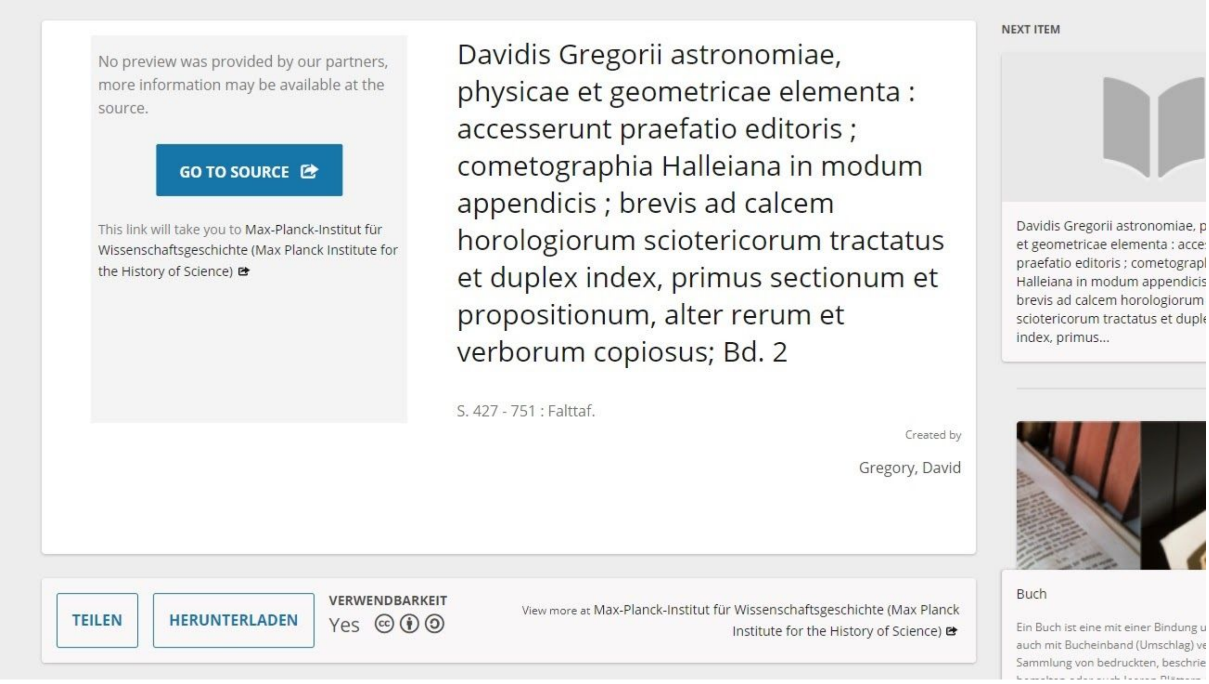
An additional highlight, which was not part of the original objectives of this task, is the high quality metadata we received from the project partners. Once we were able to accurately measure metadata tiers, from July 2019, we were able to work with the project partners to advise and support them in providing high quality metadata with their records. The chart below demonstrates the distribution of metadata tiers for all the datasets submitted by project partners over the course of the project. More than 85% of all metadata is in tiers A-C. This was not an explicit objective of the project; however, it significantly improves the services to our users. Improving the metadata quality makes the collections more discoverable, also powering the new browse experience of Europeana Collections.



Data quality examples

Here some examples of data quality improvements under Task 3.2 are provided. These examples demonstrate that data quality is not a goal in itself, but serves a greater purpose, as stated in the EPF motto - the more you give in terms of data quality, the more you get - like audience reach and overall visibility. Activity 5 used the DPQs and the work in T3.2 to highlight new high quality tier 3+ material. This material enriches the Europeana Thematic Collections and was used for editorials, blogs, galleries and an exhibition. Overally, this will encourage other NAs and CHIs to also provide high quality data.

German National Library - German Digital Library (DDB) The DDB fixed the broken links in their datasets, for instance in the Max Planck Institute for the History of Science set. After updating the set, all objects of the set are Tier 4.



No preview was provided by our partners, more information may be available at the source.

This link will take you to Max-Planck-Institut für Wissenschaftsgeschichte (Max Planck Institute for the History of Science)

Davidis Gregorii astronomiae, physicae et geometricae elementa : accesserunt praefatio editoris ; cometographia Halleiana in modum appendicis ; brevis ad calcem horologiorum sciotericorum tractatus et duplex index, primus sectionum et propositionum, alter rerum et verborum copiosus; Bd. 2

S. 427 - 751 : Falttaf.

Created by

Gregory, David

VERWENDBARKEIT

Yes

View more at Max-Planck-Institut für Wissenschaftsgeschichte (Max Planck Institute for the History of Science)

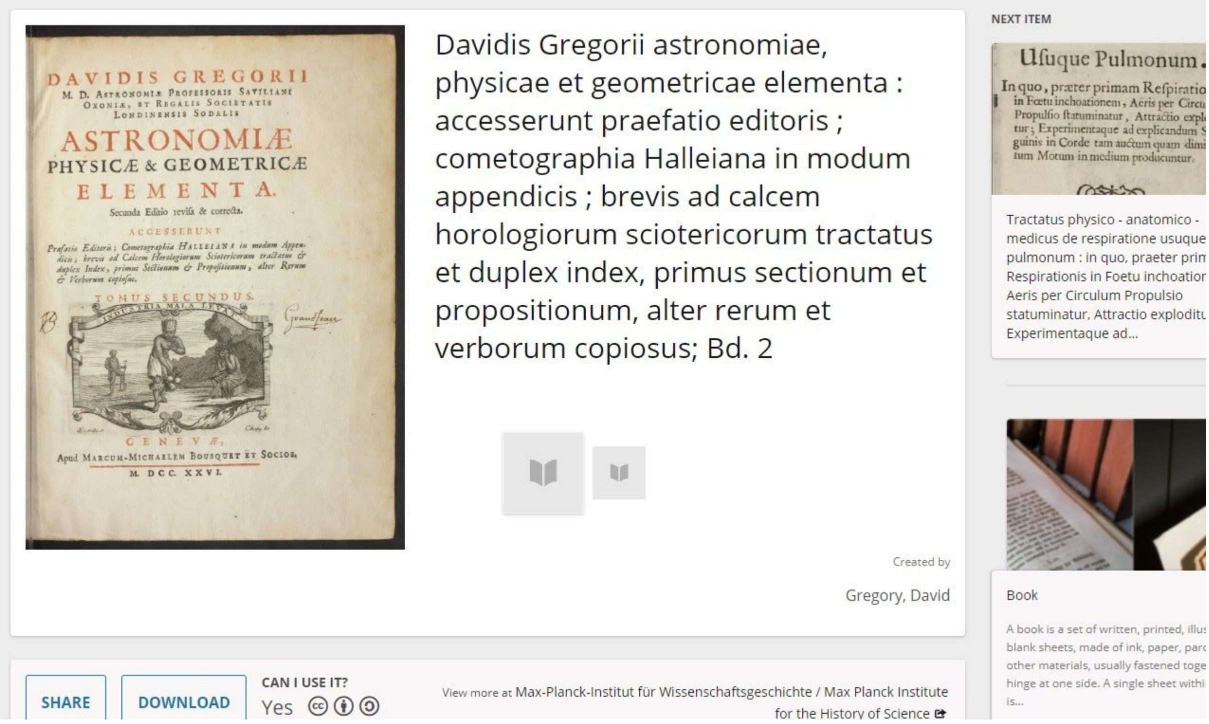
NEXT ITEM

Davidis Gregorii astronomiae, p
et geometricae elementa : acce
praefatio editoris ; cometograp
Halleiana in modum appendicis
brevis ad calcem horologiorum
sciotericorum tractatus et dupl
index, primus...

Buch

Ein Buch ist eine mit einer Bindung u
auch mit Bucheinband (Umschlag) ve
Sammlung von bedruckten, beschrie

Image 1: Before update: Tier 0 because of broken links.



Davidis Gregorii astronomiae, physicae et geometricae elementa : accesserunt praefatio editoris ; cometographia Halleiana in modum appendicis ; brevis ad calcem horologiorum sciotericorum tractatus et duplex index, primus sectionum et propositionum, alter rerum et verborum copiosus; Bd. 2

Created by

Gregory, David

CAN I USE IT?

Yes

View more at Max-Planck-Institut für Wissenschaftsgeschichte / Max Planck Institute for the History of Science

NEXT ITEM

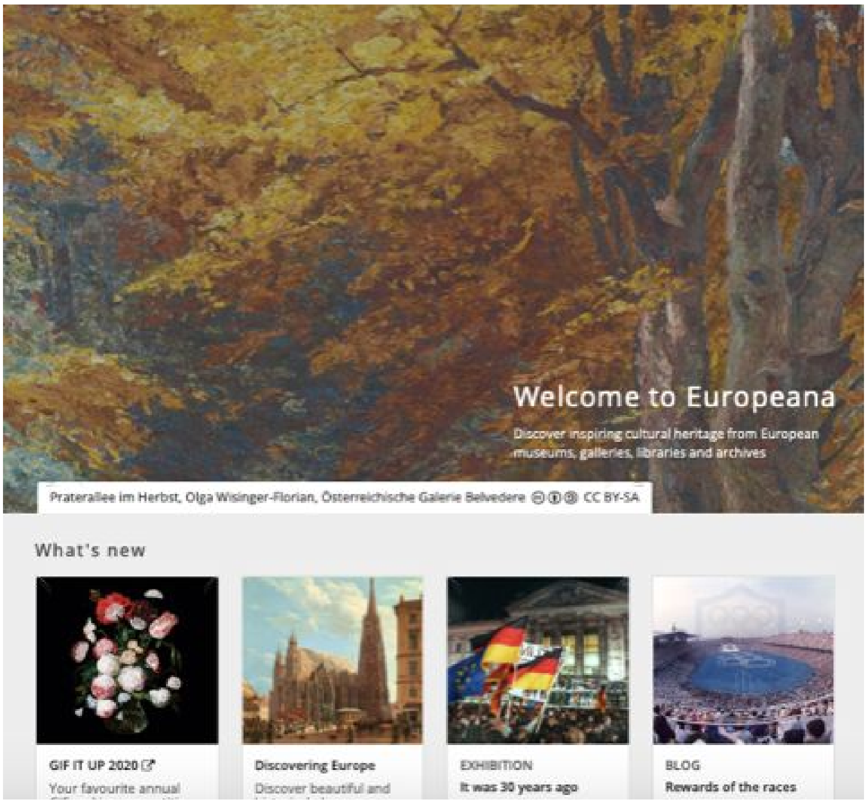
Tractatus physico - anatomico -
medicus de respiratione usuque
pulmonum : in quo, praeter prim
Respirationis in Foetu inchoatior
Aeris per Circulum Propulsio
statuminatur, Attractio exploditu
Experimentaque ad...

