EUROPEANA SOUNDS

Project Number: 620591

MS8 Functional design of semantic enrichment

Document Identifier: EuropeanaSounds-MS8-Functional-design-of-semantic-enrichment-v1.0.docx

Document link: http://pro.europeana.eu/web/europeana-sounds/documents

Date: 03/11/2014

Abstract

This Milestone contributes to the specification of functional requirements for the crowdsourcing infrastructure, by analysing the following major concerns: the target audience for semantic enrichments; the acquisition and improvement of semantic enrichments; and the system architecture and data flows exchanged between software components. Improvements of existing enrichments include both correction and refinement and can be applied to four types of existing enrichments: (1) automatic enrichments that are applied during the ingestion of metadata; (2) semi-automatic enrichments, resulting from the application of the ontologies by Data Providers; (3) assisted structured enrichments by users, as a result of crowdsourcing micro-tasks; (4) manual enrichments by users, as a result of crowdsourcing micro-tasks .This enables a constant process of improvement of the metadata quality, through crowdsourcing. The crowdsourcing micro-tasks that will be designed for the improvements of existing enrichments are aimed at enabling the crowd to improve the metadata quality in a structured way, leveraging their domain knowledge. This Milestone also provides a graphical representation of the current state of the system architecture and data flows.

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	Maarten Brinkerink	NISV	16/06/2014	
0.2	Draft	Maarten Brinkerink	NISV	23/09/2014	First draft
0.3	Draft	Maarten Brinkerink	NISV	27/10/2014	Draft for review
0.4	Draft	Sergiu Gordea	AIT	29/10/2014	Additions
1.0	Final	Maarten Brinkerink	NISV	31/10/2014	Final

III. DELIVERY SLIP

	Name	Partner/WP	Date
Document Author	Maarten Brinkerink mbrinkerink@beeldengeluid.nl	NISV / WP2	31/10/2014
Reviewed by	**Reviewers:** Pavel Kats Remy Gardien	 EF EF	27/10/2014
Approved by	**Coordinator & PMB**		31/10/2014

IV. DISTRIBUTION

No.	Date	Comment	Partner / WP
1	03/11/2014	Submitted to the Europeana Commission	BL/WP7
2	03/11/2014	Posted on Europeana Pro	BL/WP7
3	03/11/2014	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: END-USER CONTRIBUTIONS DEFINED

This Milestone contributes to the specification of functional requirements for the crowdsourcing infrastructure, by analysing the following major concerns; the target audience for semantic enrichments (1), the acquisition and improvement of semantic enrichments (2) and the system architecture and data flows exchanged between software components (3).

As described in MS7 *End-user contributions defined*, this distinction has been further developed, into “culture snackers” and “culture vultures”.

Improvements of existing enrichments include both correction and refinement and can be applied to four types of existing enrichments; automatic enrichments that are applied during the ingestion of metadata (1) semi-automatic enrichments, resulting from the application of the ontologies by Data Providers (2) assisted structured enrichments by users, as a result of crowdsourcing micro-tasks (3) manual enrichments by users, as a result of crowdsourcing micro-tasks (4).This enables a constant process of improvement of the metadata quality, through crowdsourcing.

The crowdsourcing micro-tasks that will be designed for the improvements of existing enrichments are aimed at enabling the crowd to improve the metadata quality in a structured way, leveraging their domain knowledge.

The list of controlled vocabularies selected to enrich the sound content contributed by the project is defined in D1.3 *Ontologies for Sound*. These will be primarily used for metadata enrichment during the ingestion process or in assisted processes. Apart from these, there are web resources which are important in particular application contexts.

This Milestone also provides a graphical representation of the current state of the system architecture and data flows.

TABLE OF CONTENTS

1 INTRODUCTION

As stated in the Description of Work (DoW, Part B, page 9) [REF 1] end-user engagement lies at the very heart of Europeana Sounds. This Milestone contributes to the specification of functional requirements for the crowdsourcing infrastructure, by analysing the following major concerns:

- The targeted audience for semantic enrichments (see Section 2)
- The acquisition and improvement of semantic enrichments (see Section 3)
- The system architecture and data flows exchanged between software components (see Section 4)

As such this document focuses on the 'Task 2.2: Semantic enrichment'. This Milestone provides the foundation for D2.2 *Functional design of semantic enrichment*, a more technical report, enhancing the content of this Milestone with a technical design, including an assessment of the suggested LOD sources and guidance for measuring the quality of enrichments.

