Project Acronym: **Europeana Sounds**
Grant Agreement no: **620591**
Project Title: **Europeana Sounds**

MS28: Sounds thesaurus and metadata cleaning and normalization module complete

Revision: Final

Date: 03/08/2015

Authors: Natasa Sofou, NTUA

Vassilis Tzouvaras, NTUA

Abstract: This milestone reports on the Europeana Sounds thesaurus and normalization and cleaning functionalities delivered and integrated into the MINT platform.

Dissemination level	
Public	X
Confidential, only for the members of the Consortium and Commission Services	

Coordinated by the British Library, the Europeana Sounds 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/

Revision history

Version	Status	Name, organisation	Date	Changes
0.1	ToC	Natasa Sofou, NTUA	15/07/2015	
0.2	1st draft	Natasa Sofou, NTUA Vassilis Tzouvaras, NTUA	20/07/2015	
0.3	2nd draft	Natasa Sofou, NTUA	29/07/2015	Input from reviewers
0.4	Final draft	Eva Hayles-Gledhill, BL	31/07/2015	Minor changes
1.0	Final	Laura Miles, BL Richard Ranft, BL	31/07/2015	Layout, minor changes

Review and approval

Action	Name, organisation	Date
Reviewed by	Nikos Simou, NTUA / WP5 Dimitra Atsidis, NISV/ WP2	29/07/2015 29/07/2015
Approved by	Coordinator and PMB	31/07/15

Distribution

No.	Date	Comment	Partner / WP
1	31/07/2015	Submitted to the European Commission	BL/WP7
2	31/07/2015	Posted on Europeana Pro website	BL/WP7
3	31/07/2015	Distributed to project consortium	BL/WP7

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.

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This document contains original unpublished work except where clearly indicated otherwise.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

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

Copyright notice

Contents

Executive summary: MS28 Sounds thesaurus and metadata cleaning and normalization module complete

The aim of WP5 is to enable metadata aggregation by extending and enhancing the existing Europeana aggregation infrastructure. This milestone reports on the Europeana Sounds vocabulary and normalization and cleaning module as delivered by NTUA.

1 Introduction

The aim of WP5 is to enable metadata aggregation by extending and enhancing the existing Europeana aggregation infrastructure. This milestone reports on the normalization and cleaning module and the integration of the sounds thesaurus in the MINT platform.

Europeana Sounds focuses specifically on audio and audio-related content, primarily music and speech audio, including out of commerce recordings and a large number of unpublished works from Europe's major sound archives that are not widely available. In WP1, the current EDM profile has been expanded to facilitate richer mappings and thus richer metadata for Europeana. Additionally, controlled vocabularies have been defined to be used in the enrichment step of the aggregation processes. The vocabularies designed and delivered within WP1 (D1.3 *Ontologies for sound*) are converted to SKOS (Simple Knowledge Organization System), a W3C recommendation designed for representation of structured controlled vocabulary on the semantic web that is already in use within the Europeana Sounds data provider community (for example, by the French and German national libraries). The conversion is performed under the umbrella of WP5 and the outcome is deployed by the MINT platform to facilitate richer metadata.

The Europeana Sounds data repository is populated by metadata records following the Europeana Data Model sounds profile defined and delivered within WP1 Task 1.3 as reported in D1.4 *EDM profile for Sound*. As foreseen by T5.4 *Metadata cleaning and normalization*, a web service for group handling and editing of records is provided to facilitate enrichment and reconciliation. This service enables providers to clean their metadata and normalize it using selected SKOS thesauri. Providers also have the option to browse their contributed items in the EDM profile, filter them based on specific elements, values and search criteria and group edit the resulted item set in order to perform a set of predefined actions such as; 1) data cleaning, correcting typographic errors and 2) data reconciliation; align metadata elements with the sounds thesaurus.