Book

A book is a set of written, printed, illus
blank sheets, made of ink, paper, parc
other materials, usually fastened toge
hinge at one side. A single sheet withi
is...

Image 2: After update: content Tier 4.

Kulturpool (UMA) New material submitted over the course of the project in Tier 4 was used on the homepage of the Europeana website.



Welcome to Europeana

Discover inspiring cultural heritage from European museums, galleries, libraries and archives

Praterallee im Herbst, Olga Wisinger-Florian, Österreichische Galerie Belvedere CC BY-SA

What's new

GIF IT UP 2020

Your favourite annual

Discovering Europe

Discover beautiful and

EXHIBITION

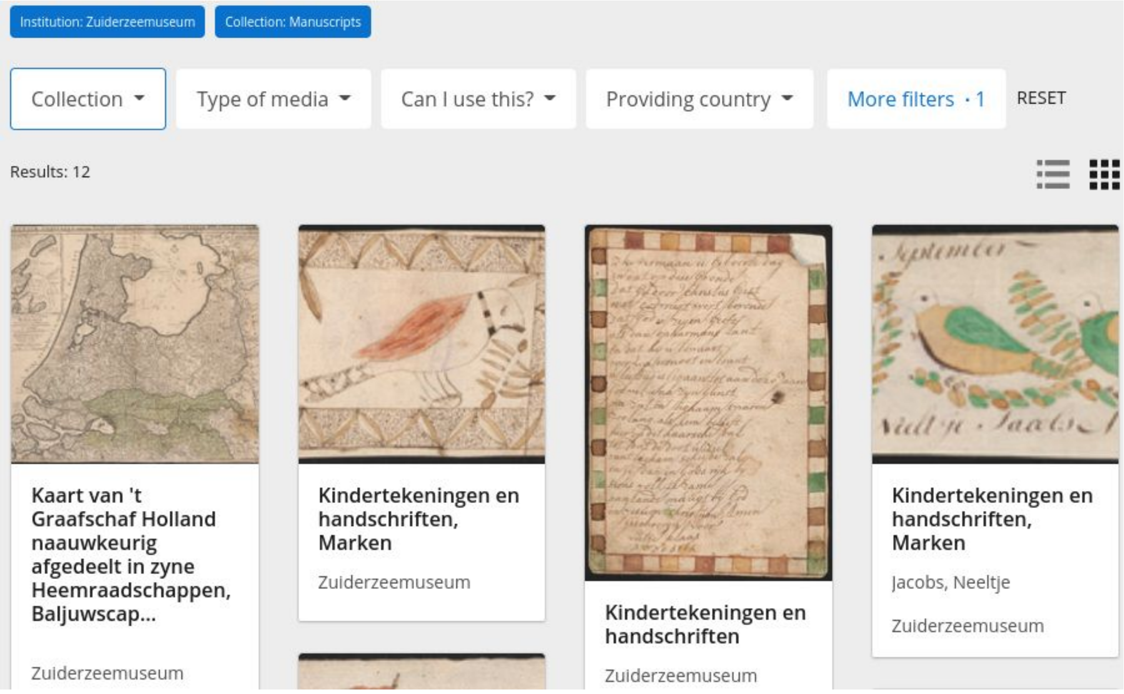
It was 30 years ago

BLOG

Rewards of the races

Image 3: The homepage of the Europeana website, showing a Kulturpool object.

Digitale Collectie (NISV) Digitale Collectie provided a new dataset from the Zuiderzeemuseum. Most of the 17,000 items are content Tier 4, and the set is part of several Thematic Collections, like Manuscripts.



Results: 12

Kaart van 't Graafschaf Holland naauwkeurig afgedeelt in zyne Heemraadschappen, Baljuwscap...

Zuiderzeemuseum

Kindertekeningen en handschriften, Marken

Zuiderzeemuseum

Kindertekeningen en handschriften

Zuiderzeemuseum

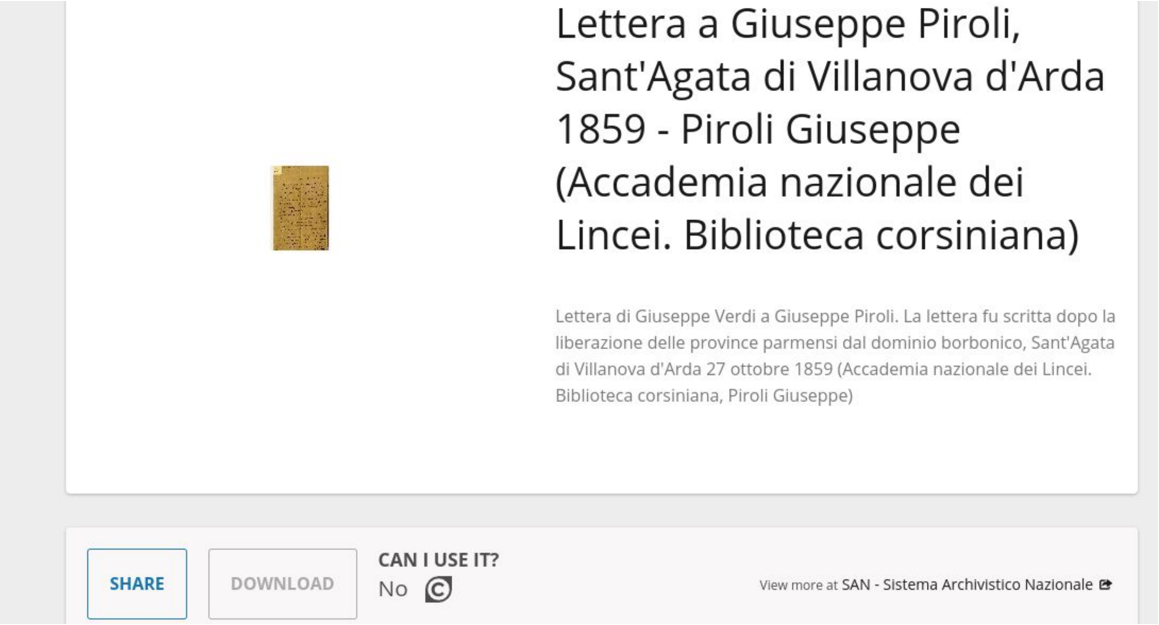
Kindertekeningen en handschriften, Marken

Jacobs, Neeltje

Zuiderzeemuseum

Image 4: A new collection of Tier 4 material, added to different thematic collections, like Manuscripts.

CulturaItalia (ICCU) The dataset of SAN - Sistema Archivistico Nazionale was taken from an inactive aggregator and resubmitted by CulturaItalia. With this update the content quality was improved considerably to Tier 3+.



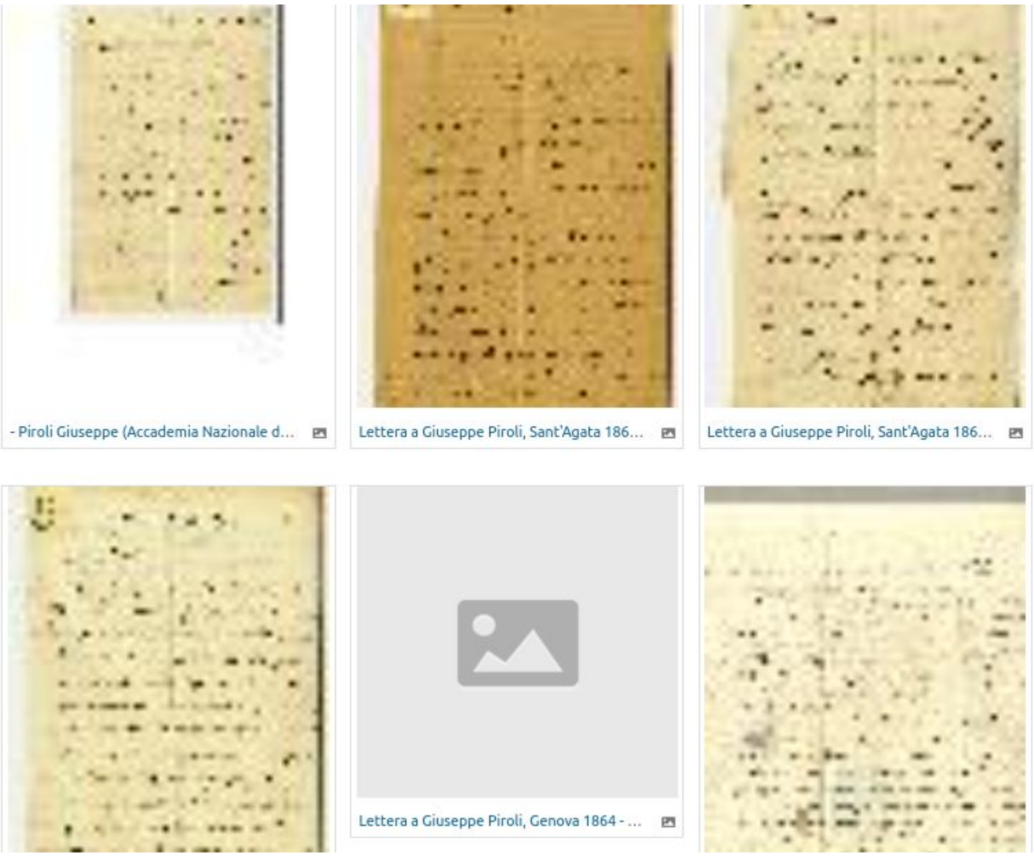
Lettera a Giuseppe Piroli, Sant'Agata di Villanova d'Arda 1859 - Piroli Giuseppe (Accademia nazionale dei Lincei. Biblioteca corsiniana)

