EUROPEANA SOUNDS

Project Number: 620591

D2.2 Functional design of semantic enrichment

Document Identifier:	EuropeanaSounds-D2.2-Functional-design-of-semantic-enrichment-v1.0.docx
Document link:	http://pro.europeana.eu/web/europeana-sounds/documents
Date:	12/12/2014

Abstract

This Deliverable contributes to the specification of functional requirements and the technical design for the crowdsourcing infrastructure that will be developed within WP2. This deliverable describes these functions from the point of view of the software that will be developed in different project scenarios within the Europeana Sounds project. The classification of User Stories led to the identification of three different scenarios that take into account the environment in which the end users interact with the resources of the object: namely creating annotations. This deliverable contains a system architecture overview, which presents the components used to build the workflows for the creation and modification of various annotation types. Finally this document mentions the main LOD sources that can be used for semantic enrichment and possible strategies for assessing the quality of the user's contribution to the enrichment of the objects.

Co-funded by the European Union
Europeana Sounds is coordinated by the British Library

The project is co-funded by the European Union, through the **ICT Policy Support Programme** as part of the **Competitiveness and Innovation Framework Programme (CIP)**.
http://ec.europa.eu/information_society/activities/ict_psp/

Dissemination level		
P	Public	X
C	Confidential, only for the members of the Consortium and Commission Services	
I	Internal, only for the members of the Consortium	

I. COPYRIGHT NOTICE

II. REVISIONS

Version	Status	Author	Partner	Date	Changes
0.1	ToC	Alessio Piccioli, Sergiu Gordea	Net7, AIT	21/11/2014	
0.2	First draft	Maarten Brinkerink, Remy Gardien	NISV, EF	24/11/2014	Section 2 and 5
0.3	Second draft	AP, SG, MB, RG, Hugo Manguinhas	Net7, AIT, NISV, EF	28/11/2014	Bulk of sections added / revised
0.4	Third draft	AP, SG, MB, RG, HM	Net7, AIT, NISV, EF	5/12/2014	Revision of all sections
0.5	Fourth draft	Breàndan Knowlton, Vassilis Tzouvaras	Historypin, NTUA	8/12/2014	Added sections 2.3 and 2.4
1.0	Final version	AP, MB	Net 7, NISV	10/12/2014	Final - following review

III. DELIVERY SLIP

	Name	Partner/WP	Date
Document Author	Alessio Piciolli piccioli@netseven.it Maarten Brinkerink mbrinkerink@beeldengeluid.nl	Net7 / WP2 NISV / WP2	 12/12/2014
Reviewed by	**Reviewers:** Pavel Kats Jon Voss	 EF / WP4 Shift / WP2	 9/12/2014 9/12/2014
Approved by	**Coordinator & PMB**		12/12/204

IV. DISTRIBUTION

No.	Date	Comment	Partner / WP
1	12/12/2014	Submitted to the European Commission	BL / WP7
2	12/12/2014	Posted on Europeana Pro	BL / WP7
3	12/12/2014	Distributed to the Project Consortium	BL / WP7

V. APPLICATION AREA

This document is a formal output for the European Commission, applicable to all members of the Europeana Sounds project and beneficiaries. This document reflects only the author's views and the European Union is not liable for any use that might be made of information contained therein.

VI. DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors named in the Delivery Slip.

VII. TERMINOLOGY

A complete project glossary is provided at the following page:
http://pro.europeana.eu/web/guest/glossary

Further terms are defined below as required:

TERM	DEFINITION
AB	Advisory Board
APEX	Archives Portal Europe network of excellence
EC-GA	Grant Agreement (including Annex I, the Description of Work) signed with the European Commission
GA	General Assembly
PC	Project Coordinator
PI	Performance Indicator
PM	Project Manager
PMB	Project Management Board
PSO	Project Support Officer
TEL	The European Library
TD	Technical Director
UAP	User Advisory Panel
WP	Work Package

VIII. PROJECT SUMMARY

Europeana Sounds is Europeana's 'missing' fifth domain aggregator, joining APEX (Archives), EUscreen (television), the Europeana film Gateway (film) and TEL (libraries). It will increase the opportunities for access to and creative re-use of Europeana's audio and audio-related content and will build a sustainable best practice network of stakeholders in the content value chain to aggregate, enrich and share a critical mass of audio that meets the needs of public audiences, the creative industries (notably publishers) and researchers. The consortium of 24 partners will:

- Double the number of audio items accessible through Europeana to over 1 million and improve geographical and thematic coverage by aggregating items with widespread popular appeal such as contemporary and classical music, traditional and folk music, the natural world, oral memory and languages and dialects.
- Add meaningful contextual knowledge and medium-specific metadata to 2 million items in Europeana's audio and audio-related collections, developing techniques for cross-media and cross-collection linking.
- Develop and validate audience specific sound channels and a distributed crowd-sourcing infrastructure for end-users that will improve Europeana's search facility, navigation and user experience. These can then be used for other communities and other media.

- Engage music publishers and rights holders in efforts to make more material accessible online through Europeana by resolving domain constraints and lack of access to commercially unviable (i.e. out-of-commerce) content.

These outcomes will be achieved through a network of leading sound archives working with specialists in audiovisual technology, rights issues, and software development. The network will expand to include other data-providers and mainstream distribution platforms (Historypin, Spotify, SoundCloud) to ensure the widest possible availability of their content.

For more information, visit http://pro.europeana.eu/web/europeana-sounds and http://www.europeanasounds.eu.

IX. STATEMENT OF ORIGINALITY

This document contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

X. EXECUTIVE SUMMARY: D2.2 FUNCTIONAL DESIGN OF SEMANTIC ENRICHMENT

This deliverable contributes to the specification of functional requirements and the technical design for the crowdsourcing infrastructure that will be developed within WP2. The purpose of this deliverable is to describe the relevant scenarios for the Crowdsourcing Infrastructure and to provide a preliminary functional design for the functionality that is required to support these scenarios.

The classification of User Stories led to the identification of three different scenarios that take into account the environment in which the end users interact with the resources of the object: namely creating annotations. These three scenarios are:

1. Manual creation of annotations
2. Manual correction of semantic annotations
3. Crowdsourcing through specialised platforms

This deliverable contains a system architecture overview, which presents the components used to build the workflows for the creation and modification of various annotation types. The system architecture is divided between "Client/Render" (what will be implemented as a user front-end, to perform crowdsourcing actions), "Services" (the back-end services for processing the results of crowdsourcing) and "Storage & Index" (where the resulting data will be stored and accessed from).

The document describes the User Stories and Epics related to the three scenarios. These User Stories and Epics are based on the previous work documented in D2.1, followed by a technical analysis. The User Stories were analysed by identifying the subjects involved, and mapping them to the Open Annotation terminology. Then they were linked to the crowdsourcing scenarios. The next step was the definition of the corresponding functions. And the last step was the assignment of a priority. By applying a general Gap Analysis, this Deliverable also indicates what basic functionality is lacking, in order to enable the User Stories and Epics required for the Crowdsourcing Infrastructure.

The deliverable provides an inventory of main LOD sources that can be used for semantic enrichment. It is based on the previous work that was documented in D1.3 *Ontologies for sounds* and adds web resources which are important in particular application contexts, either for culture vultures or culture snackers and which might be of interest for the crowdsourcing scenarios.

This document concludes with possible strategies for assessing the quality of the user's contribution to the enrichment of the objects. This will be based on previous research in this area and will be part of a Europeana-wide holistic approach to quality assessment of enrichments. Such an holistic approach will be established within a Taskforce on Evaluation Methods for Automatic Enrichments, organized in the context of EuropeanTech. Europeana Sounds will be represented in this Taskforce, in order to include the evaluation of crowdsourced improvements to enrichments to this holistic evaluation approach.

TABLE OF CONTENTS

1 INTRODUCTION

As stated in the *Description of Work* (Ref 1: DoW, Part B, page 9) end-user engagement lies at the very heart of Europeana Sounds. This deliverable contributes to the specification of functional requirements and the technical design for the crowdsourcing infrastructure that will be developed within WP2. The purpose of this deliverable is to describe the relevant scenarios for the Crowdsourcing Infrastructure and to provide a preliminary functional design for the functionality that is required to support these scenarios. This functional design will form the basis for the implementation of the Crowdsourcing Infrastructure within T5.5. The preliminary nature of this design refers to the understanding that current insights may be adjusted during the implementation phase (an intrinsic aspect of the Agile development process, described later in this document). User stories for this crowdsourcing infrastructure have been described in detail in deliverable *D2.1 Crowdsourcing infrastructure and exchange policy* (Ref 2). This deliverable describes these functions from the point of view of the software that will be developed in different project scenarios within the Europeana Sounds project.

In order to write this document all technical partners have been involved in a process that, starting with the creation of a shared vision of the overall architecture of the project (section 2), allowed to classify the User Stories identified in D2.1 (Ref 2) and "convert" them into functions. This conversion process is described in detail in section 3.1.1. Among other things, this has led to an initial Gap Analysis[1]. All identified gaps will be further investigated, scoped and detailed in Task 5.5.

The classification of User Stories led to the identification of different scenarios that take into account the environment in which the end users interact with the resources of the object: namely creating annotations. These scenarios are described briefly in section 1.2, and more detailed descriptions are provided in sections 3.2 and 3.3.

Finally this document mentions the main LOD sources that can be used for semantic enrichment (section 4) and possible strategies for assessing the quality of the user's contribution to the enrichment of the objects (section 5).

1. Users

Enrichment through crowdsourcing, as supported by WP2, will be designed in the form of “micro-tasks” (Ref 1, page 10) for two types of audiences; the general public and experts. As described in *MS7 End-user contributions defined* (Ref 3), this distinction has been further developed, along the line of Chenchen Shen’s paper (written in the context of the Channels research for WP4) *Design for User Engagement on Europeana Channels* (Ref 4) into “culture snackers” and “culture vultures”.

To summarise, “culture snackers” are perceived as users with a casual interest in Europeana or the type of content it serves. In contrast, the “culture vultures” are perceived as users who work

[1] https://en.wikipedia.org/wiki/Gap_analysis

professionally with the Europeana portal, or are culture enthusiasts for the type of content it serves. In the context of Europeana Sounds, special attention will be paid to users with a specific interest in audio and/or music domains. It is important to realise that culture snackers and culture vultures are not static or mutually exclusive roles. Depending on context, a single user is sometimes a snacker and sometimes a vulture.

Since domain expertise is required in order to be able to make improvements to the existing enrichments, the targeted user group for performing improvements of existing enrichments corresponds to the culture vulture profile. In Chenchen Shen's final report, these users are described as follows:

> *"They are the culture enthusiasts and professionals. They have a strong interest in cultural heritage and probably a good knowledge in a specific area(s). They are likely to work professionally with culture in one form or another, or to be a lifelong culture enthusiast, including researchers, students, professionals and interested laymen. While having a broad general interest a culture vulture has a special interest in, and knowledge of, one or a small number of specific topics, subjects, styles or genres."*
>
> (Ref 4, page 20)

2. Scenarios

By taking into account the classification of users and the individual scenarios described in D2.1, we identified three main categories of application scenarios, which have different requirements. Consequently, we identified that different toolsets should be used in order to implement their specific functionality. These are the three scenarios:

1. ***Manual creation of annotations**:* This category of scenarios focuses on the creation of simple annotation types (e.g. Semantic tagging, Image tagging) by end users. This is expected to happen within the future Europeana Channels (Sounds/Music & Fashion) by employing an embeddable solution based on the Annotorious framework from project partner AIT (see Section 2.3.2).
2. ***Manual correction of semantic annotations**:* This category of scenarios aims at enhancing existing annotations, which might be manually created by users or created using specific tools during metadata ingestion. The Pundit client is the application that provides the functionality needed to implement this scenario (see Section 2.3.1).
3. ***Crowdsourcing through specialised platforms**:* Many potential audiences for digitised heritage content do not spend significant amounts of time on the websites of cultural heritage institutions (Ref 5, Ref 6). Yet, there is significant latent knowledge to be found in knowledge communities, and the project will harness this to enrich heritage collections. Following the principle of putting content where users are, this project will use the

Historypin platform[2] and others (see Section 2.4.1) to link collected traditional music content to the more general crowdsourcing and discussion platforms where knowledge communities are found, then harvesting these links in the form of enrichments to the source objects. Since the Task T2.1.2 effort of partners involved in this scenario does not fall directly under the activities related to this deliverable, detailed documentation will not be provided in this document of the project but developed in subsequent user story documentation in parallel with software development. These user stories will be iterated continuously throughout the project, with the first preliminary evaluation and report made before project month M17 (D2.4) and the final by M30 (D2.7).

[2] https://www.historypin.org/

2 SYSTEM ARCHITECTURE

The objective of this section is to present a general overview of the system architecture for the proposed crowdsourcing infrastructure. This overview has been produced to provide a better understanding of the particular context and the data flows involved in the definitions of annotation scenarios, epics, stories and functionality (see Section 3). The system architecture overview presents the components used to build the workflows for the creation and modification of various annotation types.