2 USERS

Enrichment through crowdsourcing, as supported by WP2, will be designed in the form of "micro-tasks" (DoW, page 10) for two types of audiences; the general public (1) and experts (2). As described in MS7 *End-user contributions defined* [REF 2], this distinction has been further developed, along the line of Chenchen Shen's paper "Design for User Engagement on Europeana Channels" into "culture snackers" and "culture vultures". [REF 3]

To summarise, this means that the 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 professionally with the Europeana portal, or 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 following:

> *"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."*
>
> (Final report, page 20)

3 IMPROVING EXISTING ENRICHMENTS THROUGH CROWDSOURCING

3.1 Types of existing enrichment

Improvements of existing enrichments include both correction and refinement. These improvements can be applied to four types of existing enrichments, in the context of Europeana Sounds:

1. Automatic enrichments that are applied during the ingestion of metadata [REF 4]
2. Semi-automatic enrichments, resulting from the application of the ontologies by Data Providers, as specified in D1.3 *Ontologies for Sound* [REF 5]
3. Assisted structured enrichments by users, as a result of crowdsourcing micro-tasks
4. Manual enrichments by users, as a result of crowdsourcing micro-tasks

The latter two types of existing enrichments show that the improvement of existing enrichments not only leverages a crowd of culture vultures in order to correct and/or refine the result of (semi-) automatic enrichments, but also to correct and/or improve the results of its own collective contributions to the enrichments (community sourcing). This enables a constant process of improvement of the metadata quality, through crowdsourcing.

The automatic and assisted enrichments are certified by the domain experts and represented directly into the EDM metadata, while the user contributed enrichments are acquired from the larger public in a less controlled scenario (i.e. annotations are not certified as being correct, the representation is not fixed in a rigid format and no restrictions are imposed on the type of resources referenced in the annotations). Therefore, the general Open Annotation Model[1] was chosen to represent these enrichments.

3.2 Relevant types of annotations for improvement of existing enrichments

In MS7, the following figure depicting different types of annotations relevant for the technical infrastructure of Europeana Sounds was introduced. This figure is the result of discussions carried out within bi-weekly teleconferences involving the technical partners in Europeana Sounds (the Technical Coordination Group), as organised by the Technical Coordinator of the project Johan Oomen:

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

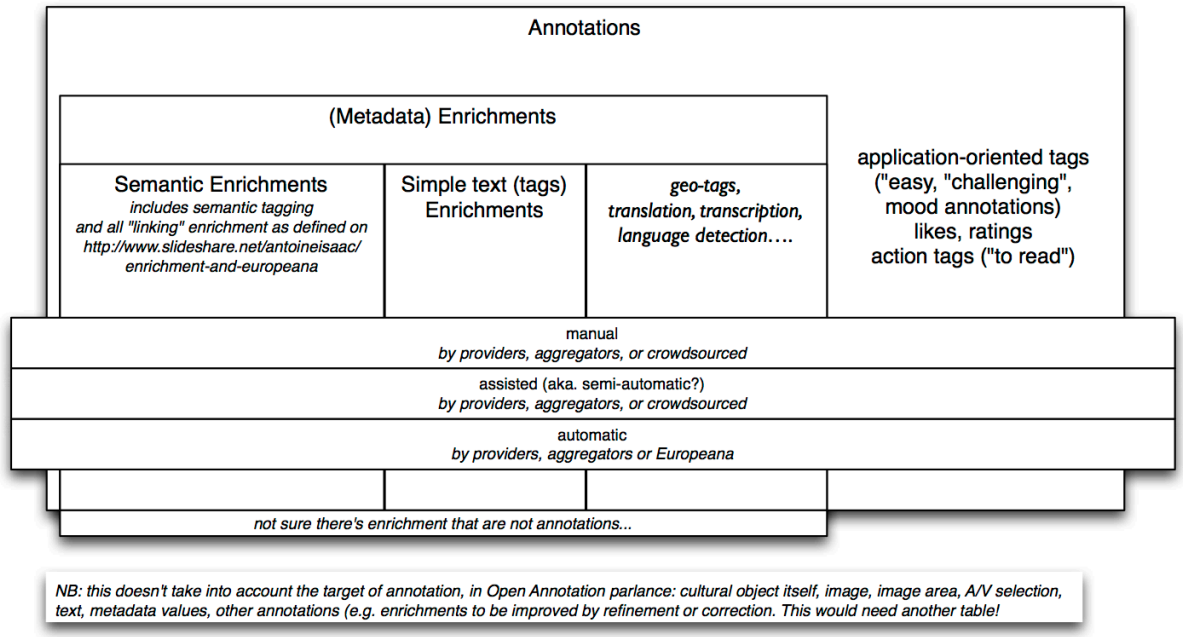

Figure 1: Different types of annotations within the technical infrastructure

The crowdsourcing micro-tasks that will be designed for the improvement of existing enrichments fall within the realm of annotations that can be considered (metadata) enrichments of an assisted nature. These are aimed at enabling the crowd to improve the metadata quality in a structured way. They allow users to contribute their domain specific knowledge, by leveraging structured information represented in Linked Open Data sources, and existing methods and tools from the domains of statistical data analysis, graph search and named entity extraction.