Lettera di Giuseppe Verdi a Giuseppe Piroli. La lettera fu scritta dopo la liberazione delle province parmensi dal dominio borbonico, Sant'Agata di Villanova d'Arda 27 ottobre 1859 (Accademia nazionale dei Lincei. Biblioteca corsiniana, Piroli Giuseppe)

SHARE DOWNLOAD CAN I USE IT? No ©

View more at SAN - Sistema Archivistico Nazionale

Image 5: Before update: small, low resolution images.



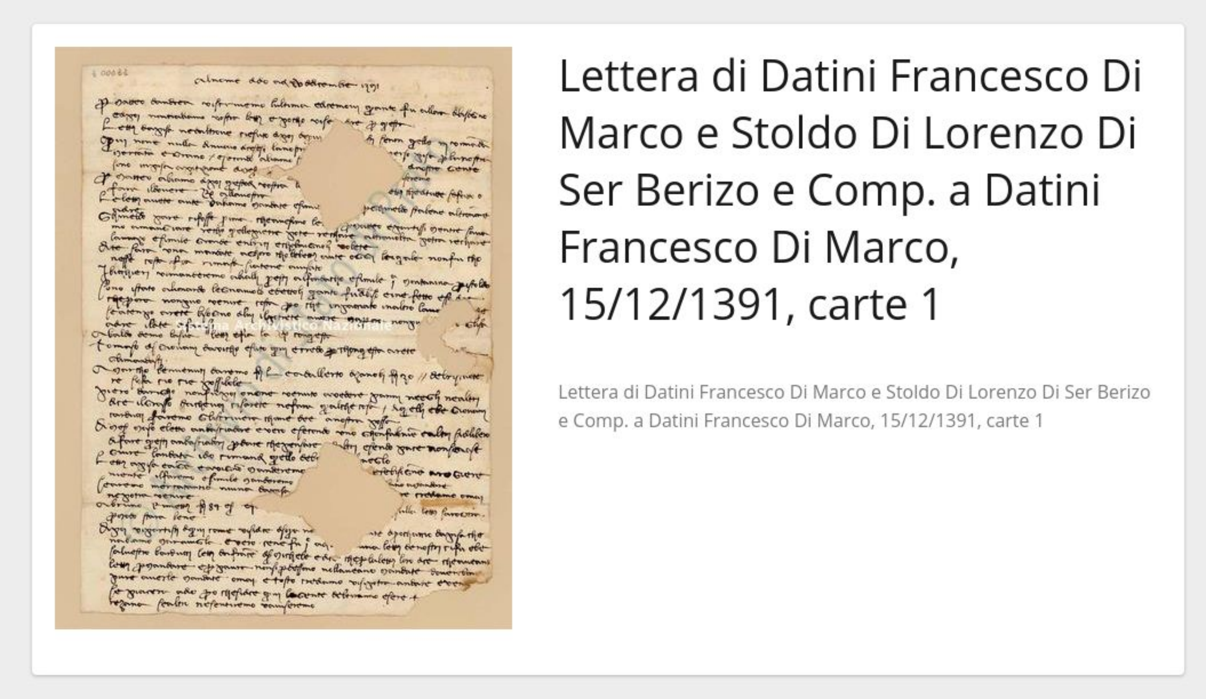
- Piroli Giuseppe (Accademia Nazionale d...

Lettera a Giuseppe Piroli, Sant'Agata 186...

Lettera a Giuseppe Piroli, Sant'Agata 186...

Lettera a Giuseppe Piroli, Genova 1864 - ...

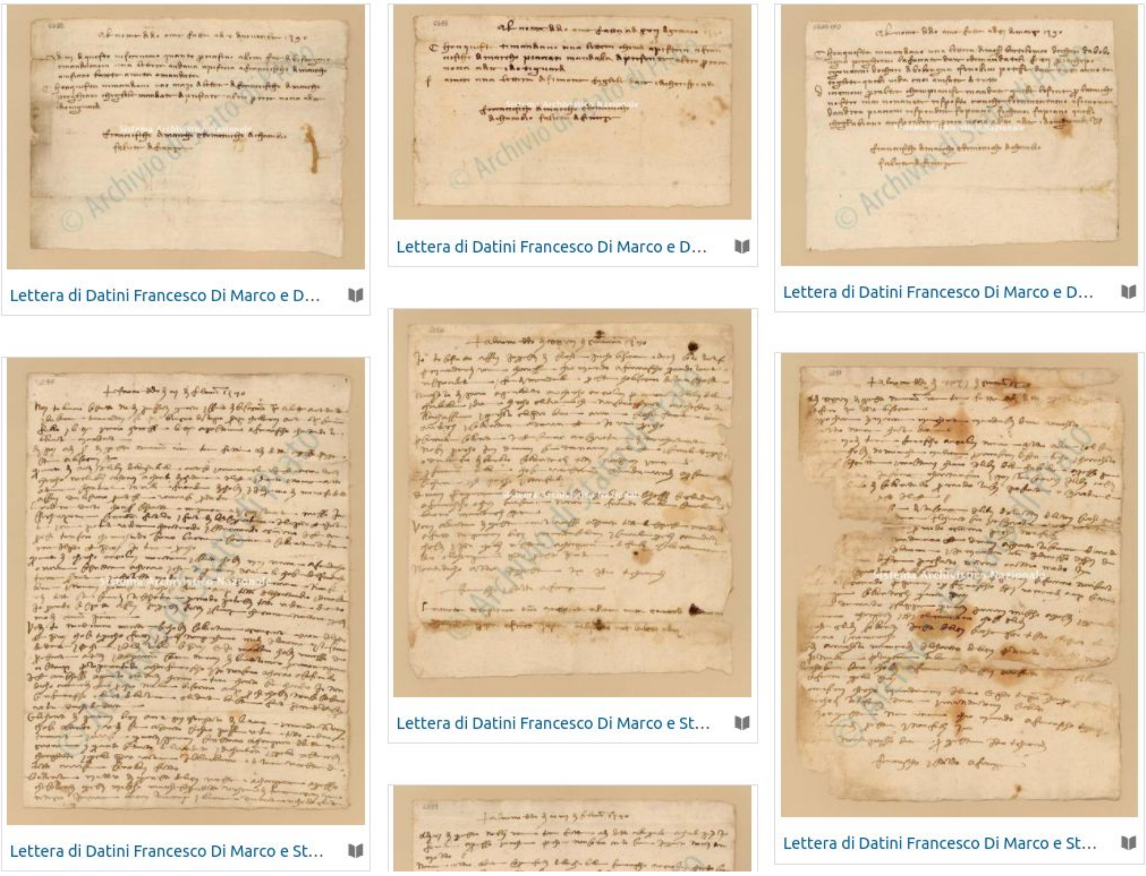
Image 6: Grid view before update with low quality previews.



Lettera di Datini Francesco Di Marco e Stoldo Di Lorenzo Di Ser Berizo e Comp. a Datini Francesco Di Marco, 15/12/1391, carte 1

Lettera di Datini Francesco Di Marco e Stoldo Di Lorenzo Di Ser Berizo e Comp. a Datini Francesco Di Marco, 15/12/1391, carte 1

Image 7: After update: larger, higher resolution images.



Lettera di Datini Francesco Di Marco e D…

Lettera di Datini Francesco Di Marco e D…

Lettera di Datini Francesco Di Marco e D…

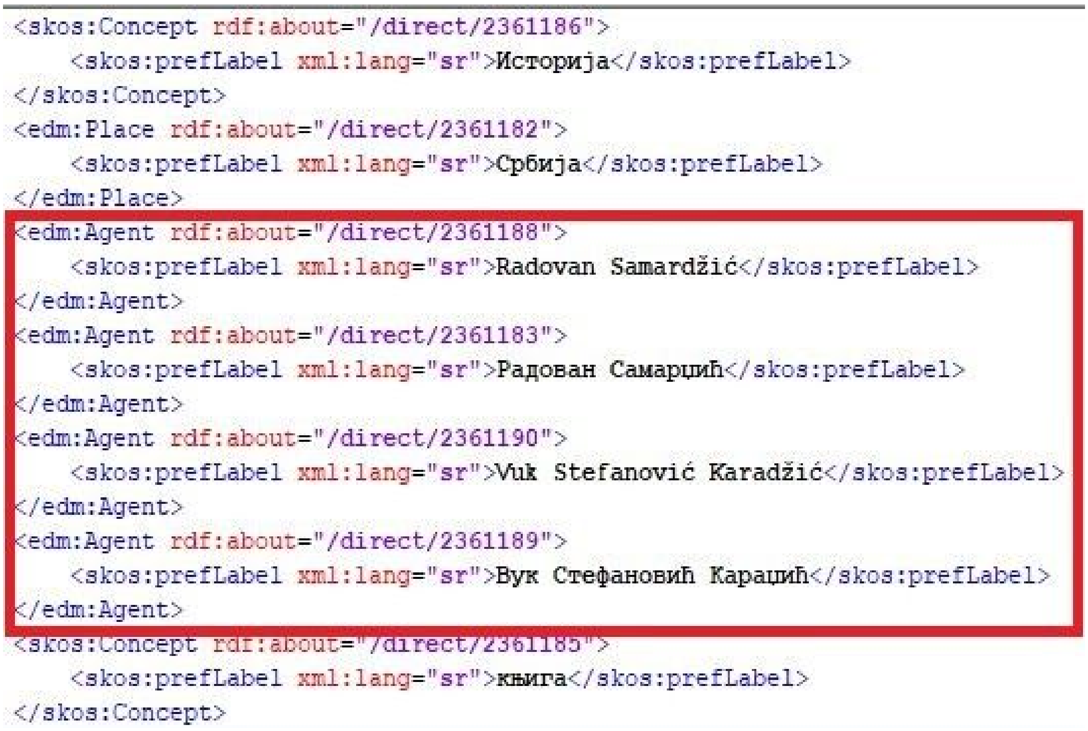
Lettera di Datini Francesco Di Marco e St…