The figure below depicts this system architecture:

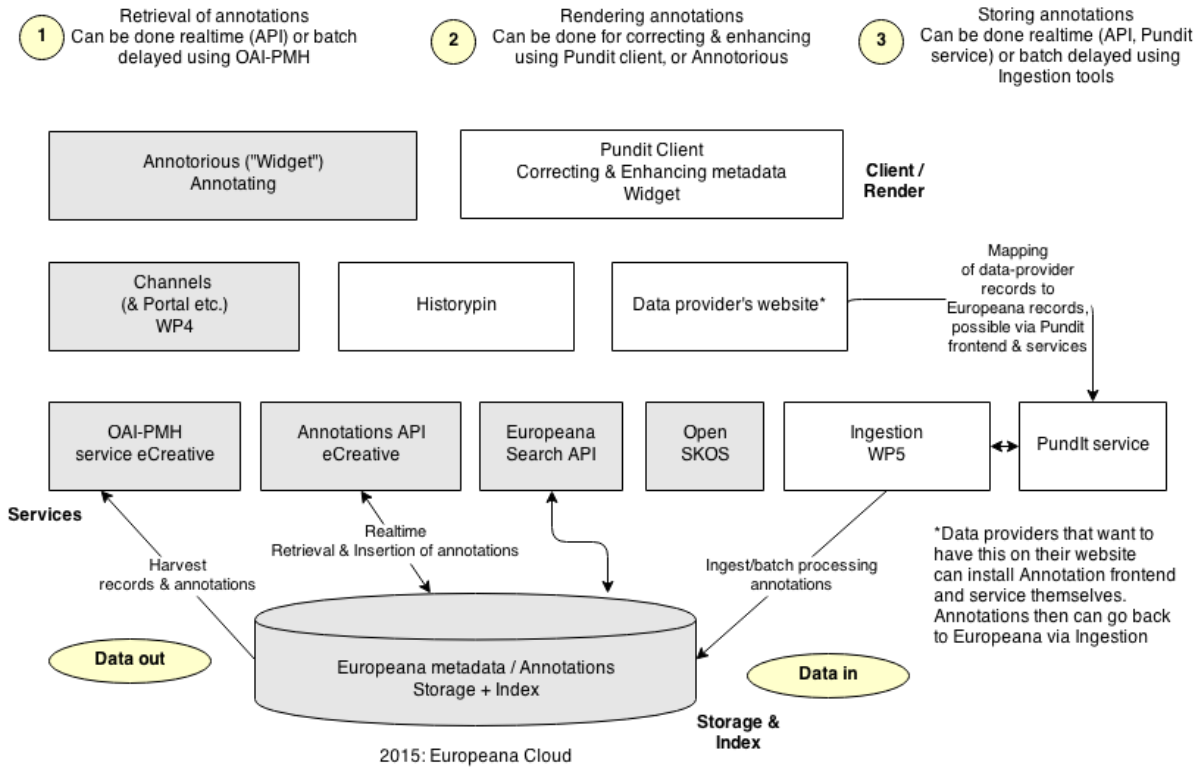


Figure 1: Draft of the system architecture

The system architecture is divided between "Client/Render" (what will be implemented as a user front-end, to perform crowdsourcing actions, see section 2.3), "Services" (the back-end services for processing the results of crowdsourcing, see section 2.2) and "Storage & Index" (where the resulting data will be stored and accessed from).

An outcome of the decision not to be a separate aggregator for the Europeana Sounds project, is that Europeana's instance of data and index will be re-used and no copy will be created to support

indexing of annotations and other crowdsourcing data. The capabilities required to support the type of annotations and functionalities, as defined and scoped in this document, will therefore be developed based on existing tools and platforms.

The Europeana core services include an OAI-PMH service (for mass-retrieval of metadata records), an Annotations API (for real-time inserts and retrieval of annotations), a Search API (for full metadata record search and filtering) and an OpenSKOS API[3] (vocabularies that will support the semantic enrichment). The Ingestion component will support (delayed) mass-ingestion of the results of crowdsourcing. Ultimately, this architecture allows for different "platforms" (Europeana Channels, Historypin, websites belonging to Data Providers) to implement support for crowdsourcing and the different annotation types.

The connections between the platforms and services are not made yet in the diagram as in theory all platforms could work with any service. The expected data flow for this case starts with retrieving metadata records from Europeana, followed by enrichment on one of the platforms, and finally resubmitting the enrichments through Ingestion. Further research and technical exploration is required to make these connections and thus create a final system architecture.

1. Core Architecture

The core foundation of the Crowdsourcing Infrastructure for the Europeana Sounds project is the Europeana storage platform, which manages the metadata describing the cultural heritage objects and the associated annotations. This storage is the persistence layer used by all back-end services.

The storage layer currently consists of data storage (MongoDB[4]) with an indexing layer (Solr[5]). In the next two years, within the course of the Europeana Sound project, the storage layer will be replaced by Europeana Cloud[6]. This means that, for integration, different (Cloud-specific) requirements may come into play at a later phase in the project. Europeana will monitor this and will assess the impact for the Europeana Sounds project and will act accordingly to incorporate these requirements into the software being developed.

To retrieve data from Europeana, Europeana adopted an API-first approach: various APIs have been, and will, be developed in order to support current and future functionality. The Europeana Channels will serve as example implementation of these APIs. These will be developed as part of WP4. Examples of APIs are search and full record APIs, which allow third party developers and Europeana services to search for objects in the Europeana platform and display object details. The Europeana Ingestion process ensures aggregated data is inserted into the Europeana platform.

[3] http://openskos.org/api
[4] http://www.mongodb.org/
[5] http://lucene.apache.org/solr/
[6] http://pro.europeana.eu/web/europeana-cloud

Within the scope of the work for the Crowdsourcing Infrastructure, Europeana's primary focus will be to help to integrate the various components developed within the scope of Europeana Sounds within the Europeana architecture. Within the scope of WP2, Europeana can commit to integrating the following components to its infrastructure, that support crowdsourcing in the form of:

- Tags as annotations (Object/Semantic tagging
- Sets (User collections)
- Image annotations (based on Annotorious)
- Flagging of annotations and flagging of semantic enriched data (i.e. flag incorrect semantic enrichments)
- Importing annotations made on Europeana objects on other platforms (Historypin being pilot implementation of this type of data exchange)

See the epic overview in section 3.1.3, for further clarification of the listed functionality. The first Europeana Channels beta release (mid 2015) will be focussed on Search & Display. All crowdsourcing functionality, such as the ones listed above, are to be added in later iterations in the course of this project implemented in the Europeana Channels as part of the work lead by WP4.

The following components will play a role in supporting crowdsourcing:

Europeana OAI-PMH Service

Europeana's OAI-PMH[7] service allows third parties to harvest Europeana objects en masse. The OAI-PMH service allows harvesting tools to harvest the entire Europeana collection, objects and results of crowdsourcing in accordance with the OAI-PMH protocol. The OAI-PMH Service uses the Europeana API to retrieve full records. This service can be used in the context of Europeana Sounds to enable third-party platforms to retrieve annotations and objects from Europeana on a mass-scale, for instance on a collection basis or even Europeana in its entirety.

The Europeana OAI-PMH service is expected to be ready for production in Q1, 2015.

Europeana Search API

The Europeana Search API[8] is a REST-API[9] that can search and output Cultural Heritage Objects (CHOs) within Europeana. The gap for Europeana Sounds crowdsourcing is:

- The API calls must be able to retrieve annotations and outputs from crowdsourcing

[7] http://www.openarchives.org/pmh/
[8] http://labs.europeana.eu/api/
[9] https://en.wikipedia.org/wiki/Representational_state_transfer

Europeana Channels

Europeana Channels is a user-facing application that is powered by the Europeana APIs. One of the Channels will be the Europeana Sounds Channel, which provides specific functionality for the data related to sound objects (and audio-related objects). The Sounds Channel will be the first implementation of the crowdsourcing tools developed within this Work Package. The details of the Channels concept and work are outlined in WP4, as reported in MS19 Audio Channels First Prototype (Ref 7).

The first version of the Channels application will revolve around Search & Display. All Crowdsourcing features (partly to be developed as an outcome of WP2) will become available in later versions. The front-end framework Annotorious will be used to help users perform the required crowdsourcing features, and will integrate the various components from the project as backend services. In particular the Annotations API will support the real-time processing and retrieval of the different types of annotations.

Europeana Core: Infrastructure and metadata storage

The Europeana Core, including storage and index, must be able to support the crowdsourcing functionality laid out within this Work Package, limited to the scope of functionalities that Europeana can support as listed in section 2.1. The main work will include integrating the various components and coordinating efforts.

Other components that play a role, but will be described later in this document, are:

- Annotations API (AIT)
- Ingestion (NTUA)

Europeana will have a role in integrating these components into its infrastructure, in particular as part of Channels (WP4).

2. Annotations Backend

The Annotations Backend component will be implemented by reusing and extending the User Generated Content service developed within the scope of Europeana Creative[10] project and described in the Europeana Creative D2.2 *Services and Messaging API*. This implements the basic functionality for managing web annotations compliant with the Open Annotation data model[11]. The service implements a REST[12] API and a JSON[13] serialization of annotations. The current implementation supports creation of annotations on Europeana objects or images using simple text comments, tags or semantic tags. An overview of this infrastructure is presented in Figure 2 (below).

[10] http://pro.europeana.eu/web/europeana-creative
[11] http://www.openannotation.org/spec/core/
[12] http://en.wikipedia.org/wiki/Representational_state_transfer
[13] http://en.wikipedia.org/wiki/JSON

The REST API acts as the interface of the Annotation Backend, which encapsulates the Java API that implements the storage and retrieval of annotations. The service architecture is similar to the one proposed within this document (see Figure 2), in the sense that it uses the same client-server separation (i.e. Frontend-Backend functionality) and the same technologies for implementing the storage and retrieval of annotations. However, there is a list of functional Gaps at the technical level:

- In Europeana Sounds the support for the new Web-Annotation[14] model using JSON-LD serialization will be implemented. Also more annotation and body target types will be supported as presented in section 3 and *Appendix 1: D2.2 Functional design of semantic enrichment - Working sheet - From User Stories to functions*.
- Support for the de-referencing of semantic tags and several controlled vocabularies will be implemented
- Support for time-based media annotations is required, however this is only possible when direct links to content are provided (together with the metadata)

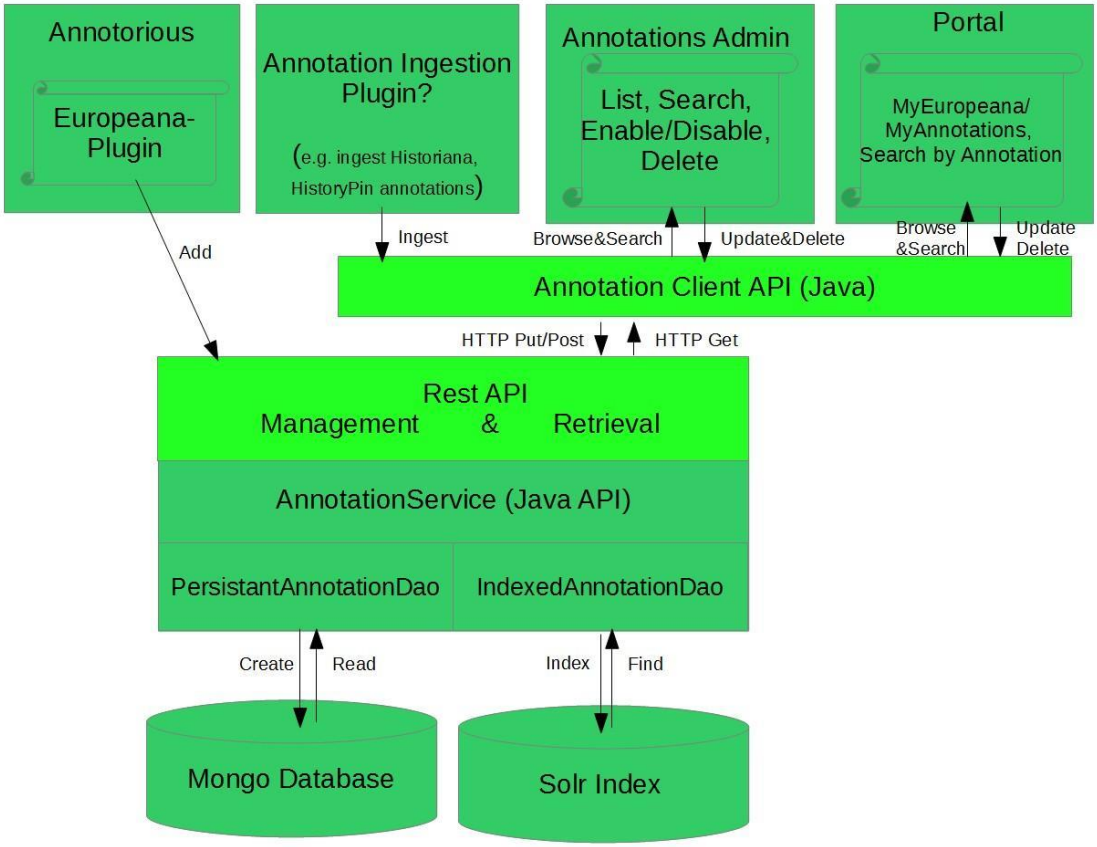


Figure 2: Overview of user generated content service in Europeana Creative

[14] http://w3c.github.io/web-annotation/model_fpwd/static.html

3. Annotations front-end

In this section the "Client Render" component of the system architecture is described. This component will be realised in the form of two types: built-in or widget. In the first case, code is added to a page, which offers the annotation service (creation or modification) in an integrated way, by adding the necessary elements or functions to the page. In the second case it requires the addition of a single piece of HTML code that allows access to a dedicated interface. Both types will be implemented in order to provide effective and efficient solutions for collecting end user contributions annotations from regular users (Annotorius used in the Europeana websites) and from culture vultures and culture snackers (Pundit used in the Data Provider websites).