3.3 Relevant LOD sources for improvement of existing enrichments

Semantic enrichments that aim to inter-link Europeana objects or link Europeana objects with semantic web resources are a particular type of annotation. They form the basis for preparing rich contextual descriptions for Europeana objects, which are required for building thematic portals, applications and services (e.g. Europeana channels, Music retrieval pilots, extended search API, etc.).

The list of controlled vocabularies selected to enrich the sound content contributed by the project is defined in D1.3 *Ontologies for Sound* [REF 5]. These will be primarily used for metadata enrichment during the ingestion process or in assisted processes. Apart from these, there are web resources which are important in particular application contexts, either for culture vultures or culture snackers (e.g. IMSLP, TheSession.org, Freebase, etc.) and which might be of interest.

A list of web repositories that are relevant for the project related activities is presented in Table 1 below. Linking heterogeneous resources (e.g. 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 1: 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).
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.
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, based on the wiki principle.
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.

4 SYSTEM ARCHITECTURE DRAFT

The figure below depicts the current status of an ongoing discussion about the system architecture within the Technical Coordination Group. This draft will be finalised with the delivery of D2.2 *Functional design of semantic enrichment*:

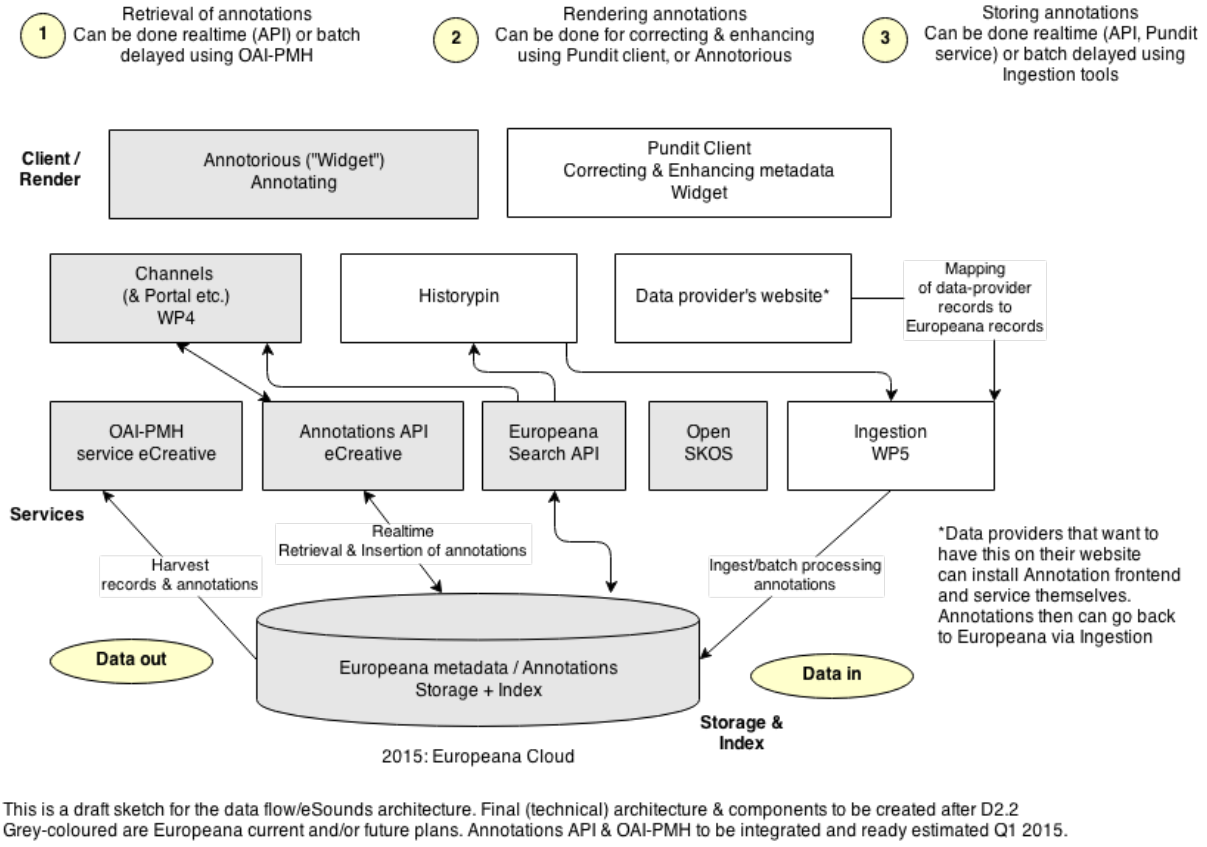

Figure 2: Draft system architecture

The figure shows how the enrichments flow between the different components within the system architecture.