Lettera di Datini Francesco Di Marco e St…

Lettera di Datini Francesco Di Marco e St…

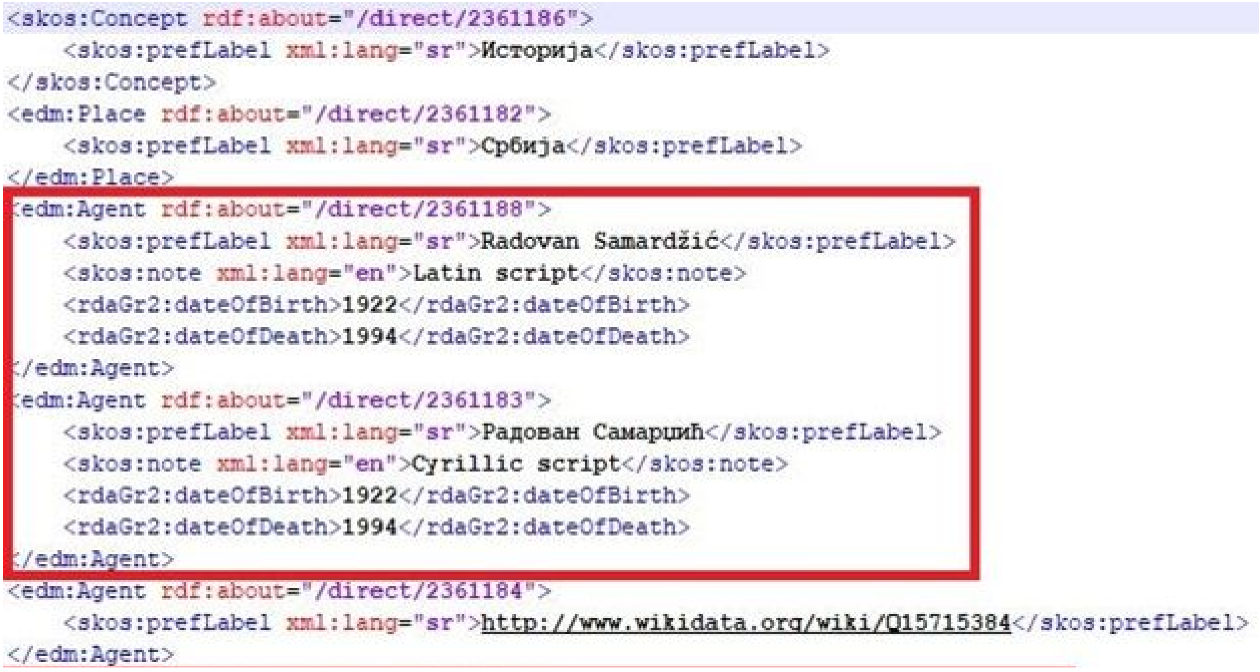
Image 8: Grid view after update.

National Library of Serbia (NLS) Besides content improvements, NLS in Serbia, also worked on metadata improvements, for example by adding extended metadata and enrichments with the wikidata vocabulary.



```
<skos:Concept rdf:about="/direct/2361186">
    <skos:prefLabel xml:lang="sr">Историја</skos:prefLabel>
</skos:Concept>
<edm:Place rdf:about="/direct/2361182">
    <skos:prefLabel xml:lang="sr">Србија</skos:prefLabel>
</edm:Place>
<edm:Agent rdf:about="/direct/2361188">
    <skos:prefLabel xml:lang="sr">Radovan Samardžić</skos:prefLabel>
</edm:Agent>
<edm:Agent rdf:about="/direct/2361183">
    <skos:prefLabel xml:lang="sr">Радован Самарџић</skos:prefLabel>
</edm:Agent>
<edm:Agent rdf:about="/direct/2361190">
    <skos:prefLabel xml:lang="sr">Vuk Stefanović Karadžić</skos:prefLabel>
</edm:Agent>
<edm:Agent rdf:about="/direct/2361189">
    <skos:prefLabel xml:lang="sr">Вук Стефановић Караџић</skos:prefLabel>
</edm:Agent>
<skos:Concept rdf:about="/direct/2361185">
    <skos:prefLabel xml:lang="sr">књига</skos:prefLabel>
</skos:Concept>
```

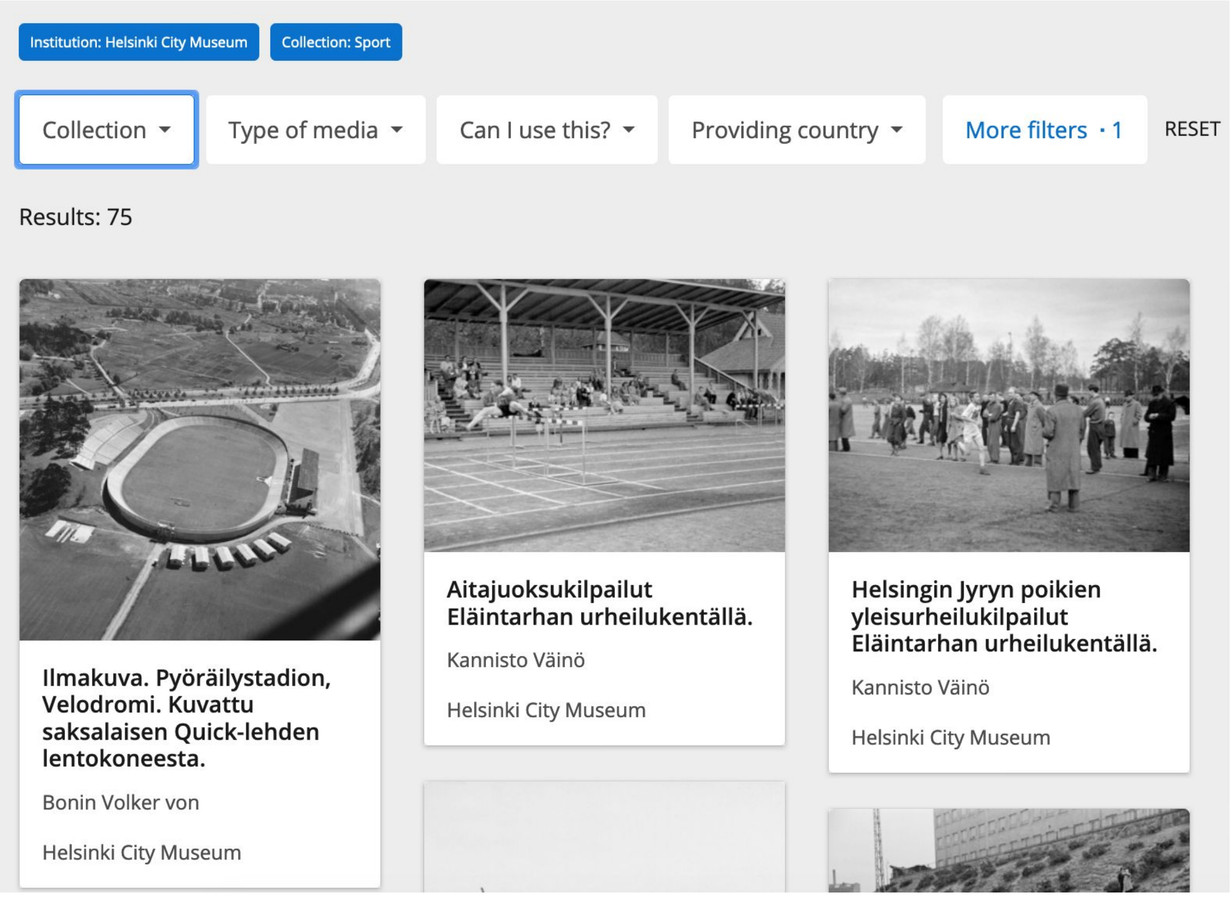
Image 9: metadata enrichments before update



```
<skos:Concept rdf:about="/direct/2361186">
    <skos:prefLabel xml:lang="sr">Историја</skos:prefLabel>
</skos:Concept>
<edm:Place rdf:about="/direct/2361182">
    <skos:prefLabel xml:lang="sr">Србија</skos:prefLabel>
</edm:Place>
<edm:Agent rdf:about="/direct/2361188">
    <skos:prefLabel xml:lang="sr">Radovan Samardžić</skos:prefLabel>
    <skos:note xml:lang="en">Latin script</skos:note>
    <rdaGr2:dateOfBirth>1922</rdaGr2:dateOfBirth>
    <rdaGr2:dateOfDeath>1994</rdaGr2:dateOfDeath>
</edm:Agent>
<edm:Agent rdf:about="/direct/2361183">
    <skos:prefLabel xml:lang="sr">Радован Самарџић</skos:prefLabel>
    <skos:note xml:lang="en">Cyrillic script</skos:note>
    <rdaGr2:dateOfBirth>1922</rdaGr2:dateOfBirth>
    <rdaGr2:dateOfDeath>1994</rdaGr2:dateOfDeath>
</edm:Agent>
<edm:Agent rdf:about="/direct/2361184">
    <skos:prefLabel xml:lang="sr">http://www.wikidata.org/wiki/Q15715384</skos:prefLabel>
</edm:Agent>
```

Image 10: metadata enrichments after update, with wikidata vocabulary and extended metadata

FINNA A new dataset with 79,680 records from the Helsinki City Museum was provided, with the majority of records in Tier 4. Some of these objects were added to the Sports thematic portal and used in an Europeana blog as well[10].