These two types correspond to the first two scenarios described in section 1.2. They will be further detailed in Section 3. They correspond respectively to the use of the Annotorius software (integrated), and the use of the Pundit tool.

2.3.1 Pundit

The main idea behind Pundit[15] is to enable users not only to comment, bookmark or tag web pages, but also to *create semantically structured data* while annotating, thus enriching the so-called Web of Data[16].

The ability to express semantically typed relations among resources, relying on ontologies and controlled vocabularies, not only enables users to express unambiguous and precise semantics it also, and more interestingly, fosters the reuse of such collaboratively created knowledge within other web applications. For example: to provide a powerful semantic search, build innovative ad-hoc data visualizations which ultimately improve the way users explore the web.

[15] http://thepund.it
[16] http://www.w3.org/2013/data/

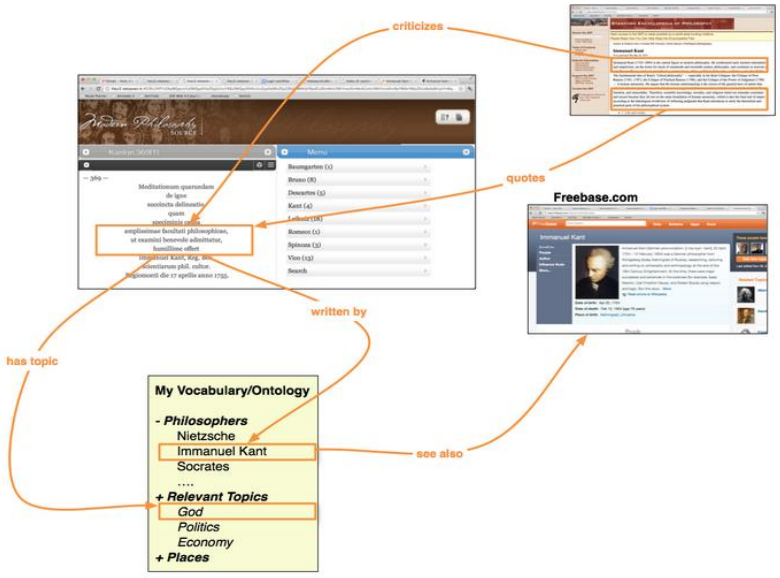


Figure 3: This picture visualises what we mean by semantically structured annotations: the ability for users to create knowledge graphs where web content fragments, concepts and entities are meaningfully connected.

Technically, such a knowledge graph is represented using the Resource Description Framework (RDF)[17], the main building block of the so called Semantic Web[18].

Pundit is a tool for semantic annotation that is developed in a framework consisting of four different components:

- Client: Pundit Client consists of a set of JavaScript libraries and CSS. Version 2 is currently being developed[19] and is based on the Angular.js[20] framework. There will be a first stable release at the beginning of 2015
- Component: feed.thepund.it. The role of this component is to feed Pundit client with annotatable content and configurations. It is a REST application that, through the mechanism of scraping of content (RDF or HTML), is able to generate and display a suitable interface for the creation and modification of semantic annotations.
- Annotation Server: the Annotation Server is the central repository where knowledge is stored. It is based on a triple-store[21] and a relational database to enable fast access to commonly used data. The Annotation Server is a Java based web application for storing and retrieving annotations and notebooks (collections of annotations). RDF is used as the default

[17] https://en.wikipedia.org/wiki/Resource_Description_Framework
[18] http://www.w3.org/standards/semanticweb/
[19] http://dev.thepund.it
[20] https://angularjs.org/
[21] http://rdf4j.org/

data model and a set of REST APIs are provided, both for creating and for consuming annotations (see below).

- Askthepund.it: is a general purpose discovery and visualization tool, with a couple of native visualisations.

The annotation server API[22] services consist of:

- Open API: This API allows clients to access and consume public annotations on the server. API calls do not require authentication, but operate only on annotations that have been explicitly marked as "Public" by their creators.
- Authenticated APIs/Consuming annotations: This API allows clients to read Notebooks and Annotations supporting different formats: plain text (that can also be rendered in HTML pages using stylesheets), JSON, RDF+XML and RDF+N3. All API calls in this set require the implementation of an OpenID[23] workflow at client side. API calls are then authenticated and authorized according to the client identity.
- Authenticated API/Creating and editing annotations: This API allows clients to create new notebooks and annotations as well as edit or delete existing ones and modify their metadata. All API calls require the implementation of an OpenID workflow on the client side.
- Searching API: This set of APIs allows users to search for Notebooks and Annotations based on specific search parameters. The prototype version of the Annotation Server provides only a single search API with simple searching parameters.
- Users API: This set of APIs can be used to obtain all information about a registered user on the Annotation Server, or to perform specific operations regarding user management.

2.3.2 Annotorious

Annotorious[24] is a lightweight software framework, developed in JavaScript, providing support for annotating web resources. The main development stream concentrated on building a simple, intuitive, and embeddable tool for annotating image content, for example, art objects, historical maps or manuscripts. However, it was developed using a plugin based architecture which facilitates adaptations required for handling different types of web resources like text in HTML web pages, music scores or online maps.

The most important non-functional characteristic provided with the tool is a simple mechanism for developing new plugins and integrating them into web applications. These properties position Annotorious as an appropriate solution for integrating it in existing portals for collecting simple semantic enrichments (e.g. tags, image annotations), within a seamless user experience.

[22] http://thepund.it/documentation/pundit-server-api/
[23] http://openid.net/
[24] http://annotorious.github.io/index.html

4. Historypin

Historypin.org is a crowdsourcing platform that focuses on the general problem of geo-tagging content collections. The organisation behind Historypin takes a special interest in any form of user-contributed content within knowledge communities that relates to the development of stronger ties between community members or between community members and local places.

Historypin's infrastructure is powered by Google App Engine and related cloud storage repositories. The front-end of the historypin.org website is written in the Python/Django framework. The platform is widely used on a daily basis with a user base of over 50,000 individuals and 2,100 institutional partners in more than one thousand cities around the world.[25] In Europeana Sounds, Historypin and other popular social platforms will be used to explore how material from sound archives can be made available to knowledge communities, who coalesce around the contemporary performance and transmission of music, specifically traditional music[26].

By making archival content available on websites that are used regularly by practising musicians and knowledge community members ('engagement platforms'), the sound archives can be presented as the authoritative sources that they are, while enhancing the experiences of a broad base of musicians, including those who would not normally be aware of archival institutions and holdings.

2.4.1 Current User Scenario

The specific user scenario to be addressed is a problem, common to traditional music practitioners, concerning music identification. This scenario is often observed in the discussion and music repository site TheSession.org[27], used regularly by 50,000 traditional musicians.

The way this question is generally posed and answered at the moment is as follows:

1. A musician records an unfamiliar tune being played at an informal musical session (usually on their mobile phone).
2. The musician goes home and either uploads the audio/video recording to YouTube or SoundCloud or transcribes it into a simplified musical notation known as 'ABC'.
3. The musician posts the YouTube link or the ABC transcription to www.thesession.org.
4. The musician asks members of The Session for help identifying the tune or tunes in the video or transcription.

[25] On http://yearofthebay.org, Historypin has developed and is testing a prototype UI for the crowdsourcing of metadata pertaining to tags (unstructured), geolocation and date. Initial mapping of this data using the OA/EDM data model is in development, and exports and APIs are being explored as parts of other projects (EV3 as well as some other non-related Europeana projects).

[26] Sometimes grouped within "intangible cultural heritage", reflecting that such musical forms tend to be transmitted through oral rather than written processes, and thus contain inherent ambiguities and variations related to the social process of remembering and transmitting such culture.

[27] https://thesession.org/

5. Helpful community members of The Session will respond with tunes they think match the tune in question, sometimes linking to other tunes on www.thesession.org, sometimes to other YouTube videos and very rarely to an archive record held at traditional music archives, such as the Europeana Sounds partners ITMA, Comhaltas or Tobar an Dualchais. This identification is performed based solely on the memory of the helpful community member, and on what resources they can find through individual research on websites of various archives.
6. The original "searching" musician, following these links, can develop a list of tune names associated with the piece of music, and, using these resources, eventually discover reference recordings of the tune and/or published tune books containing sheet music for the tune. From these sources the musician can learn how to play the tune, and also discover background information such as alternate names, alternate settings (variants) and identification of people or regions strongly associated with the particular piece of music. This furthers the transmission of traditional music and allows the musician to participate more fully in session culture.

An alternate way of discovering the name or context of a tune is to use a specialist application, such as TunePal[28], which performs a statistical match of a live recording against a known corpus of traditional tunes and returns the most likely matches to the user. TunePal is generally accessed via a mobile iOS[29] or Android[30] application, and is typically used directly on the musician's phone during the session. TunePal also supports a browser-based tune detection application at tunepal.org, however, this mechanism is not used as often. TunePal is used regularly by approximately 20,000 musicians.

2.4.2 New User Scenario

Some of this user scenario can be automated and improved through the tools developed in the Europeana Sounds project, as follows:

1. User A (the Searcher) posts a transcription of a piece of music (the Query) to a third-party Platform. This could be Historypin, TheSession.org or another third-party platform. The user requests help from the community in identifying the name or context for the piece of music.
2. User B (the Expert) decides to help, and invokes an automated "Linker" widget provided by Europeana Sounds.
3. The widget queries a web service (for example, TunePal) to identify possible name matches (Candidates) for the transcription.
4. The widget sends each of the statistically likely name matches (Candidates) to the Europeana Search API, limiting the search to pre-selected Data Providers, likely to have

[28] http://tunepal.org
[29] http://itunes.apple.com/us/app/tunepal/id356935033?mt=8&uo=6
[30] https://play.google.com/store/apps/details?id=org.tunepal

matching material.

5. The widget presents a list of potentially matching Europeana Objects to User B (the Expert) with some basic metadata. The user has the chance to preview (listen to, and examine the full descriptive metadata record) each potential match to an audio object, in order to evaluate if the Object matches the original Query transcription. The user selects one or more Objects to be Linked, and optionally provides textual Commentary explaining their choice.
6. The Links are published to the Platform in an easy-to-use format, such that all users can easily follow the Link to the original Object, and also be able to preview any linked audio recordings in-situ.
7. User A (the original Searcher) follows the Links posted by the Expert, immediately finding archival recordings that match the transcription of the musical piece that they posted. User A learns more about the context for these archival recordings and objects by following links to the object pages of the data providers.
8. The widget automatically exposes these published Links and the Commentary in a form that can be ingested as Annotations by Europeana, and potentially also can be queried by the original data provider.

A more sophisticated variant of this scenario is that in which a user posts a new recording rather than a transcription in step 1. In this case, the recording is first automatically transcribed (also using the TunePal web service) before the rest of the steps are followed.

The options for the provisioning of annotation data to Europeana from Historypin or the described widgets, is covered in section 2.5 "Ingestion". A decision on the use of technology and software will be made during the development process, based on the options listed in this section.

5. Ingestion

The role of ingestion is to aggregate metadata and store them in the Europeana servers. We distinguish between two types of metadata in Europeana Sounds. Metadata in the form of XML records harvested from Data Providers repositories and metadata in the form of annotations created by users. Annotations can also be considered enrichments on providers' metadata. Each metadata type will use its own ingestion mechanism and workflow.

Metadata coming from Data Providers will be ingested using the Europeana Sounds aggregation mechanism, allowing for the ingestion of semi-structured data and offering the ability to align and take advantage of well-defined, machine understandable schemes in an intuitive manner. The MINT[31] platform is used to map and transform the metadata into EDM. Repox[32] is providing the harvesting mechanism, based on OAI-PMH technology, for the metadata to be ingested into Europeana servers.