In the table below a short description of the main components is given:

Table 2: Description of main components within the system architecture

Name	Description
Pundit client	Dedicated end-user application for crowdsourcing improvements of existing enrichments, by using the Pundit[2] frontend and the Annotation API
Annotorious widget	Embeddable solution that supports acquisition of end-user annotations integrated into existing web portals (e.g. Europeana Portal, Europeana Channels, Content Provider websites). The embeddable widget is based on Annotorious[3]
Channels	End-user website to explore the material aggregated within the context of Europeana Sounds - developed in the context of WP4 - that includes crowdsourcing functionality
Historypin	End-user website that includes crowdsourcing functionality for collection thematic annotations
Data Provider's website	Websites of the Data Providers that are part of the Europeana Sounds consortium, that can include a 'widget' to include crowdsourcing functionality
OAI-PMH service	Webservice for harvesting metadata and annotations from the main Europeana data aggregation
Annotations API	Webservice for (realtime) retrieval and insertion of annotations to the main Europeana data aggregation
Europeana Search API	Webservice for (realtime) retrieval of metadata records and annotations from the main Europeana data aggregation
OpenSKOS	Repository for controlled vocabularies that can be utilised for assisted types of (metadata) enrichments, within the crowdsourcing micro-tasks
Ingestion	The process used for ingesting semantic enrichments from third party repositories into Europeana.
Europeana metadata/annotations storage + index	Main Europeana repository used for storage and retrieval of europeana metadata and annotations

[2] https://github.com/net7/pundit
[3] https://annotorious.github.io/

5 SUMMARY

This Milestone contributes to the specification of functional requirements for the crowdsourcing infrastructure, by analysing the following major concerns: the target audience for semantic enrichments; the acquisition and improvement of semantic enrichments; and the system architecture and data flows exchanged between software components.

As described in MS7 *End-user contributions defined*, this distinction has been further developed, into "culture snackers" and "culture vultures".

Improvements of existing enrichments include both correction and refinement and can be applied to four types of existing enrichments: (1) automatic enrichments that are applied during the ingestion of metadata; (2) semi-automatic enrichments, resulting from the application of the ontologies by Data Providers; (3) assisted structured enrichments by users, as a result of crowdsourcing micro-tasks; (4) manual enrichments by users, as a result of crowdsourcing micro-tasks.

The latter two types of enrichments show improving existing enrichments not only leverages a crowd of culture vultures to correct and/or refine the result of (semi-) automatic enrichments, but also uses them to correct and/or improve the results of their own collective contributions to the enrichments (community sourcing). This enables a constant process of improvement of the metadata quality, through crowdsourcing.

The crowdsourcing micro-tasks will be designed for the improvement of existing enrichments and are aimed at enabling the crowd to improve the metadata quality in a structured way, leveraging their domain knowledge.

Semantic enrichments that aim to inter-link Europeana objects or link Europeana objects with semantic web resources are a particular type of annotation. They form the basis for preparing rich contextual descriptions for Europeana objects, which are required for building thematic portals, applications and services (e.g. Europeana channels, Music retrieval pilots, extended search API, etc.).

The list of controlled vocabularies selected to enrich the sound content contributed by the project is defined in D1.3 *Ontologies for Sound*. These will be primarily used for metadata enrichment during the ingestion process or in assisted processes. Apart from these, there are web resources which are important in particular application contexts.

6 REFERENCES

Ref 1	EC-GA including Annexe I ("Description of Work") http://pro.europeana.eu/documents/2011409/8d0e9833-4608-494e-af77-681e68f8a8c8
Ref 2	MS7 End-user contributions defined http://pro.europeana.eu/documents/2011409/ed3b1dbb-ca33-4754-8d6f-abb921e38b46
Ref 3	Shen, C. "Design for User Engagement on Europeana Channels". Delft University and Europeana, Master of Science Graduation Project-Design for Interaction. 2014. https://basecamp.com/1936492/projects/4984678/attachments/100527076/download
Ref 4	Frequently Asked Questions: 'How does Europeana enrich metadata?' See information underneath header 'Data Enrichment' on Europeana pro: http://pro.europeana.eu/providers-faq
Ref 5	D1.3 Ontologies for sound http://pro.europeana.eu/documents/2011409/ed7f27af-65ff-4ea3-9d00-8345c31489cf