Results: 75

Ilmakuva. Pyöräilystadion, Velodromi. Kuvattu saksalaisen Quick-lehden lentokoneesta.
Bonin Volker von
Helsinki City Museum

Aitajuoksukilpailut Eläintarhan urheilukentällä.
Kannisto Väinö
Helsinki City Museum

Helsingin Jyryn poikien yleisurheilukilpailut Eläintarhan urheilukentällä.
Kannisto Väinö
Helsinki City Museum

Image 11: Some new Tier 4 items from FINNA are added to the Thematic Collection for Sport

[10] https://blog.europeana.eu/2020/11/lonkero-the-finnish-long-drinks-invented-for-the-1952-olympics/

Task 3.2 Findings and Lessons Learned

The following are a few discussion points that arose during the process of improving each participant's data quality according to a data quality plans methodology and the implementation of these plans.

Collaboration with CHIs is vital

The results of the extension gave us time to work with individual partners more closely and for them in turn to support their local CHIs in submitting datasets. Despite the additional slow down of activity all around due to the Covid-19 pandemic. The cooperation of CHIs is essential for NAs and EF to improve data quality in this Activity, but also for future quality targets. CHIs have, however, their own requirements, goals and capabilities. CHIs are not part of the project and didn't receive any funding for their work. NAs worked very closely with CHIs and helped wherever they could: with a helpdesk or support function, technical infrastructure and knowhow, depending on the role the NA has in their respective country. However, they can often not control the time and effort CHIs can and will spend on their digital collections, or have any say in the (commercial) collection systems CHIs use. This makes it difficult to estimate if data quality improvements and data submissions can actually be done in a certain timeframe, creating a risky dependency on CHIs.

Peer-to-peer model as best practice

Aggregators working in collaboration with each other and sharing knowledge as peers is a model which we tested out with good results in the project. The model has merit and should be investigated for use in future projects and capacity building activity. It is an effective knowledge sharing structure that alleviates the dependency on one core entity for guidance and support, in this case the Europeana Foundation team.

Within the project, we divided partners into small clusters of 3-4 partners each based on shared experiences, tools or methodologies. This was especially beneficial in the case of certain partners learning how to use particular software and mapping tools from other partners with more experience.

Data quality plans are an effective tool

Working with Data Quality Plans gave a very detailed overview of the work that NAs would take on in the project. It gave NAs the chance to think about the objectives in depth from the start. However, developing and working with the DQPs over the course of the project also created challenges that we can learn from and use to refine such methods in the future. Midway through the project, NISV and EF created a publication timeline, based on the DQPs, to evenly distribute the ingestion of datasets over the remaining course of the project. It was, however, difficult for NAs to commit to this schema, mostly because of the challenge in planning data submissions with their CHIs, as described above. The plan therefore did not succeed in alleviating the bottleneck for the ingestion of datasets at the end of the project.

Even so, we deem the development of the DQP methodology successful in this project. The implementation and roll-out of this new methodology was solid and gave a good indication of what is possible when working with aggregators towards shared project goals. Lessons learned and adjustments for future development of the methodology are further addressed in the Sustainability Plan of the project.

Improvement of data quality is a complex endeavour

Content improvements need to be discussed with CHIs and then adopted on their side. For instance, if digitizations done years ago are now considered low quality, upgrading these or even redoing digitizations might be costly and not in scope for CHIs at this moment. Metadata improvements can pose issues as well. NAs in many cases collect the source data from CHIs, and while exports, metadata mappings and enrichments can often be improved, getting more data that currently absent (e.g. descriptions, contextual information) from the CHIs is not possible if they already provide everything they have. Improving the way metadata is captured in the (commercial) collection management systems of CHIs, or adding metadata like descriptions is a time consuming and possibly costly process for CHIs.

These points are part of a broader discussion and policy recommendations on national aggregation policies and national digital transformation strategies across Europe. This will be further elaborated on in the conclusions below.

Nevertheless, NAs faced these challenges and managed to negotiate many content and metadata improvements with the CHIs. New and upcoming NAs learned from the partners in the project and strengthened their position, establishing their aggregation processes with quality goals in mind, and this Activity profited from this immediately. NAs now have a much better understanding of the quality of their datasets due to the creation and the implementation of the DQPs. This is very useful for further collaboration on aggregation with Europeana and with the CHIs. NAs now have more awareness of the strengths and points of improvements of their aggregation processes, which will help in future efforts to keep improving quality.

A few NAs share their stories and lessons learned below:

DDB: “Improving the content was overall difficult because it implies improving licenses and the image resolution which is not easy. When the CHI has the images with a higher resolution, but they only send us lower resolution images, it can be successfully negotiated that we get the “better” images. If they don’t have “better” images, it is basically impossible in a short period of time (1,5 years) and without financing to scan the images again to get a better resolution. The licenses issues were more flexible, although not very, because if willing, the people responsible have to take this decision through many hierarchies and committees and usually (especially museums) are very reluctant to use a permissive license. The Deutsche Fotothek accepted the DDB proposal and changed the InC license to InC-Edu for more than a million objects. The metadata mapping was also improved.”

For the National Library of Serbia (NLS), taking part in the Europeana Common Culture project has been a rewarding experience as an emerging aggregator. For NLS reaching high quality standards both for content and metadata and continually learning and improving existing practices would not have been possible without the project, because they rely on external support for the aggregation infrastructure and upgrading metadata conversion tools. NLS valued the Metis sandbox environment for the significance of checking the tiers and tracking errors at all levels.

In Sweden it is up to each data partner to develop a technical solution that SOCH can work with. Metadata is not further processed or enriched. This means that the SOCH team has very limited opportunities to increase the quality of the data that is aggregated since the data partners have to do the work on their side. One lesson learned within the project is that it takes a long time for many institutions to work on the data quality. They often have limited resources to work with the databases and often focus on new digitization or on targeted initiatives that benefit their own institution. Taking general quality-enhancing measures is very difficult for most institutions to implement. This lesson led SOCH to deviate from the decision not to enrich data and work out a solution for language tagging. SOCH also noticed that it takes a long time for CHIs to become an aggregation partner. Generally, it can take more than a year from when they contact an institution until they become a partner and even longer before they start to provide data. This is something to bear in mind for future outreach. They also learned that a direct approach and close cooperation works much better than general mailings.

EKT found it rather difficult to improve the data quality as an aggregator. Most of the datasets are provided by CHIs with limited resources to work on optimizing the quality of the digital files or to improve their infrastructures. Regarding licencing, Greek CHIs have a very proprietary relationship with their content and they are reluctant to use open licences. Additionally the Greek Archeological Law imposes further legal restrictions to the reuse of historical and archeological content. This implicated the process of getting more open licenses significantly.

Task 3.3 Semantic enrichment processes

Introduction

As specified in the grant agreement, the overall objectives for task 3.3 were as follows:

- Create an inventory of semantic resources and datasets to be enriched from amongst five task partners;
- Enrich the metadata of at least 8 datasets by connecting to semantic resources available as linked data;
- Engage a 15 person user group (curators, entrepreneurs, educators, general public) to assess the benefits of the enrichments;
- Develop a semantic enrichment production-ready software for the Europeana Network for individual memory organisations and national aggregators which makes use of existing partners' technology;
- Report on the assessment of the process.

In this section, first the tasks and its objectives are addressed. Following this, the semantic enrichment tools and datasets are introduced and the testing process described. Finally, the results are evaluated, an assessment is given and a conclusion is drawn. While setting up the methodology to implement the task objectives, it became clear that deviations were needed in order to better execute the tasks and maintain the integrity of the final results. These deviations are described in the section Methodology - setting up the task.

The lead for task 3.3 was NISV who coordinated the activities and liaised with each task partner and Europeana. The partners who provided datasets were EKT, DBB, UMA and FINNA. All partners, except FINNA, were also able to make their tools available for the testing and evaluation process. The process and the tools are detailed in the section Inventory of Tools.

Over the past few years there have been several projects and initiatives in which semantic enrichment by metadata analysis and linking values to semantic sources was explored[11], for instance by the EuropeanaTech Task Force on Enrichments and

[11] https://pro.europeana.eu/project/evaluation-and-enrichments

Evaluation. The work done by the task force focused on improving interoperability and collecting metrics for the enrichment workflows. More recently, the Europeana XX project[12] is investigating semantic enrichment processes on datasets from several DSI-partners. What sets the tasks, and in particular the testing done in this Activity apart is the manner of execution. The partners in this task have experience with the development or use of semantic enrichment software, these experiences will be further explored in this task.