[31] http://mint-projects.image.ntua.gr/sounds
[32] http://repox.sysresearch.org/

Annotations can be of many types and can be created in many ways, within and outside Europeana. Some of the tools that will be built will use their own annotation servers and must have a way to ingest annotations to the Europeana servers. The tools built on top of Europeana's annotations service will use the annotation API. Below we define the two ways that we will provide support for ingesting annotations:

1. OAI-PMH: Annotations will be ingested in batches using the OAI-PMH protocol[33]. This functionality is already covered by the Repox tool. What is left to implement, is the connection between Repox and the Annotation Server. Repox currently is connected to the metadata server. This process will be identical to the one that Data Providers are following to ingest metadata. This type of ingestion is preferable, as it has been tested and can be easily adapted for annotations.
2. RSS: Annotations will be ingested using the RSS protocol[34]. RSS will enable annotations to be ingested automatically in the form of RSS feeds from selected websites.

[33] http://www.openarchives.org/pmh/
[34] http://en.wikipedia.org/wiki/RSS

3 FUNCTIONAL DESIGN: FROM USER STORIES TO FUNCTIONS

1. Introduction

As mentioned in the Description of Work (Ref 1, page 88), the development methodology of the project will be based on the principles of Agile software development method for iterative and incremental development, where requirements and solutions evolve through collaboration between cross-functional teams. It enables adaptive planning, evolutionary development and delivery and encourages rapid and flexible responses to change.

The aim of Agile[35] development is to break tasks into small increments with minimal planning and does not directly involve long-term planning. Iterations are over short time frames that should last from one to four weeks. Each iteration involves a team working through a full software development cycle, including planning, requirements analysis, design, coding, unit testing, and acceptance testing. This minimises overall risk and allows the project to adapt to changes quickly.

In this document we use the terms, "User Stories" and "Epics". These User Stories and Epics are based on the previous work documented in D2.1 (Ref 2), followed by a technical analysis. A definition in accordance with the Agile software development methodology is given[36] below:

User Story:

> *"A user story is simply something a user wants. User stories are more than just text written on an index card but for our purposes here, just think of user story as a bit of text saying something like, "Paginate the monthly sales report" or, "Change tax calculations on invoices." Many teams have learned the benefits of writing user stories in the form of: "As a <type of user> I <want/can/am able to/need to/etc.> so that <some reason>." But it is not necessary that a user story be written that way.*

Epic:

> *"A Scrum epic is a large user story. There's no magic threshold at which we call a particular story an epic. It just means "big user story." I like to think of this in relation to movies. If I tell you a particular movie was an "action-adventure movie" that tells you something about the movie. There are probably some car chases, probably some shooting, and so on. It tells you this even though there is no universal definition that we have agreed to follow, and that an action-adventure movie must contain at least three car chases, at least 45 bullets must be shot, and ….*
>
> *So, "epic" is just a label we apply to a large story. Calling a story an epic can sometimes convey additional meaning. Suppose you ask me if I had time yesterday to write the user stories about*

[35] http://agilemanifesto.org/
[36] http://www.mountaingoatsoftware.com/blog/stories-epics-and-themes

the monthly reporting part of the system. "Yes," I reply, "but they are mostly epics." That tells you that while I did write them, I didn't get the chance to break most of them down into stories that are probably small enough to implement directly."

3.1.1 The Process

In order to prepare this Agile development, we have performed a classification of user stories with the purpose of converting them into functions. The result of this work can be found in *Appendix 1: D2.2 - Working sheet - From User Stories to functions*[37].

Table 1: Description of 'D2.2 Working sheet – From User Stories to functions'

Sheet	Field Name	Field Description
Stories	ID	Identification User Story number (S1,...,SN). This ID number has been added in order to have a short unique way to refer to the user Story.
Stories	As a / an	User Story field, value taken from D2.1
Stories	I want to	User Story field, value taken from D2.1
Stories	Because	User Story field, value taken from D2.1
Stories	Epic Number	Epic reference number used to group Stories by Epic
Stories	Motivation	Open Annotation Field[38]
Stories	Target Type	Open Annotation Field[39]
Stories	Body Type	Open Annotation Field[40]
Stories	Selector	Open Annotation Field[41]
Stories	MAS Function	This field is used to describe the functions corresponding to the user story in the "Manual Annotation Scenario (MAS)"
Stories	MAS Priority	This field is the priority in the context of the implementation in accordance with the Agile development (Must have MH, Nice to Have NTH, Out of Scope OOS). The values entered in this field are the result of discussions that involved directly from one side the technical partners (NET7, AIT) on the other side partners that deal with content (NISV, EF) and took into account the technical feasibility and end-user needs. These assessments were made for the MAS scenario.
Stories	MAS Function	This field is used to describe the functions corresponding to the user story in the "Enrichment Annotation Scenario (EAS)"

[37] The spreadsheet will continue to be used during the development, as a 'living document' but with persistent references, in the form of a Google Spreadsheet, which can be found here: https://docs.google.com/spreadsheets/d/11Rh0aR3eikc_A-k1R0nVp3WW80dL4T_J4IRFpq_7cdU/edit?usp=sharing

[38] http://www.openannotation.org/spec/core/core.html#Motivations

[39] http://www.openannotation.org/spec/core/core.html#BodyTarget

[40] http://www.openannotation.org/spec/core/core.html#BodyTarget

[41] http://www.openannotation.org/spec/core/specific.html#Selectors

Sheet	Field Name	Field Description
Stories	MAS Priority	This field is the priority in the context of the implementation in accordance with the AGILE development (Must have MH, Nice to Have NTH, Out of Scope OOS). The values entered in this field are the result of discussions that involved directly from one side the technical partners (NET7, AIT) on the other side partners that deal with content (NISV, EF) and took into account the technical feasibility and end-user needs. These assessments were made for the EAS scenario.
Epics	ID	Identification Epic number (E1,..., EN). This ID number has been added in order to have a short unique way to refer to the Epic.
Epics	Short name	Brief description (one or two words) of the Epic
Epics	Title	Brief description (one sentence) of the Epic
Epics	Description	Full description of the Epic
Epics	MAS Priority	Same as "User Stories/MAS Priority" for the Epic (see Above)
Epics	EAS Priority	Same as "User Stories/EAS Priority" for the Epic (see Above)

To decide what features will be implemented during the course of the Europeana Sounds project, the following process has been adopted:

1. We analysed the user story by identifying the subjects involved in it, and we mapped them to the Open Annotation[42] terminology (Motivation, Target Type, Body Type, Selector)
2. Once we had agreed on this classification, we decided on which scenarios (MAS, EAS) the individual stories belonged.
3. The next step was the definition of the corresponding functions, by the technical partners involved in the integration of annotation tools.
4. The last step was the assignment of a priority to each individual epic, user story and function by using Must Have (MH), Nice to Have (NTH) and Out of Scope (OOS).

The descriptions of the meaning of the values that we have assigned to the priorities are given below:

Must Have (MH): We start with these stories, and aim to deliver at least a minimal viable product

Nice to have (NTH): We continue with these stories when there are resources left

Out of Scope (OOS): These stories will not be dealt with.

We think that this approach is good, because it allows us to reach a shared understanding of the meaning of the individual user stories through their classification, and consequently it allows us to agree on the prioritisation, while keeping into account the technical feasibility in future implementation.

[42] http://www.openannotation.org/spec/core/

3.1.2 Describing the scenarios

In the following sections the abovementioned scenarios and the corresponding functions that need to be implemented are described in more detail. In particular, it documents the following aspects:

- **Scope:** This sub-section reports the general description of the scenario: type of annotations, the annotation environment, and the tools used for the annotation.
- **Functions:** Here we report a table with the "Must Have" functions that will be implemented in the corresponding scenario. Every row contains the User Story Number (Appendix 1, with the same table from D2.1, with an ID number included) the classification fields (Motivation, Target Type, Body Type, Selector) and the corresponding functions. Some functions are classified with "technical requirement": this means that the corresponding function is already present in another user story, or is already provided by the current version of the software which is used for the implementation (Annotorius / Pundit).
- **Data flow:** In this sub-section we report on the description of the part of the data flow that is required in order to implement the "Must Have" functions described in the previous sub-section.
- **Gaps:** In this sub-section we describe the new features of the tool that must be implemented in order to reach the implementation of the functions described in the sub-section "Functions".
- **Nice to have functions:** Same table as sub-section "Functions", with a description of the "Nice to have" functions.

3.1.3 Describing Epics

The user stories collected in the scope of Task 2.1 were analysed by the project stakeholders in order to identify the functionality required to implement them, and also to determine their relationships (see: *Appendix 1: D2.2 - Working sheet - From User Stories to functions*). There are several user stories that were identified that describe the same functionality, and there are other stories that were identified to represent different views of a more complex story. By following the Agile development model, these related stories were linked to a well-defined Epic[43] [44].

The full list of identified Epics is shown in the Table 2, while the related functions are presented in Sections 3.2.2 and 3.3.2. The last two columns in the table indicate the priority for development of the individual Epics within the scope of manual annotation (MAS) and enhancement of annotations (EAS) scenarios, by using the must have (MH), nice to have (NTH) and out of scope (OOS) conventions.

[43] http://www.mountaingoatsoftware.com/blog/stories-epics-and-themes
[44] http://www.agilemodeling.com/artifacts/userStory.htm#Epics

Table 2: Overview of Annotation Epics

ID	Title	Abstract	Description	MAS	EAS
E1	Complete metadata	As an end-user I would like to indicate specific information of Europeana objects	When I as an end-user see an object that is missing certain specific information, such as the date of creation of an object or the date of birth of an author, I would like to be able to indicate this information through the usage of annotations.	NTH	NTH
E12	Correcting metadata	As an end-user I would like to correct specific existing information of Europeana objects	When as a regular end-user I spot information from an object that in my view is not correct (it is also not part of auto-enrichment) I want to be able to propose a correction through the usage of annotations.	NTH	NTH
E2	Object tagging	As an end-user I would like to tag Europeana objects through controlled vocabularies	I want to be able to tag objects in order to relate objects and enrich their information by adding tags that are made with the help of controlled vocabularies. E.g. I want to be able to tag an object "Paris" when it is a photo taken in Paris, through controlled vocabularies I will be able to select "Paris, France" in order for the system to know that it is the city Paris, in France.	MH	NTH
E3	Linking objects	As an end-user I would like to relate objects to each other and identify their relationship	I want to be able to link two objects together with a certain type of relation in order to group objects together.	MH	MH
E4	Discussion	As an end-user I would like to discuss and comment on Europeana objects	I want to be able to add free text to objects in order to discuss them with other users, such as by having a comment field or a separated forum-like discussion.	OOS	NTH
E5	Like and dislike	As an end-user I would like to have the ability to review/mark/like previously made annotations	I want to be able to "like" or "dislike" a previously made annotation when I see one that I believe can be more or less relevant to other users and make the system able to prominently show the annotations according to their popularity.	MH	MH
E6	Literature enrichment	As an end-user I would like to add bibliographical references and literature to Europeana objects	I want to be able to enrich Europeana objects by adding bibliographical references or literature to objects, by either adding links or domain specific information such as ISBN numbers.	OOS	NTH
E7	Image annotations	As and end-user I would like to annotate images and part of images	I want to be able to annotate part of images with information; I want to do this in a visual way by drawing shapes in order to annotate part of an image to reference my annotation to a part of the image.	MH	OOS
E11	Audio annotations	As an end-user I would like to annotate sounds and selections of sounds	I want to be able to annotate a timed selection of a sound file in order to add references or remarks about a sound	NTH	NTH

ID	Title	Abstract	Description	MAS	EAS
E8	User sets	As an end-user I would like to create my own sets of Europeana objects	I want to be able to create my own sets of Europeana objects, either privately or publicly, in order to group objects by a certain theme or just group the ones that I like or are of a particular interest. I also want to be able to recommend items for other public sets.	MH	OOS
E9	Moderation	As an end-user I would like to suggest the removal of auto enrichment or highlight an enrichment that is very relevant	When I see detailed metadata of a Europeana object that comes through enrichment, and I spot something that does not relate to the object, I want to be able to report that in order to improve the quality of the metadata. On the other hand, when I spot something that is very relevant, I want to be able to highlight this in order for Europeana to tune its enrichment process based on what works and what doesn't work well.	NTH	MH
E10	Administration	As a Channels administrator I would like to be able to moderate annotations	As a Channels administrator or super-user, I want to be able to update or hide annotations that do not meet the terms of use or are not related to the object they are linked to in order to	MH	OOS

2. Manual annotation scenario (MAS) or manual creation of annotations

3.2.1 Scope

The manual creation of annotations scenario (MAS) focuses on collecting information that users would like to add to Europeana in an efficient, lightweight implementation. Therefore, this scenario concentrates on simple annotation functionality like tagging, commenting, providing feedback (e.g. like, rate) that can be easily embedded in the Europeana portal and/or channels. The information provided is strictly related to the content of Europeana objects or to their associated metadata.