Methodology - setting up the task

The main objective of task 3.3 was to develop semantic enrichment production-ready software for individual memory organisations and national aggregators which are part of the Europeana Network, using the project partners' technology. It became clear as the pilot progressed that an adjustment to this objective was necessary. As such, identifying common requirements for a stand-alone enrichment tool was not efficient or cost effective in the long run. On one hand, partners and CHIs already have their own domain or task specific enrichment solutions integrated in their workflows, along with fundamental differences in their metadata.

At the same time, progress was made on the Europeana's self-service testing environment called Metis Sandbox. This tool enables CHIs and Aggregators to test the Europeana workflow as performed by Metis and allows them to resolve issues before sending their data to Europeana. An aggregator might transform records to EDM[13] as a part of their ingestion process, before indexing the record. Semantic enrichment can be added at multiple points in the process. These developments indicated that it would be better to investigate how to reuse and improve existing tools in the network or to integrate with the future services of the Sandbox rather than develop a stand-alone semantic enrichment tool that would quickly be eclipsed.

The outcome of this task was adjusted to provide a set of mappings and scripts which can be integrated into any pipeline which supports XSLT mappings and shell scripts. In the long run this is more sustainable and practical for partners and will lead to improved metadata mappings with semantic enrichment and, therefore, higher metadata quality tiers.

[12] https://pro.europeana.eu/project/europeana-xx

[13]https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf

As a first step, a number of suggested methods of tool analysis were proposed and the question of how to establish a baseline was discussed. The establishing of a baseline is necessary in order to quantitatively compare performances of enrichment or tagging tools. Establishing a baseline by using a single representative testing record was considered. However, this idea was eventually rejected as it would not yield enough data to draw meaningful results. Finally, it was decided to proceed by using a selection of 8 datasets, from each partner, amounting to a rich selection of 237,395 records, and to cross-process them (or a subset in the case of very large sets) using each other's tools. Rather than comparing the properties of individual tools in this task, the tools were used to process a set of unfamiliar data from another provider and the results were compared. By having each tool enrich metadata from another partner's datasets, we were able to derive an indication of how domain-specific each tool is. Although a drawback to this approach is that the results are not comparable in a quantifiable way, it does provide insights into how well the tools work in a real-life scenario.

There was also a discussion as to how best to evaluate the findings. Instead of engaging an external group from the general public, the partners instead agreed to a different approach which would offer more insightful feedback. Each partner evaluated the enrichments of their data, as they have the most knowledge about their datasets. Following this, the results as a whole were brought to the Europeana Data Quality Committee[14], consisting of representatives from CHIs, National and Thematic Aggregators from across Europe (as well as Europeana Foundation staff), for a final review. Feedback is integrated in the final version of this report.

Testing semantic enrichment tools and datasets

This section provides an evaluation of the partner tools and answers the question how well the relative tools perform when faced with datasets outside its normal domain or scope. The datasets that were used to evaluate the tools are described in the next sections. The assessment is further worked out in the section Evaluations of enrichments.

Inventory of Tools

This section describes and compares three tools used by the Common Culture partners to facilitate semantic enrichment and tagging. While some of the tools are developed by the partner in-house, other tools are provided by third parties.

[14] https://pro.europeana.eu/project/data-quality-committee

	EKT - Semantics.gr	**Kulturpool - Tag Extraction Tool SAM**	**DDB - OpenRefine**
What languages does the tool support as input?	English and Greek	English and German	Any language
What input schema types does the tool support as input?	oai_dc, qdc, ese, edm	Plaintext over REST API	DC, ESE, OAI-DC, EDM (not extensively tested). Also: XML, SQL exports, csv
Which data format does the tool use as output?	edm	SAM tool as REST API use its own XML structure as output.	DC, ESE In testing: EDM
What are the source fields used for enrichment?	dc:type, dc:subject, date (dc:date, dcterms:created, dcterms:temporal)	SAM extracts persons, places, dates, organizations.	People, concepts, places, organizations
What LOD vocabularies does the tool use?	ekt-item-types[15] ekt-unesco[16] historical-periods[17]	None	Gemeinsame Normdatei[18] Geonames[19] Wikidata[20]
Is the tool open-source?	No	No	Yes
Does this tool support harvesting via OAI PMH?	Yes	No	No
Does this tool require input datasets to be published already?	No, but it helps if they are (the curator uses links to explore items from the publication site)	No	No
Can the tool process datasets from outside of Europeana?	Yes	Yes	Yes

OpenRefine (DDB)

[15] https://www.semantics.gr/authorities/vocabularies/ekt-item-types
[16] https://www.semantics.gr/authorities/vocabularies/ekt-unesco
[17] https://www.semantics.gr/authorities/vocabularies/historical-periods
[18] https://lov.linkeddata.es/dataset/lov/vocabs/gndo
[19] https://www.geonames.org/
[20] https://www.wikidata.org/wiki/Wikidata:Main_Page

OpenRefine[21] is an open-source application developed by third-parties and contributors. OpenRefine is a standalone tool that runs like a web server on an individual's computer. It uses a web browser as its interface but the data stays local. It works with local files or data from web addresses in a number of file formats, including CSV, TSV, XLS, XML, and other formats. OpenRefine was used by the DDB first for use cases, to process data in DC and ESE formats. The data was processed for:

- Consistency - This is usually the case especially when data is aggregated from multiple sources, or when a very flexible format is used (especially Dublin Core);
- Accuracy - Especially the case when data was mapped or transformed many times or is old;
- Completeness - For the cases where compulsory values or elements are missing (like license or type);
- Enrichment (reconciliation) - with controlled vocabularies;
- Exporting in one common format.

Advantages:

- Possibility of extensive transformations across the data set. The results of transformation expressions are previewed interactively with live data;
- Effective analysis of the data through facetting and clustering;
- Good visualisation of the data in a tabular format;
- Complete history of all modification, with the possibility of saving and using them on sets that have the same structure;
- Can be used by non-professionals - there is no need for extensive programming background;
- Open source;
- Strong support offered by the community, there is a good amount of documentation and tutorials available;
- It can find duplicate entries, empty cells, entry variations, inconsistencies, and patterns of errors for bulk fixing and cleaning.

Disadvantages:

- No formal technical support is available;
- Big datasets (over 200.000 records) are difficult to process without having to improve the hardware;
- The enrichment has to be often intellectually controlled to reduce the risk of false positives.

Semantics.gr (EKT)

[21] https://openrefine.org/

Semantics.gr was initially created by EKT as a platform where institutions can create and publish RDF-based vocabularies and thesauri of any kind (concepts, timespans, agents, places) or any schema (parametric schema definition support). The platform was enhanced with a mapping tool that allows aggregators to enrich their collections with vocabulary and thesaurus terms. The tool has a GUI environment with advanced automated functionalities that help the curator easily define Enrichment Mapping Rules (EMR) per collection from distinct metadata values to vocabulary terms.

The tool accesses collection metadata via OAI-PMH harvesting. After setting the EMR, they can be served on request via a REST API in JSON format. Subsequently, the EMR can be used by an aggregator to enrich the collection in a bulk and straightforward one-pass fashion. The EMR are defined per distinct value of a predefined metadata field (for example dc:type or dcterms:temporal), which is called primary field.

In special cases the curator can choose a second metadata field (for example dc:subject) to create more precise EMR in case the documentation of the primary field is poor. This metadata field is called secondary field and its values filters. For example, a metadata record may have a dc:type value “folklore object” but a dc:subject value “Jewel” that reveals a much more accurate type. The enrichment tool supports automatic suggestion of EMR which by default is based on string similarity matching between metadata field values and indexed labels of vocabulary entries (e.g. skos:prefLabel and skos:altLabel). The automatic mapping suggestion is very effective and efficient leveraging the indexing system of semantics.gr search engine, namely Apache Solr[22].

The tool can be configured to be loosely coupled to the aggregator search portal (using deep linking) allowing the curator to search the collection for items having the specific values on primary and secondary fields. The curator can create complex logical expressions on the filters of a vocabulary entry assignment in order to create finer and more precise EMR and avoid false positives. For instance, they can use the logical NOT operator for setting exceptions. When the automatic suggestion function fails to produce correct rules, the curator can set mappings manually. The enrichment tool “remembers” manual assignments in order to improve the effectiveness of auto-suggestion in future.

In certain cases, the curator can choose a highly selective descriptive field (the number of its distinct values approaches the number of all items) as a secondary field, such as dc:title or dc:description, if the values contain words or phrases that can reveal the appropriate vocabulary entry. For example a dc:title “An amphora from Attica” implies that the item is a vase. The tool searches inside such values for specific words or

[22] https://lucene.apache.org/solr/

phrases derived from the vocabulary terms and then exposes only the matches as filters (instead of the entire field values).

SAM (Kulturpool)

SAM is a REST service that allows to extract semantic information from texts, like named entities: persons, places, dates, organizations and Web addresses and keyphrases (tags). SAM offers support for German and English, but can be expanded with other languages. Because of the REST interface, this service can be integrated to any project. SAM output is data in XML format. Extracting entities from free-text descriptions with further vocabulary providing can help to make data the most extensive.

SAM is a GATE-based modular pipeline allowing to extract semantic information from texts, like for example keyphrase extraction and named entity recognition. SAM offers support for German and English.

Named Entity Extraction

- Automatic extraction of named entities from unstructured text;
- Recognizes e.g. persons, places, dates, organizations, and web addresses;
- Can use Sparql queries as source for the named entities.

Datasets

This section introduces the datasets used to evaluate the tools. Each partner was asked to provide two datasets. The datasets were then tested to see whether or not they could be enriched by the tools. The selected datasets were representative of the objects (content, metadata and language of the metadata) the partners have available for Europeana. In most cases, the datasets were published already on the Europeana website.