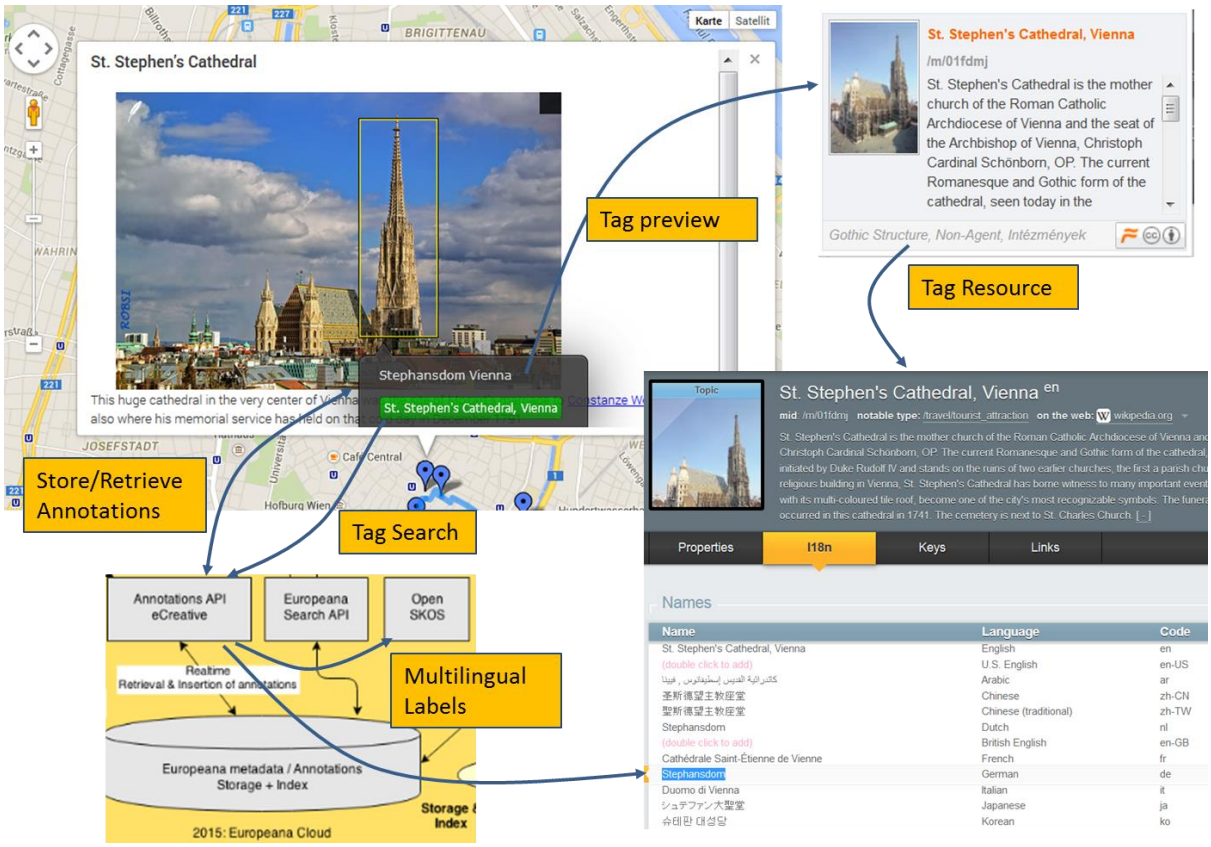


Figure 4: Overview of semantic tagging process

Figure 4 presents a sample scenario, representing the creation of semantic tags which link the items to semantic web resources. Consequently, they facilitate access to rich, multilingual descriptions, regarding the content and the subjects of Europeana objects. This particular use case represents the process of explaining the content of the image by using the Semantic Tagging functionality of Annotorious[45], in combination with the Freebase[46] semantic repository. Adding semantic tags to the image representing the St. Stephen's Cathedral in Vienna is showcased here. The acquisition of other types of information as specified in the next sections and follows a similar process.

[45] http://annotorious.github.io/plugins.html#semantic-tagging
[46] https://developers.google.com/freebase/v1/topic-overview

3.2.2 Functions

Table 3: Description of the functions in the MAS scenario

ID	Epics	Functions	Prio
S6, S55	E2, E11	Tagging Europeana objects with names of artworks managed through controlled vocabularies	MH
S7	E2	Add semantic tags to the subject field of Europeana Objects	MH
S10	E5	The users should be able to mark annotations that are inappropriate or not correct. (Dislike function)	MH
S11	E5	Technical requirements for editing or deleting own annotations	MH
S17	E3	Create collections of related items to be used in third party applications	MH
S18	E3	Add semantic or simple tags to connect Europeana objects that are related to the same thematic or used in an particular context	MH
S21	E3	Add simple tags connecting objects related to a given thematic. Simple tags must have a type	MH
S27	E7	Add semantic tags to Europeana Objects indicating the related media genres	MH
S31	E7	Add annotations to image content that become publicly available	MH
S33	E8	Create user collections by providing a set of metadata regarding the purpose of the collection	MH
S34	E8	Create a user collection that will aggregate only sound content	MH
S37, S38, S42	E8	Publishing (private) user collections	MH
S41	E8	Define multilingual labels for User Created Collections	MH
S43	E8	Changing the visibility of the User Collections between private/group/public	MH
S1	E1	Suggest dates to add in the empty metadata fields of Europeana objects.	NTH
S2	E1	Tagging Europeana objects with time periods managed through controlled vocabularies	NTH

ID	Epics	Functions	Prio
S4	E1	Tagging Europeana Objects with location where the Media object was produced	NTH
S5	E1	Tagging Europeana objects with place names (mentioned in metadata or media objects) managed through controlled vocabularies	NTH
S8	E7	Tagging Europeana objects with names of persons (mentioned in metadata or media objects) managed through controlled vocabularies	NTH
S25	E8	Add feedback on Europeana Objects in form of Rate (binary or 1-5 scale) and a short explanation	NTH
S26	E1	Add semantic tags to Europeana Objects indicating the related media genres	NTH
S29	E1	Add tags to Europeana objects indicating their media genres	NTH
S36, S40, S44, S47	E8	Add Europeana Object to User Collections that can be accessed only by a given group of users	NTH
S48	E8	Add feedback on user collections in form of like/dislike	NTH
S49		Show the profile picture of the annotation creators in the graphical user interface	NTH
S50	E5	Technical requirement to sort by popularity	NTH
S52		Add a Rate and a short comment to music items. (similar to S25)	NTH
S54	E9	Add support for indicating automatic enrichments that are incorrect	NTH

3.2.3 Data flow

The functionality for the Manual Annotation Scenario, as identified in the previous section can be grouped by the motivation that drives the users to provide semantic enrichments of Europeana objects, into categories that include: *Tagging, Linking, Commenting, Feedback, Curating*. Except for *Curating* user stories, which include a more complex process, the other annotations workflows are fairly similar to the sample scenario presented in Figure 4 (Section 3.2.1). All of them aim at providing small informational texts, including (or not) links to semantic concepts identified by their URIs. The main difference lies in the semantic and the structuring of “annotation bodies”. However,

the information flows exchanged between different components of the system follow the same paradigm.

The blue arrows present in the Figure 4 indicate the information flows involved in the creation of semantic tags on St. Stephen's Cathedral in Vienna. This included the invocation of the following front-end and backend functionality:

- User assistance for concept selection: the user is provided with support for searching candidates from a list of existing semantic tags through the invocation of the Annotations API.
- Semantic Tag Preview: a preview of individual tags will allow users to select the most appropriate resource to enhance the semantics of the new annotation.
- Search on multilingual labels: given the nature of content stored in Europeana repository and the diversity of the languages used by the web users, it is important that the tag search functionality is able to break the language barriers and open up the crowdsourcing infrastructure for the large and diverse public. This is achieved by making use of controlled vocabularies and semantic web resources, which can expose labels in multiple languages.
- The storage of the new annotations and their retrieval is implemented within the Annotation backend and exposed through a RESTful API, or another method listed in section 2.5 (if deemed more appropriate during development). The new created annotations can be searched by URL of their targets or by free text search.

3.2.4 Gaps

Gaps regarding the implementation of this crowdsourcing scenario include the following functionality that is not yet available in the individual components:

- An Annotorious Plugin for the invocation of the Annotations backend needs to be implemented as an extension of the REST Plugin[47]. This must support the creation and representation of annotations stored in an Web Annotation[48] format using JSON(-LD)[49] serialization.
- A plugin for semantic tagging, using the vocabularies provided by the current project as described in *D1.3 Ontologies for Sound* (Ref 9), are required. The end users must be provided with support for searching appropriate concepts when tagging objects.
- A plugin for handling Europeana Sounds specific content like music scores in handwritten or printed forms. The scans of music scores are special types of images, for which automatic detection of measures or beats may be used when creating annotations.
- De-referencing functionality is needed for supporting free text search on semantic concepts which are identified by their URIs. This implies invocation of LOD repositories in order to

[47] https://github.com/annotorious/annotorious-vanilla-rest-plugin
[48] http://w3c.github.io/web-annotation/model_fpwd/static.html
[49] http://json-ld.org/

retrieve the preferred labels, multilingual labels or abstracts describing the concepts.

3.2.5 Supporting the creation of user collections

There is an increasing interest for reusing Europeana content within the Europeana platform (i.e. portal, exhibitions, My Europeana, Labs), or in third party applications[50]. There is a functional requirement that is common for the most of these scenarios, namely the selection of the content that fits the purpose of these applications. Therefore, an important type of Annotations is the one related to the creation of User Collections. Epic 8 in Table 2 describes the application scenario and the required functionality is presented in Table 3.

However, the process proposed for aggregating User Collections involves more components of the Europeana platform, for example, the Search API, My Europeana and the Content Reuse Framework. Detailed technical requirements are available within the Assembla space used by Europeana for management and tracking of the development process[51].

3. Enrichment Annotation Scenario (EAS) or manual correction of semantic enrichment

3.3.1 Scope

Manual corrections of semantic enrichment (EAS) consist of annotations added by users of Data Provider web portals that provide Data for Europeana Sounds. Such annotations have the following targets:

- the object itself
- the object's metadata as obtained from the process Ingestion process (see section 2.5)
- other annotations previously created by other users of Europeana, or the Data Provider

The resources involved in these types of annotations are resources that must first be loaded into Europeana. The tool used to create annotations of these types is Pundit.

3.3.2 Functions

Table 4: Description of the functions in the EAS scenario

ID	Epic	Functions	Prio
S9	E5	In the annotation preview I use the edit button to propose a new value of the content of the annotation metadata.	MH

[50] Take for example the very successful http://vangoyourself.eu/ , or other examples documented on http://labs.europeana.eu/blog/ .
[51] https://www.assembla.com/spaces/europeana-creative/tickets/43#/activity/ticket

S10	E5	In the annotation preview (sidebar) a button allows to mark an annotation as "TO BE REMOVED BY USER X" or “USER X propose to remove the annotation Y”.	MH
S11	E5	In the annotation preview a button allows to modify the value of the annotation created by the user. Predicate or object.	MH
S18	E3	In the annotation preview (sidebar) a edit button allows to add new tags to the annotated objects. A special tag (label predicate "has_same_author", "has_same_work", "disagree" to specific the link between objects) allows the user to search in the Europeana Sounds provider and create a link between objects.	MH
S27	E7	Technical requirement	MH
S32		Technical requirement	MH
S49		Technical requirement	MH
S1	E1	The values of the metadata model of Europeana cannot be edited directly from the interface of semantic enrichment: the function then again concerns the proposed modification of existing values. Same as S9. It can be considered a technical requirement for specific values of type date.	NTH
S4	E1	Technical requirement	NTH
S5	E1	Technical requirement	NTH
S6	E2	In the annotation preview (sidebar) a edit button allows to add new tags to the annotated objects. CONSIDER THE Following Interpretation: "tag works of art in a way to link the same works together"	NTH
S7	E2	In the annotation preview (sidebar) a edit button allows to add new tags to the annotated objects	NTH
S8	E7	Technical requirement	NTH
S12	E4	In the annotation preview (sidebar) a button open a text area where the user can add a comment on the annotation	NTH
S15	E4	In the "annotation environment" dedicated to the resource, the resource itself will be annotatable. A "Comment" Button will open an area that allows user to add his own comment.	NTH
S16	E6	Add links to a Europeana Objects pointing to related information available in wiki pages (URLs)	NTH
S19	E6	Technical requirement. Same as S11.	NTH

S20	E6	Use the DbPedia Pundit Selector in order to build semantic annotation where the triple has as subject the resource itself, as predicate “has tag”, as object a Wikipedia URI	NTH
S21	E3	Technical requirement. Same as S11.	NTH
S23	E4	Technical requirement: In the annotation preview (sidebar) while the user is adding a comment to a specific element (S16), the comment made by previous users (ordered by date DESC) are displayed.	NTH
S25	E8	Enhanced edit function with template (like + predicate "why I like", comment)	NTH
S8bis	E11	An advanced players allows to build a timespan and activate the triple composer in order to create an annotation where the target is a timespan of the source (source must be available)	NTH

3.3.3 Data flow

In order to perform the manual correction of semantic enrichment scenario, Pundit can be integrated in the Data Provider websites, starting with a pilot implementation (based on selected Data Providers that already meet the technical requirement involved), followed by an open-call and by providing reference material to enable Data Providers to implement Pundit with limited technical support (as part of T2.1.2).

We propose here a data flow where all components in the EAS scenario are involved:

1. Europeana ingests resource (metadata) from the Data Provider. The ingestion creates the link between the Europeana resource ID and the Content Provider Resource ID. This link is stored in the Europeana Storage through the appropriate EDM properties. Through the Europeana Search API it is possible to retrieve the Europeana resource ID.
2. The Data Provider adds the Pundit Client Widget (PCW) to the resource web page. Technically it is HTML code that allows you to add the button “Annotate” to the page.
3. When the user of the Data Provider page clicks on the “Annotate" button, the widget opens an iFrame without leaving the context of the page source. The iFrame shows an environment that allows the user to make manual corrections of semantic enrichment. At this stage the Pundit Client Widget makes the necessary calls: to retrieve the Europeana ID (using Europeana Search API), to retrieve the metadata (using Europeana Search API), to retrieve annotations already imported with the mechanism of ingestion (using Annotation Europeana API), and to retrieve annotations not yet imported to the mechanism of Ingestion (using Pundit Annotation server).
4. When users save a manual correction of semantic enrichment, the PCW will write to the Pundit annotation server (using Europeana resource identifier, URI).
5. Europeana ingests annotations from the Pundit Annotation Server to the Europeana storage and annotations will be removed from Pundit Annotation Server.

3.3.4 Gaps

According to the functions that have been reported as MH in the table in section 3.3.2 the following features will be implemented in the Pundit software:

- Communication between the Pundit Annotation Server and the Europeana server
- Pundit client customisation (annotation environment with feed.thepund.it, configuration, annotation templates, etc.) according to the data model and the functional requirements
- Add the edit function as “user proposes to modify an annotation value”
- Add the remove function as “user proposes to remove an annotation value”
- Add the possibility to create annotation where the object is the annotated resource

4 RELEVANT LOD SOURCES FOR IMPROVING EXISTING ENRICHMENTS

The semantic enrichments that aim at inter-linking Europeana objects, or linking Europeana objects with semantic web resources, are the basis for preparing rich contextual descriptions for Europeana objects. These rich contextual descriptions enable the creation of thematic portals, applications and services (for example, Europeana channels, Music retrieval pilots, extended search API calls, etc.), and greatly improved discoverability.

The list of controlled vocabularies selected to enrich the sound content contributed by the project is defined in *D1.3 Ontologies for Sound* (Ref 9) and *D1.4 EDM profile for sound* (Ref 10). These will be primarily used for metadata enrichment during the ingestion process, or in assisted processes. Apart from these vocabularies, there are web resources which are important in particular application contexts, either for culture vultures or culture snackers (for example, IMSLP[52], TheSession.org, Freebase, etc.) and which might be of interest for the crowdsourcing scenarios.

A list of web repositories that are relevant for project related activities is presented in Table 5. Linking heterogeneous resources (for example, music scores manuscripts, digital music scores and audio files) is particularly relevant for T2.4.1 Linking Music to Scores and T2.4.2 Music Information Retrieval, as well as in WP4 Channels Development.

Table 5: Relevant LOD Sources for improvement of existing enrichments

Name	Description
Europeana Open Skos Repository http://skos.europeana.eu/	This repository stores the controlled vocabularies defined by Europeana Aggregators, which do not have an authority responsible for publishing and maintaining them, e.g. music genres vocabulary aggregated by the Europeana Sounds project (Ref 9, Ref 10).
Freebase http://www.freebase.com/music	Currently, Freebase contains 29M topics related to the music domain. The Freebase music commons contains recording artists, albums, and songs. Data found here is a combination of information sourced from MusicBrainz and Wikipedia - further information sources will be integrated in the future.

[52] http://imslp.org/

TunePal http://tunepal.org/tunepal/zeitgeist.php	A software system and mobile application that performs automated transcription and identification of traditional music audio recordings (a Query By Playing interface), based on a corpus of 25,000 items supplied by various archives and musical scholars.
The Session http://thesession.org/	An online community which collects tunes, recordings, *sessions* and events related to the interpretation of Irish traditional music.
Petrucci Music Library http://imslp.org/	The International Music Score Library Project (IMSLP), also known as the Petrucci Music Library after publisher Ottaviano Petrucci, is a project for the creation of a virtual library of public domain music scores.
YouTube https://www.youtube.com/	YouTube is a video-sharing website which allows users to upload, view, and share videos, both user-generated and corporate media video. Available content includes video clips, TV clips, music videos, and other content such as video blogging, short original videos, and educational videos. A large share of YouTube content is music related.

5 ENRICHMENT EVALUATION

> *The outcomes of semantic enrichment must be validated and so must the outputs from crowdsourcing before crowd and semantically-enriched data can be used to improve discovery of content via Europeana, on search engines and within providers own sites.*
>
> (Ref 1 Part B, page 13)

For the evaluation phase of the crowdsourcing infrastructure, performance monitoring and reporting on progress, a solid qualitative measurement method is required. It is important to distinguish between the different types of annotations as they might require different methodologies. This section first pays attention to previous and existing work related to evaluation enrichment in the context of Europeana and the different platforms of the technical partners. The section ends with a strategy to develop a solid qualitative measurement method for practices that still requires a solution.

Europeana performs semantic and automatic enrichment on top of the data delivered by its data providers. This process implies the addition of information to the data about certain resources such as agents, places, concepts and timespans; and the creation of new links between these enriched resources and other reference datasets. Europeana currently performs enrichment using open and multilingual vocabularies such as Geonames[53], Dbpedia[54] and Gemet.[55] For more information on the current semantic and automatic enrichment processes performed by Europeana, please refer to *Appendix 2: Semantic Enrichment Framework*.

The semantic and automatic enrichment on top of the data that is ingested has previously been analyzed. The EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy set out to analyze datasets in Europeana and to evaluate them with regard to their enrichment potential and the enrichments that were executed. The goal was to drive a strategy for enriching metadata fields adding value for users. To achieve this, the members of the task force held a one-day workshop in Berlin where they analyzed randomly selected datasets from Europeana, their metadata fields and their enrichment potential. The Task Force on Multilingual and Semantic Enrichment Strategy has shown in its report *EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: final report* (Ref 11) that good quality data is crucial for preventing enrichments errors and flaws in the various stages the metadata undergoes in Europeana. In the context of this deliverable, it is also interesting to highlight the fact that this report also concludes in one of its recommendations that crowdsourcing could help further improve the enrichments:

[53] http://www.geonames.org/
[54] http://dbpedia.org/
[55] http://www.eionet.europa.eu/gemet/en/themes/

> *Another point of discussion was to leverage the users' input to crowd-source and validate links, filter ambiguous meaning and relations. Overall, users could be more involved in improving metadata quality and enrichment quality.*
>
> (Ref 11, page 28)

A workshop similar to the one organized by the Taskforce mentioned above for the Data Providers involved in Europeana Sounds could shed some light on the current quality of the semantic and automatic enrichment of the data in this domain. It will be included as part of the agenda for the WP1 Third Metadata Training Session, scheduled for July 2016.

Other relevant previous work related to the evaluation of the automatic (semantic) enrichment in the context of Europeana, is described in the paper *A Framework for the Evaluation of Automatic Metadata Enrichments* by Juliane Stiller, Marlies Olensky, and Vivien Petras from Humboldt University (Ref 11, 2014). The authors propose an evaluation based on four dimensions (frequency, coverage, relevance and error rate):

Table 6: Framework for evaluating enrichments

Dimension	**Object-dependent**	**Query-dependent**
Frequency	Percentage of enriched object; Enrichment per objects	Percentage of queries that retrieve enriched objects
Coverage	Distribution of enrichments across facets	Proportion of enriched objects per query; Percentage of queries retrieving enrichment facets
Relevance	Relevance of enrichment to the object	Relevance of enrichments to queries
Error rate	Percentage of incorrect enrichments; Percentage of retrieved objects with incorrect enrichments	Percentage of queries retrieving incorrect enrichments

In order to apply the evaluation framework to Europeana data, a sample corpus of queries and enriched objects for evaluation, the 1,000 most frequent queries in Europeana were extracted from Google Analytics. Based on a random selection of these queries searches in Europeana were performed and their first result pages were assessed.

Since automatic and manual enrichment are getting more attention in Europeana projects, it is now pertinent to explore the (abovementioned) recommendations from previous research. Previous

Taskforces in the Network and EuropeanaTech[56] context have already concluded that a holistic approach to quality improvement is crucially needed. This in turn calls for a holistic approach to quality assessment of enrichments. This is why in the first two quarters of 2015 a Taskforce on Enrichments and Evaluation will focus on establishing such a holistic approach, in the context of EuropeanTech (Europeana V3). Europeana Sounds will be represented in this Taskforce (by WP2 lead Maarten Brinkerink), in order to include the evaluation of crowdsourced improvements to enrichments to this holistic evaluation approach.

Evaluation of the improvement of automatic (semantic) enrichments through crowdsourcing specifically has less precedent, and has not been in scope of the previous research mentioned above. However, it makes sense to apply an A/B testing based method on a sample dataset, before and after improvement. Another possibility would be to perform evaluations, based on the Like/dislike and Approve/disapprove types of annotations. These methods - and other options and best-practices - will be explored further in the abovementioned Enrichments and Evaluation Taskforce.

[56] http://pro.europeana.eu/web/network/europeana-tech

6 SUMMARY

This deliverable contributes to the specification of functional requirements and the technical design for the crowdsourcing infrastructure that will be developed within WP2. The purpose of this deliverable is to describe the relevant scenarios for the Crowdsourcing Infrastructure and to provide a preliminary functional design for the functionality that is required to support these scenarios. This functional design will form the basis for the implementation of the Crowdsourcing Infrastructure within T5.5. This deliverable describes these functions from the point of view of the software that will be developed in different project scenarios within the Europeana Sounds project.

In order to write this document all technical partners have been involved in a process that, starting with the creation of a shared vision of the overall architecture of the project, allowed to classify the User Stories identified in D2.1 and "convert" them into functions. The classification of User Stories led to the identification of three different scenarios that take into account the environment in which the end users interact with the resources of the object: namely creating annotations. These three scenarios are:

- Manual creation of annotations (based on Annotorious, integrated in Europeana
- Manual correction of semantic annotations (based on Pundit, embedded on websites from Data Providers)
- Crowdsourcing through specialised platforms (supported by Historypin, including other specialized platforms)

This deliverable contains a system architecture overview, which presents the components used to build the workflows for the creation and modification of various annotation types. The system architecture is divided between "Client/Render" (what will be implemented as a user front-end, to perform crowdsourcing actions), "Services" (the back-end services for processing the results of crowdsourcing) and "Storage & Index" (where the resulting data will be stored and accessed from). All components in the system architecture overview, and their relation to the three above-mentioned scenarios, are described in this deliverable.

The document describes the User Stories and Epics related to the three scenarios. These User Stories and Epics are based on the previous work documented in D2.1, followed by a technical analysis. The User Stories were analysed by identifying the subjects involved in it, and mapped them to the Open Annotation terminology. Then they were linked to the crowdsourcing scenarios. The next step was the definition of the corresponding functions. And the last step was the assignment of a priority. By applying a general Gap Analysis, this deliverable also indicates what basic functionality is lacking, in order to enable the User Stories and Epics required for the Crowdsourcing Infrastructure.

The deliverable provides an inventory of main LOD sources that can be used for semantic enrichment. It is based on the previous work that was documented in D1.3 and adds web

resources which are important in particular application contexts, either for culture vultures or culture snackers and which might be of interest for the crowdsourcing scenarios.

This document concludes with possible strategies for assessing the quality of the user's contribution to the enrichment of the objects. This will be based on previous research in this area and will be part of a Europeana-wide holistic approach to quality assessment of enrichments. Such a holistic approach will be established within a Taskforce on Evaluation Methods for Automatic Enrichments, organized in the context of EuropeanTech. Europeana Sounds will be represented in this Taskforce, in order to include the evaluation of crowdsourced improvements to enrichments to this holistic evaluation approach.

APPENDIX 1

See separate document: 'Appendix 1 D2.2 Functional design of semantic enrichment - Working sheet - from User Stories to functions.xls'

APPENDIX 2: EUROPEANA SEMANTIC ENRICHMENT FRAMEWORK

Editor: Hugo Manguinhas
Contributors: Antoine Isaac, Valentine Charles, Yorgos Mamakis, Juliane Stiller
Version: 5 November 2014.
Based on: Semantic enrichment at Europeana – memo, Europeana Office, March 2, 2014
Attached ticket: 1219

Abstract

The semantic and multilingual enrichment of metadata in Europeana is a core concern as it improves access to the material, defines relations among objects and enables crosslingual retrieval of documents. The quality of these enrichments is crucial to ensure that highly curated content from providers gets represented correctly across different languages. To ensure that those enrichments unfold their full potential and act as facilitators of access, the semantic enrichment of the metadata is needed. This document describes the semantic enrichment process currently in place at Europeana, explaining the methods used to enrich the objects and the vocabularies that were selected for their enrichment, as well as, an assessment of the current results of enrichment and evaluation of their quality.