	EKT	**Kulturpool**	**DDB**	**FINNA**
What is the language of your dataset?	EN, EL	EN, DE	DE (maybe EN)	FI, partly EN (doctoral theses), FI, partly SV (photographs)
How many records are in your dataset?	11.339	220.597	800	4.659
Is your dataset	No	Yes	Yes	published in

published?				Finna, but not in Europeana
Is your dataset available via OAI-PMH?	Yes	Yes	FTP	No
What schema(s) is your dataset available in?	oai_dc, ese	edm	edm, dc	dc

DDB datasets

1. BINE Dataset: a dataset of born digital scientific publications offering information on the energy research funded by the Federal Ministry for Economic Affairs and Energy (BMWi) until (and including) 2018, in particular in the fields of energy efficiency technologies and renewable energies.
    - mediatype text (PDF)
    - Language: German and English language.
    - Format: Dublin Core
    - Number of records: about 700
    - No language attributes
    - No enrichment

2. Bavarikon Dataset: the set includes manuscripts of the Kaiser-Heinrich-Bibliothek Collection. It was a result of a digitization project within the University of Bamberg in cooperation with the Bayerische Staatsbibliothek München. The set was provided in EDM to the DDB, and then it was mapped to the internal EDM-DDB format. The set consists of around 100 records.

EKT datasets

1. Historical Archives of Greek Refugees (H.A.G.R.) of the Municipality of Kalamaria collection. The aim is to preserve the historical memory and present the identity of Greek refugees in modern history. The activities are mainly concerned with the large wave of refugees of the Asia Minor Catastrophe (1922) and the Treaty of Population Exchange (1923-24) from the regions of the Black Sea (Pontus), Asia Minor and Eastern Thrace, in the Ottoman Empire. The dataset provided consists of 7.093 records in OAI Dublin-Core and ESE. The metadata language was in majority Greek with English values for dc:type. Prior to this task the metadata were enriched by EKT in terms of type, chronological periods and subjects[23].

[23] https://www.searchculture.gr/aggregator/portal/search?temporalSearchMode=EKT_HISTORICAL_PERIOD&page.page=1&providerShortName=%3aiape&_strictPeriods=on&language=en

2. The American School of Classical Studies at Athens provides graduate students and scholars from a consortium of about 190 North American colleges and universities a base for research and study in Greece with programs in classical archaeology, classics, linguistic studies, Byzantine, Ottoman, and modern Greek studies, archaeological sciences, political science, history, and other social sciences. The repository publishes the photographic collection that documents the field activities of the American School from its establishment in 1881 until WWII.The dataset provided consisted of 4.246 records in EDM. The metadata language was English. Prior to this task the metadata were enriched by EKT in terms of type, chronological periods and subjects[24].

FINNA datasets

1. FINNA has provided two datasets, the first is a subset of doctoral theses of the University of Oulu. The subset consists of theses written in English and the topics mostly represent medical and natural sciences. No prior enrichments have been made to the dataset.
    - Language: Descriptions are partly in English (titles and subjects), partly in Finnish.
    - Format: Dublin Core
    - Number of records: 2.246

2. The second dataset consists of photographs collected from family albums, depicting life in Finland in the 20th century. No prior enrichments have been made to the dataset.
    - Language: Finnish
    - Format: Dublin Core
    - Number of records: 2.413

Kulturpool datasets

1. The Austrian Theater Museum presents exhibitions on the major topics in theater history - from spoken theater to dance, from puppetry to film and from pantomime to opera. It gathers hand drawings, stage and architectural models, photos, souvenirs of famous actors, authors and composers, autographs etc.
   Languages: German, English
   Format: EDM (accessible by OAI-PMH interface) Number of objects: 23.096
   This set was improved by using as much info as we could extract from raw data (provided by the museum), f.e. Subject, creator’s profession, creator’s biography, creator’s GND link were added.

[24]https://www.searchculture.gr/aggregator/portal/search?providerShortName=%3AASCSA&_strictPeriods=on&temporalSearchMode=EKT_HISTORICAL_PERIOD&page.page=1&sortByCount=false

2. The MAK – Austrian Museum of Applied Arts, Vienna is one of the most important museums of its kind worldwide. It contains collections of applied arts, design, architecture, and contemporary art which has been developed in the course of 150 years.
   Languages: German, English
   Format: EDM (OAI-PMH interface). Number of objects: 197.501
   This dataset was improved by using as much info as we could extract from raw data (provided by the museum), e.g. type, creator's profession, creator's biography, English translations, and temporary entities were added.

Evaluation of enrichments

This section takes an in-depth look at the enrichments that were performed on the datasets from the previous section. The process was different for each of the tools as each of them had different ingestion methods. Additionally, there were incompatibilities between a number of datasets and tools. This section goes over each of the datasets and describes the enrichments (or attempts thereto) on what challenges are faced when using enrichment tools outside of their intended domain.

In order to evaluate how well the tools perform with the datasets, the data providers have performed evaluations on the enrichments carried out on their respective datasets. The enrichments were done by analyzing the selected fields, and connecting them to semantic resources in the form of linked data vocabularies, a process which makes metadata more valuable. It disambiguates the value of the field by adding a link which points to an entry in a vocabulary. The grant agreement foresees that the enrichments are done by creating links to semantic resources as linked data. The tools achieve this using a number of Linked Open Data (LOD) vocabularies[25]. The LOD vocabularies used are listed below, references to the vocabularies can be found in the section 'Inventory of Tools':

- EKT Item Types
- EKT Unesco
- Historical Periods
- Gemeinsame Normdatei (GND)
- Geonames
- Wikidata

DDB dataset enrichment

[25] https://lov.linkeddata.es/dataset/lov/

Enrichments carried out by Kulturpool

In both sets provided by DDB tags were extracted from the description field. The tool processed the dataset and extracted keywords and terms that it sees as representative. The keywords that are created are not linked to a semantic resource and the created XML does not adhere to any specific data format. The terms are added inline to the incoming metadata. This makes the tool very versatile as it is based on plain-text rather than any particular data format. However, this also makes it more difficult to process the output.

Overall, the results seem meaningful. Especially the extracted keywords have a high degree of accuracy. In the first dataset some abbreviations are incorrectly tagged as organization. Additionally, there are some errors in the locations when faced with ambiguous German words.

EKT dataset enrichment

Enrichments carried out by Kulturpool

Upon inspection of the provided metadata it seems that the tool extracted phrases from dc:description and replicated them as keywords. As SAM is not an enrichment tool and no vocabulary was used, those keywords are not LOD, so no contextual classes were added to EKT content. Since the extracted terms were already searchable in the description field the content discoverability has not improved either.

Enrichments carried out by DDB

Looking at the provided metadata it appears that dc:subject, dc:type and dc:creator were enriched when possible using wikidata items. The representation of these enrichments however are not in separate contextual classes, so they can not be regarded as a valid EDM model. This issue could be addressed by using XSLT to add the enrichments as contextual classes.

As for the qualitative analysis, the accuracy of the enrichments was impressive, especially since the provided metadata language was Greek. If the enrichment was presented in a valid EDM model, the utility of the record would certainly be improved especially in terms of multilinguality.

FINNA dataset enrichment

Enrichments carried out by Kulturpool

In the first dataset (Doctoral theses) the tool has extracted information both from record descriptions and titles. Both real person names and abbreviations with capital letters have been interpreted as persons. The extracted locations seem to be meaningful. Some extracted English keywords are meaningful but others too general to be helpful, and some keywords are gibberish. The extracted organisations seem to be mostly names or words other than real organisations.

In the second dataset (Photographs), the dataset doesn’t contain many descriptions so the tool has extracted information only from record titles. Some person names were recognised correctly but most extracted keywords do not make sense. The tool does not seem to work well with Finnish language, not recognising inflections or special characters.

Enrichments carried out by DDB

The enrichment tool has recognised some place names and enriched them with Wikidata links. These links seemed to be correct. Place names with encoded special alphabets (e.g. ‘ä’) have not been recognised. License texts have been enriched with correct license URIs. Some person names have also been enriched with Wikidata links to persons with similar names, but we can’t know for sure if these actually refer to the same persons. For some reason the contents of dc:creator field have been mapped to dc:date. This most likely happened in error.

Kulturpool dataset enrichment

Enrichments carried out by DDB (OpenRefine)

Vocabulary links were added for two fields: dc:subject and dc:creator fields. These enrichments were meaningfully added to the data. The structure of the result was not exactly the same EDM as used on the side of UMA. For example, the subject value was placed in skos:prefLabel field of edm:ProvidedCHO which makes it slightly confusing, as it could belong to either context. However, as discussed at the OpenRefine webinar[26], it is possible to apply a template to correct the alignment between the different EDM formats.

Vocabulary links were extracted for the fields dc:subject and dc:creator, which enriched the data meaningfully. A randomly selected set of records was checked for correctness of the vocabulary links, all of which were correct.

[26] https://pro.europeana.eu/event/europeana-common-culture-webinar-increasing-raw-data-quality-using-openrefine

The structure of the result was not exactly the same EDM as on the UMA side. For example, subject value was placed in skos:prefLabel field of edm:ProvidedCHO which makes it confusing to what context it belongs to. But, according to the OpenRefine webinar, we can set our own template for correct transformation. Some vocabularies were extracted from keywords, f.e. edm:Place was taken from Creator label - but this place can belong to Creator, not to Object. Such cases can produce "dirty" data, so have to be checked first. The dc:type vocabularies correspond to original values.

Enrichments carried out by EKT (Semantics.gr)

The datasets were harvested and ingested[27] by Semantics.gr. In the first set (Theatermuseum Wien), the field dc:type did not contain English language text, so the tool was not able to enrich that field with type enrichment. The contents of the dc:subject field was in English so the enrichment there was done by subject enrichment. However, the amount of distinct values was rather low (only six distinct values), and the field was not filled in all cases. Therefore, only subject enrichment was carried out[28]. The ingestion and preprocessing of the data was automated, and the following process of analyzing and assigning subjects took an approximate time of two minutes. The automated enrichment was adjusted to better reflect the subject of the dataset.