1. Introduction

In general, the process of semantic enrichment aims at adding new information at the semantic level to the data about certain resources. This is a rather vague notion, which has different interpretations depending on the disciplinary context. For example, in the Linked Data context, it chiefly refers to the creation of new links between the enriched resources and others, preferably coming from an existing, reference dataset. In Information Retrieval, it means adding new terms to a query or document and therefore reaching a higher visibility of documents within the document space.

A better way to understand the semantic enrichment process is to conceptually distinguish between three main stages that it embeds:

- **Analysis**: the pre-enrichment phase focuses on the analysis of the metadata fields in the original resource descriptions, the selection of potential resources to be linked to and derives rules to match and link the original fields to the contextual resource.
- **Linking**: the process of automatically matching the values of the metadata fields to values of the contextual resources and adding contextual links (whose values are most often based on equivalent relationships) to the dataset.
- **Augmentation**: the process of selecting the values from the contextual resource to be added to the original object description. This might not only include (multilingual) synonyms of terms to be enriched but also further information, for example broader or narrower concepts.

Even though these stages are present in Europeana, the name "*semantic enrichment*" is mainly used to designate only the second stage of this process as it refers to "*the creation of links to controlled vocabularies*" representing contextual resources such as places, concepts, agents and time periods.

For Europeana, the semantic enrichment of metadata is a core concern as it facilitates access to the material and contributes to the contextualization and knowledge generation beyond the original object description. Its goal is to make documents easier to be retrieved by alleviating some of the vocabulary problems in metadata and to add semantically rich resources that contextualize the document within its interpretational scope.

This document describes the semantic enrichment process and its current implementation called the "Semantic Enrichment Framework". It continues as follows: Section 2 describes the enrichment process in Europeana, in particular the contextual classes and sources, the vocabularies used, and the rules applied; Section 3 describes the current implementation; Section 4, its results; Section 5 describes the past efforts for the evaluation of enrichments; and finally, while Section 6, describes how the enrichments are displayed in the Europeana portal.

2. Semantic enrichments at Europeana

This Section describes in detail the semantic enrichment process in place at Europeana through each of its three main components:

- the **source** of enrichment i.e. the fields in EDM data from which the links are derived mostly by matching the string value of these fields to the labels of the contextual resources;
- the **target** of enrichments i.e. the set of resources to which objects in Europeana are linked;
- a **rule** that specifies how a match is obtained between the source metadata field(s) and the (labels of the) target contextual resources.

2.1 Contextual Classes in EDM

In EDM, four contextual classes are modelled as separate entities from the CHO with their own properties. The inclusion of these classes allows the exploitation of this rich data and allows data about the contextual resource to be kept separate from the data about the object of the description. The contextual classes provided by EDM are the following:

- A edm:Place to refer to a place as spatial locations identified by the provider and named according to some vocabulary or local convention;
- A skos:Concept to refer to a concept defined as a unit of thought or meaning that comes from an organised knowledge base (such as subject terms from a thesaurus or controlled vocabulary) where URIs or local identifiers have been created to represent each concept;
- A edm:Agent to refer to an agent comprising people, either individually or in groups, who have the potential to perform intentional actions for which they can be held responsible;
- A edm:TimeSpan for time periods defined as a period of time having a beginning, an end and a duration.

Presently, the semantic enrichment at Europeana is mainly focused on the properties that related a ProvidedCHO to one of the contextual resources described above. In particular, for places it focuses on dcterms:spatial and dc:coverage; for concepts, on dc:subject and dc:type; for agents on dc:creator and dc:contributor; and for time periods on dc:date, dcterms:temporal and edm:year.

2.2 Target vocabularies selected for enrichment

A target vocabulary for enrichment, is recognised as a set of resources to which objects in Europeana are linked. In order to be better exploited by Europeana services, a target vocabulary must meet the following requirements:

- technically available (through Linked Data or in dedicated repositories), properly documented, and in open access;
- well-connected together, e.g. equivalent elements in other vocabularies are indicated; and,
- multilingual

Looking at these requirements a set of target vocabularies was chosen for each of the conceptual entities being enrichment:

- For places, the Geonames57 geographical dataset was chosen since besides being publicly available, it offers a wide coverage of places with very rich content. From Geonames, only a subset58 of places from European countries that correspond to a specific Geonames feature class59 is being used for enrichment (i.e. "A", "P.PPL", "S.CSTL", "S.ANS", "S.MNMT", "S.LIBR", "S.HSTS", "S.OPRA", "S.AMTH", "S.TMPL", "T.ISL").
- For concepts, both GEMET60 and DBPedia were chosen. From GEMET, almost all concepts are used with some exceptions that were causing false matches (e.g. the case of "Drawing" which is a medical concept that was almost always used to enriched objects that have little to do with medicine). Besides this, some labels have been removed to prevent numerous harmful matches, such as linking any print to the physical “pressure” concept because of its German “Druck” alternative label (Ref 12). From DBpedia61, only a subset is used for enrichment which comprises a handful of WWI battles, the “World War I” category and the following categories:

[57] http://www.geonames.org/
[58] The vocabularies are accessible under the “places/countries” (for all European countries) and “places/EU” (for places within EU countries) subfolders of the vocabulary folder on GitHub.
[59] http://www.geonames.org/statistics/total.html
[60] GEneral Multilingual Environmental Thesaurus, http://www.eionet.europa.eu/gemet/
[61] http://dbpedia.org

"Art",	"Pre-Raphaelite_Brotherhood",	"Symbolism",
"Architecture",	"Kitsch",	"Music",
"Art_Deco"	"Still_life",	"Theatre",
"Art_Nouveau"	"Landscape",	"Painting",
"Baroque"	"Minimalism",	"Sculpture",
"Cubism",	"Modernism",	"Drawing",
"Contemporary_art	"Renaissance",	"Poster",
"Dada",	"Realism_(arts)",	"Photograph",
"Digital_art",	"Romanesque_art",	"Furniture",
"Expressionism",	"Romanticism",	"Costume",
"Fine-art_photography",	"Rococo",	"Fashion",
"Folk_art",	"Pastoral",	"Jewellery",
"Futurism",	"Portrait",	"Porcelain",
"Impressionism",	"Street_art",	"Tapestry",
"Neoclassicism",	"Surrealism",	"Woodcut"

- For agents, as for concepts, a subset from DBpedia corresponding mostly to painters is used. For this a set of dbpedia categories62 were selected and is used to obtain the agents that fit those categories.
- For time periods, Semium Time63 was chosen. It is a vocabulary of time periods which is partially automatic generated (for "objective" time divisions like the 3rd quarter of 15th century) and partly manual (for historical period like "Roman empire"). At the time of writing, Semium Time ceased to be available and for this reason an alternative needs to be identified for the enrichment of time periods - or Europeana should start hosting it.

At the time of writing, Europeana is looking at other vocabularies that can be used as targets for enrichments. Among these is Freebase[64] which is currently being implemented as an alternative to DBpedia for the enrichment of both agents and concepts.

In order to be efficiently used within the semantic enrichment framework, these vocabularies are: retrieved from the providing web sources (through a specific protocol / API or by downloading the dumps); converted into the corresponding contextual entity in EDM; stored as separate RDF files; and, finally bundled as part of the tool (see the vocabularies folder[65] for the full set of resources currently bundled with the tool). This means that when target vocabularies undergo significant change, the RDF files should be updated and a new version must be assembled and deployed.

[62] For the full set of categories see the FetcherOfPeopleFromDbpediaSparqlEndpoint.java file under the enrichment project on GitHub.

[63] Previously accessible at: http://semium.org/time.html. A snapshot is available under the vocabulary folder on GitHub.

[64] https://www.freebase.com/

[65] https://github.com/europeana/tools/tree/master/europeana-enrichment-framework/enrichment/enrichment-framework-service/converters/vocabularies/

Table 7: Summary of the target vocabularies selected for enrichment

Dimension	Target Vocabulary	Size	Details
Places	Geonames	140,097	European countries and the most relevant places within them.
Concepts	GEMET concepts	5,284	Close to all concepts.
	DBPedia	163	Subset containing a handful of WWI battles, the "World War I" and other categories.
Agents	DBPedia	14,222	Subset with mostly painters.

2.3 Matching rules

As the target vocabularies, the rules applied within the semantic enrichment framework will be different depending on the contextual entity to which the object will be enriched. Each of these rules will only be applied for a selection of properties of the object (i.e. "source fields"). For each of these source fields, a case insensitive match is performed against all of the preferred and alternative labels of the target vocabulary. For the particular case of agents, the source fields are processed prior to matching to remove birth/death dates in parentheses or roles in brackets, while for the others, no string processing of the matching fields (source and target) is done (see Table 8 below).

Table 8: Summary of the enrichment rules for each contextual entity

Contextual Entity	Source fields	Matching Rule	Target fields
Places	dcterms:spatial, dc:coverage	Case insensitive	skos:prefLabel, skos:altLabel
Concepts	dc:subject, dc:type	Case insensitive	skos:prefLabel, skos:altLabel
Agents	dc:creator, dc:contributor	Case insensitive with: • **source field**: Removal of birth/death dates in parentheses or roles in brackets	skos:prefLabel, skos:altLabel
Time Periods	dc:date, dcterms:temporal, edm:year	Case insensitive	skos:prefLabel, skos:altLabel

3. Current implementation

The current and previous[66] implementations of the Semantic Enrichment Framework are historically based on the Annocultor[67] software. However, the new version of the framework was subject to a significant refactoring of the code and architecture, as well as, the software libraries being used, but maintaining most of the functionality supported by the previous version. The code is available at Europeana's GitHub account under the tools folder[68]. As part of the new version, a REST API[69] was designed to allow for machine access to the semantic enrichment framework.

The full overview of the semantic enrichments in Europeana is not completed without understanding its use within the core infrastructure for ingestion currently in place at Europeana. This infrastructure is based on a workflow system called UIM[70] which runs activities implemented as pluggable software components and which the Enrichment Plugin[71] is one. This plugin besides being responsible for interacting with the enrichment framework it is also responsible for the pre-processing of the source fields (as explained in Section 2.1) and also generating the corresponding enriched fields. In the future, this functionality will go inside the enrichment framework itself.

An installation is available at this address[72] which can be used for testing the current implementation of the Semantic Enrichment framework. It is a deployment of the *enrichment-framework-gui*[73] maven module and uses the latest version of the target vocabularies.

4. Current results of enrichments

The process of enrichment is applied to every dataset being ingested in Europeana. The number of enriched objects depends both on how the target vocabularies are able to cover the terms referred by the objects and the ability of the rules applied in the enrichment to recognize these terms within the target vocabularies. A quantitative analysis of the number of objects being enriched through the Semantic Enrichment Framework can be depicted in Figure 6. It is important to note that these measurements were obtained by querying the portal (more precisely, by querying for the generated property of each dimension and filtering by a keyword present in the URI of the target vocabulary,

[66] https://github.com/europeana/tools/tree/master/annocultor_solr4
[67] http://sourceforge.net/projects/annocultor/
[68] https://github.com/europeana/tools/tree/master/europeana-enrichment-framework/enrichment
[69] For further documentation on the API and instructions for its installation see: https://docs.google.com/document/d/1fjVG-F4K7xl4oHuv1COGOVP1H6PhM2Fd6hMSwQdvDYY
[70] https://github.com/europeana/uim-europeana
[71] https://github.com/europeana/uim-europeana/tree/master/workflow_plugins/europeana-ui