Unfortunately, the tool was unable to automatically process the second set[29]. Therefore, it was necessary to manually curate the set, which took an approximate eight hours. This is quite a lot of time, especially in contrast to the previous set. This illustrates the variety in time spent processing various datasets. Although much of this time was spent correcting suggested enrichments by manually curating them[30]. The usage of preselected mappings in the form of XSLT mappings could reduce the amount of time spent by providing a peer-reviewed version-controlled mapping. Further reasons for using generic mappings became apparent when differences were found between the structure of the EDM that was used by EKT and the structure of the EDM on the side of UMA. As the field dc:subject is not present, subject enrichment could not be carried out.

Dc:type was enriched with vocabularies, main link in Greek, but the English vocabulary is also present in the entity, as shown in the code snippet below:

```
<skos:Concept rdf:about="http://semantics.gr/authorities/ekt-item-types/fwtografia">
```

[27] https://aggregator.ekt.gr/acceptance-oai/request?verb=ListRecords&set=kulturpool&metadataPrefix=edmNoEnrichments

[28] https://aggregator.ekt.gr/acceptance-oai/request?verb=ListRecords&set=kulturpool&metadataPrefix=edm

[29] https://aggregator.ekt.gr/acceptance-oai/request?verb=ListRecords&set=kulturpool2&metadataPrefix=edmNoEnrichments

[30] https://aggregator.ekt.gr/acceptance-oai/request?verb=ListRecords&set=kulturpool2&metadataPrefix=edm

```
<skos:prefLabel xml:lang="el">Φωτογραφία</skos:prefLabel>
<skos:prefLabel xml:lang="en">Photo</skos:prefLabel>
<skos:broader
rdf:resource="http://semantics.gr/authorities/ekt-item-types/disdiastata-grafika"/>
<skos:narrower rdf:resource="http://semantics.gr/authorities/ekt-item-types/fwtoarnhtika">
<skos:narrower
rdf:resource="http://semantics.gr/authorities/ekt-item-types/fwtografikh-diafaneia"/>
<skos:exactMatch rdf:resource="http://vocab.getty.edu/aat/300046300"/>
</skos:Concept>
```

Recommendations

Based on the results described above there are a number of recommendations and lessons learned. This section presents them, as well as describes what future work could benefit the enrichment tools.

Early on in the project the discovery was made that the tool developed by FINNA would not be able to be tested by using datasets provided by the partners in this activity. The reason for this was that the enrichment process couldn't be decoupled from the main workflow tool of FINNA and be applied to external datasets. A lesson might be derived from this, in the sense that it is useful when designing such systems to keep in account modularity in order to make systems more reusable and modifiable. Designing the enrichment tool in a modular way could make it easier to expand the supported metadata format. It would make it easier to change the preprocessing module to one supporting another metadata format if the pre-processing logic and enrichment logic were completely separated from each other. However, an open question not answered by this approach would be which metadata model is best to use as a generic metadata model in between the processing modules and enrichment tool. In the case of enrichment within the Europeana Network, it makes sense to use EDM.

Furthermore, such modularity can be supported even further by introducing standardization before and/or after enrichment (depending on the intended application). This brings us back to the repository of XSLTs referenced previously. The application of XSLTs can support this standardization step by separating the processing (enrichment) from the pre-processing (standardization). Pre- and post-processing of enriched data can be used as a way of standardizing the metadata being processed by enrichment tools. Similarly, Europeana maintains two schemas of their own data model, EDM-internal and EDM-external. This ensures that the tools that are used within the Europeana portal can be confident to always use the same strict schema on the data. At the time of writing, there are a few mappings and scripts available, however, the process of collecting mappings will continue.

Apart from the inability to separate out the enrichment process from the FINNA workflow, another limitation to the available tools was discovered later. During the selection of the datasets it became apparent that thesauri used by EKT are specific to cultural heritage and cannot cover scientific disciplines such as those of the PhD theses described in the FINNA dataset. Semantics.gr can do subject enrichment against a vocabulary that they created based on a subset of the UNESCO Thesaurus where they added Greek translations. The Vocabulary contains 1500 terms but is focussed mainly on Culture, Art, History and Social and Human Sciences. In order to successfully enrich this set, a large and generic scientific vocabulary and import it into the tool. However, this would cost a lot of time and was not within the scope of this task.

One of the observations made during the preparatory phase as well as evident from the results is that the relative domains of the dataset and vocabulary should match in some respects for the resulting enrichment to be relevant. A mismatch between dataset and vocabulary domains can lead to fewer or wrong enrichments. This step involved much custom work as fields were manually selected for enrichment. The enrichments were evaluated using input from each of the dataset providers as they are best placed to evaluate how valuable the enrichments are. One of the observations from the initial phase of the project was the amount of effort involved in the exchange of the data between partners for the purpose of the experiment and the various technical challenges that process entailed. Participants were limited by the features of the metadata regarding format, language and availability. Another important observation was the matter of configurability of a tool. In other words, how well can the tool be reconfigured to use a different vocabulary. This is something that must be taken into account early on in the software development process whenever developing such a tool.

Task 3.3. Outcomes

In the course of this activity the participants attempted to enrich a large number of records by using the tools the partners had available. The objective was to see how the tools performed when used with datasets different from the datasets they were originally designed for. As was revealed by the enrichments results, there was no dataset which could be enriched by all three of the tools. This is a clear indication that most tools cannot be applied right away to data created outside of the domain the tools were intended for. It would still require a lot of customisations for them to function in such situations.

The outcome of the enrichments was diverse, with some tools adding meaningful enrichments, while others were not able to process the datasets at all. Generally, the most useful enrichments resulted from the tools being customised the most. Eventually, the enrichments were made only for the benefit of this activity. The results provide insight into the versatility of the tools.

Looking at the objective of task 3.3 to create an inventory of semantic resources and select datasets to be enriched, this was done for the tools and a diverse set of datasets provided by the partners in this task. Furthermore, the results as a whole were shared for review by members of the Europeana Data Quality Committee. Originally, the grant agreement called for the development of production-ready software. This outcome was amended once it became clear that it would be impossible to develop a new application without determining which enrichment tools are currently used by participants on the data they published into the Europeana portal as well as which aspects of the enrichment tools should be present in such an application. It can be argued that the selection of XSLTs is a more valuable addition to the tools than the development of another enrichment tool, as it could be integrated into a processing pipeline easily without abandoning the existing architecture. Also, as shown by the resulting evaluations, many partners have developed software which is very good at performing enrichments in their own domain. The findings presented in this report indicate that a generic tool would perform worse than an enrichment tool which was developed for a specific task.

Overall Activity 3 Conclusions

The Activity 3 objectives as outlined below, were met or are expected to be met in a timely manner, with some carefully considered deviations for Task 3.3.

- Raise 4 million of existing content on Europeana Collections to at least Tier 2;
- c. 1.7 million new records are delivered, complying with Tier 3 or 4 specifications;
- Complete the transition to EDM of the metadata structure that are still using ESE;
- Enhance multilingual features by widening the adoption of multilingual vocabularies;
- Enrich the metadata of at least 8 datasets by connecting to semantic resources available as linked data.

The participating NAs improved 4,440,533 records to tier 2+ (111.01% of the target) and provided 2,008,486 new records in tier 3+ (118.15% of the target). The full transition to EDM was completed, with all updated and new datasets provided as EDM.

The enhancement of multilingual features by widening the adoption of multilingual vocabularies was, on one hand, part of the data quality improvement of some NAs, and, on the other hand, a practical recommendation of task 3.3 that will help further multilingual enrichment endeavours. The enrichment work in task 3.3 was done for eight datasets, evenly distributed among the participating NAs. The results were shared with a user group of about 20 target stakeholder representatives from the Europeana Data Quality Committee.

While the general task 3.3 objectives were met, there was a deviation in the task in regards to the development of semantic enrichment production-ready software. As was elaborated on in the Semantic Enrichment Processes section, existing partner's technology is often specifically designed for a certain task or domain, which was extensively evaluated in the task. Therefore, instead of one standalone software solution, a more relevant recommendation is to further build on a repository of XSLTs. These can then be adopted and integrated into an aggregation pipeline without abandoning an NAs existing architecture. In the long run, this will further improve metadata mappings with semantic enrichment and therefore the metadata quality tiers.

Another data quality highlight not directly part of the Activity objectives, was the submission of high-quality metadata with 85% of all sets provided by project partners in Tier A or above.

The outcomes of this project make it clear that data aggregation and improving data quality for the Europeana Initiative is not done in a vacuum. This entire process is part of a much broader ecosystem and discussion on policy recommendations at European level with topics that should be addressed on the national level like national digital transformation strategies, frameworks and standards, capacity building, stakeholder engagement and collaboration between aggregators. This is further elaborated in the *Recommendations to the member states on National Aggregation Policies*, part of the Activity 2 report on the *Landscape of National Aggregation in Europe*.

In the coming years the Europeana Initiative will focus even more on the digital transformation of CHIs across Europe[31] as the latter is crucial to make further work on data quality possible. CHIs play a crucial role in this process; since e.g. digitisation work resides with the CHIs, parties like EF and NAs cannot achieve all the improvements they would like to make without CHIs, no matter how solid or robust the methodologies and measurement tools, such as the Data Quality Plans. In this project it became clear that NAs are dependent on the cooperation of CHIs and what they are able to offer, and that CHIs expect clear incentives in order to improve their data. These points are further addressed in the *Sustainability Plan* of the project.

[31] https://pro.europeana.eu/page/strategy-2020-2025-summary