[72] http://testenv-solr.eanadev.org:9191/enrichment-framework-gui-0.1-SNAPSHOT/

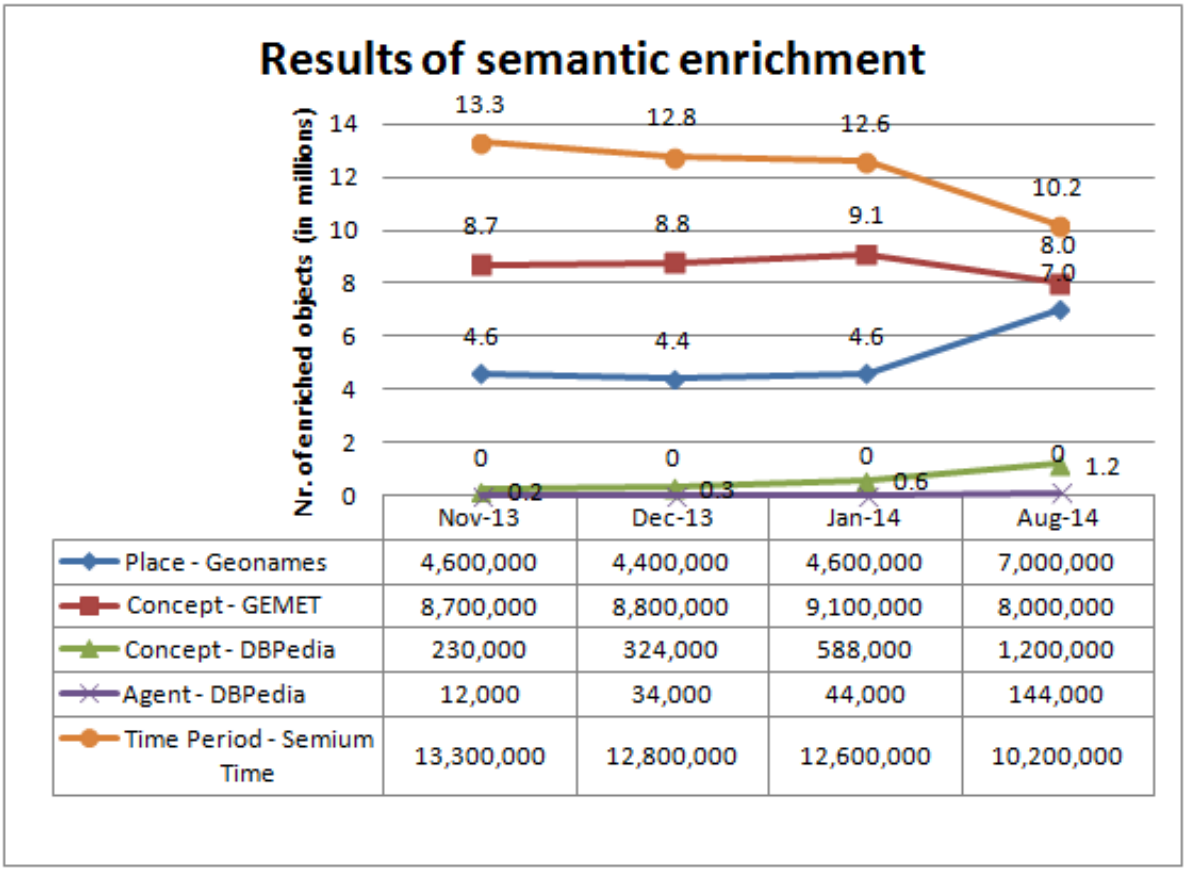
[73] https://github.com/europeana/tools/tree/master/europeana-enrichment-framework/enrichment/enrichment-framework-gui

see Figure 5) and not by actually measuring the enriched objects since there is currently no way to distinguish an enrichment coming from the data provider from the Europeana enrichment. This is especially the case for Geonames and Dbpedia since there are many data providers that already link their sources to these vocabularies. In order to improve the precision of these measurements, a way must be found to distinguish enrichments from other data.

URI syntax:
http://www.europeana.eu/portal/search.html?query=<property>:*<keyword>*

Examples:
http://www.europeana.eu/portal/search.html?query=edm_place:*geonames*
http://www.europeana.eu/portal/search.html?query=skos_concept%3A*dbpedia*

Figure 5: Query syntax used for obtaining the measurements



	Nov-13	Dec-13	Jan-14	Aug-14
Place - Geonames	4,600,000	4,400,000	4,600,000	7,000,000
Concept - GEMET	8,700,000	8,800,000	9,100,000	8,000,000
Concept - DBPedia	230,000	324,000	588,000	1,200,000
Agent - DBPedia	12,000	34,000	44,000	144,000
Time Period - Semium Time	13,300,000	12,800,000	12,600,000	10,200,000

Figure 6: Quantitative overview of the results obtained with the semantic enrichment

5. Evaluation of the enrichments

The most significant effort towards the evaluation of enrichments was performed by the EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy[74]. It set out to analyse datasets in Europeana and to evaluate them with regard to their enrichment potential and the enrichments that were executed. The goal was to drive a strategy for enriching metadata fields adding value for users. To achieve this, the members of the task force held a one-day workshop in Berlin where they analysed randomly selected datasets from Europeana, their metadata fields and their enrichment potential. The Task Force on Multilingual and Semantic Enrichment Strategy has shown in its report (Ref 13) that good quality data is crucial for preventing enrichments errors and flaws in the various stages the metadata undergoes in Europeana. In particular, it was recognised that Europeana should encourage the delivery of specialized vocabularies with resolvable URIs which would also lead to less need for enrichments by Europeana itself. With regard to the enrichment process itself, it states that clear rules for each field need to be established.

Another relevant work (Ref 14) related to evaluation was done under the Europeana v2.0 project which aimed at evaluating the impact of semantic enrichment from the perspective of its contribution to search. For the purpose of this study, a framework (see Table 9) was designed based on four dimensions (frequency, coverage, relevance and error rate) and used to assess a sample corpus of queries (i.e. a random selection of the 1,000 most frequent queries extracted from Google Analytics) and their respective results containing enrichments (only the first result page). This study showed that enrichments might not necessarily increase an object's findability or likelihood to be retrieved, but they can help in the contextualization of objects, particularly in a multilingual environment. The results from this study have contributed to identifying focus areas for future developments of enrichments.

[74] http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Task+force+multilingual+semantic+enrichment

Table 9: Framework for evaluating enrichments

Dimension	Object-dependent	Query-dependent
Frequency	Percentage of enriched object; Enrichment per objects	Percentage of queries that retrieve enriched objects
Coverage	Distribution of enrichments across facets	Proportion of enriched objects per query; Percentage of queries retrieving enrichment facets
Relevance	Relevance of enrichment to the object	Relevance of enrichments to queries
Error rate	Percentage of incorrect enrichments; Percentage of retrieved objects with incorrect enrichments	Percentage of queries retrieving incorrect enrichments

6. Displaying enrichments

The enrichments can be seen at the portal (www.europeana.eu) and are shown on the object page under the “auto-generated tags” section (see Figure 7). An example of an object that is currently enriched by Europeana’s semantic enrichment framework can be accessed at this location[75]. At the time of writing, the enrichments are displayed in a confusing way since URIs present in the original metadata and URIs resulting from enrichment are both shown in the fields representing the original values (e.g. "Geographic coverage") and the "Auto-generated tags" section.

[75] http://europeana.eu/portal/record/09003/818CA77941A28830F41B3DCD8FA25EB951E50971.html

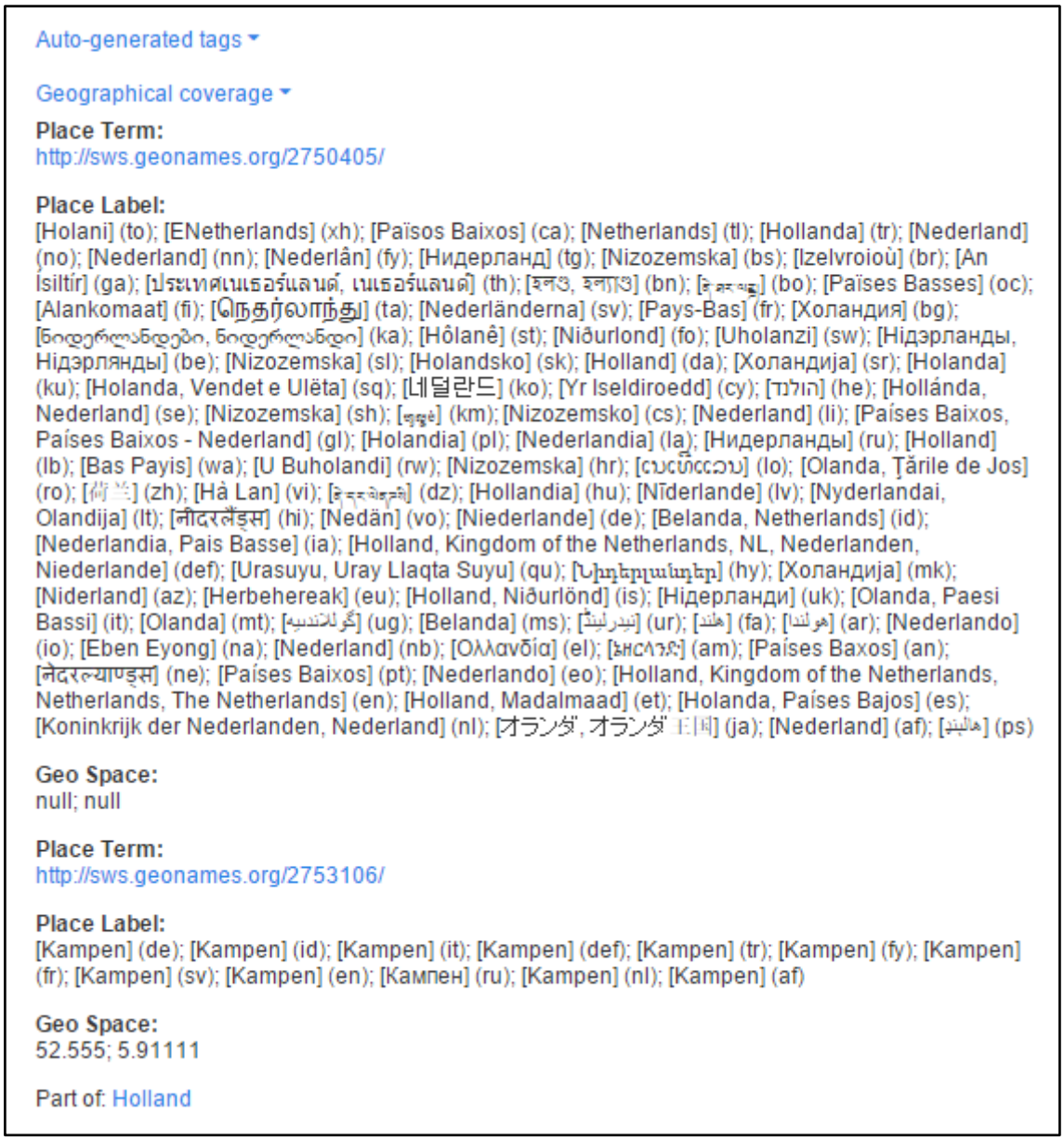


Figure 7: Display of automatic enrichments in the Europeana portal

REFERENCES

REF 1	Description of Work Link to Document on Europeana Professional: http://pro.europeana.eu/documents/2011409/2034825/Description+of+Work
REF 2	D2.1 *Crowdsourcing infrastructure and exchange policy* Link to Document on Europeana Professional: http://pro.europeana.eu/documents/2011409/2034829/EuropeanaSounds-D2.1-Crowdsourcing-infrastructure-and-exchange-policy
REF 3	MS7 *End-user contributions defined* Link to Document on Europeana Professional: http://pro.europeana.eu/documents/2011409/2089168/EuropeanaSounds-MS7-End-user-contributions-defined
REF 4	Design for User Engagement on Europeana Channels (C. Shen, 2014) Link to Document on WP4 Basecamp: https://basecamp.com/1936492/projects/4984678/attachments/100527076/download
REF 5	Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. 2011. Crowdsourcing systems on the World-Wide Web. Commun. ACM 54, 4 (April 2011), 86-96. DOI=10.1145/1924421.1924442 http://doi.acm.org/10.1145/1924421.1924442
REF 6	Sui, Daniel Z., Sarah Elwood, and Michael F. Goodchild. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Dordrecht: Springer, 2013. Print.
REF 7	MS19 *Audio Channels First Prototype* Link to Document on Europeana Professional: http://www.pro.europeana.eu/documents/2011409/2089168/EuropeanaSounds-MS19-Audio-channels-first-prototype
REF 8	Europeana Creative D2.2 *Services and Messaging API* Link to Document on Europeana Professional: http://pro.europeana.eu/web/europeana-creative/project-documents/-/document_library_display/yP06/view/1601973

REF 9	D1.3 *Ontologies for Sound* Link to Document on Europeana Professional: http://pro.europeana.eu/documents/2011409/ed7f27af-65ff-4ea3-9d00-8345c31489cf
REF 10	D1.4 *EDM profile for sound* Link to Document on Europeana Professional: http://www.pro.europeana.eu/documents/2011409/5c845fc2-dcf7-46c2-a66d-6aea9dac1c0c
REF 11	EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: final report (Juliane Stiller, Antoine Isaac & Vivien Petras, 2014) Link to Document on Europeana Professional: http://pro.europeana.eu/documents/468623/8b75b054-712e-432b-a0f7-761898e6f60e
REF 12	Olensky, M., Stiller, J., Dröge, E.: Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy. In: Dodero, J.M., Palomo-Duarte, M., Karampiperis, P. (eds.) MTSR 2012. CCIS, vol. 343, pp. 252–263. Springer, Heidelberg (2012)
REF 13	Stiller, J., Isaac, A., Petras, V. (eds.): EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: Final Report (2014) Link to Document on Europeana Professional: http://pro.europeana.eu/documents/468623/8b75b054-712e-432b-a0f7-761898e6f60e
REF 14	Stiller, J., Olensky, M., Petras, V.: A Framework for the Evaluation of Automatic Metadata Enrichments. In: Closs, S., Studer, R., Garoufallou, E., Sicilia, M.-A. (eds.) MTSR 2014 – 8th Metadata and Semantics Research Conference.