2 SKOS vocabularies

The vocabularies designed and delivered within WP1 (see D1.3 *Ontologies for sound*) are converted to SKOS (Simple Knowledge Organization System). The SKOS representation was derived from the tabular data created as described in details in D1.3 using the RDF extension of Open Refine (formerly Google

refine), one of the most used open source data curation tools. For visualization purposes, SKOSPlay[1] was chosen, an open source online tool that can be used to generate interactive hierarchical views and to create PDF documentation for each concept included in a vocabulary. OpenSKOS[2] was chosen as a SKOS vocabularies repository, which was already used by the Europeana Foundation. It is an open source solution that supports storage and consumption of SKOS vocabularies via REST-API and deploys a web interface to allow domain experts to edit the vocabularies. The REST-API facilitates integration with other applications, for example in the end-user facing portal channels, which will be developed in WP4. Vocabularies produced by T1.4 Aggregation management are hosted at the OpenSKOS instance provided and maintained by Europeana.

The resulting sounds thesaurus (which can be thought to be composed of all sounds vocabularies) is integrated into the MINT platform in order to enable data providers to align their metadata with selected thesaurus entries. The sounds thesaurus accessible through MINT platform is illustrated in Figure 1.

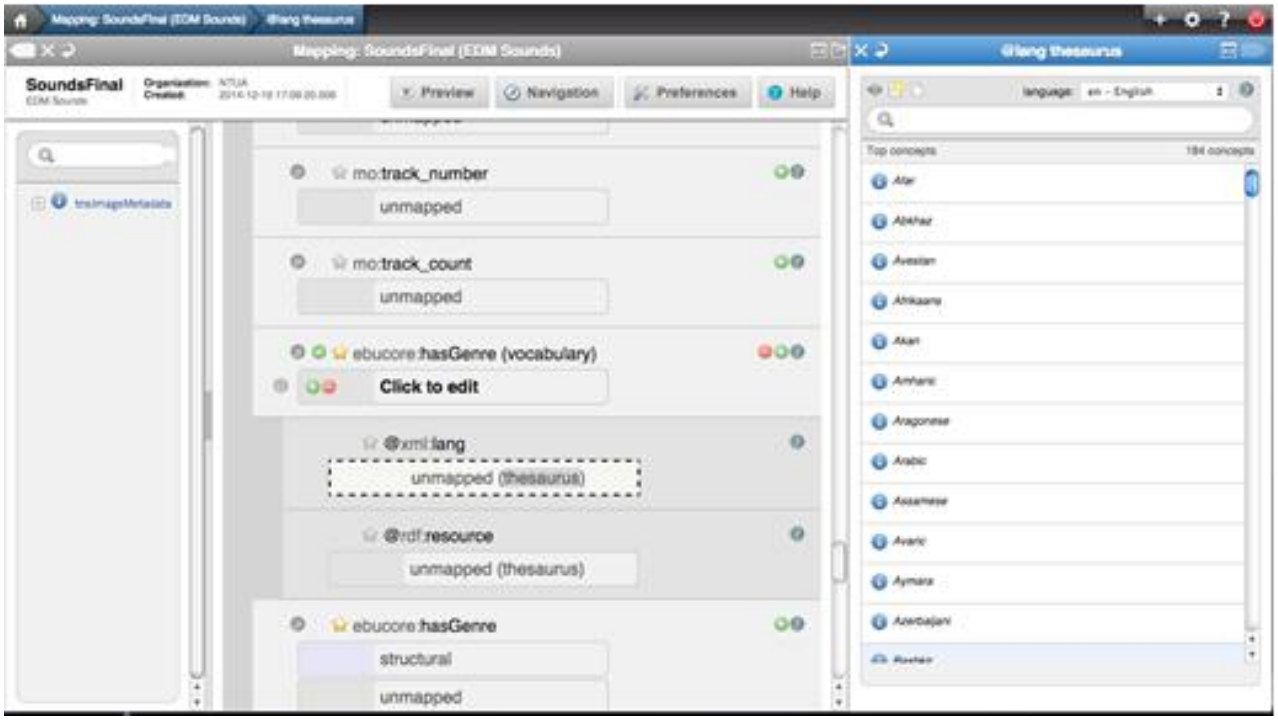

Figure 1: the Sounds thesaurus is accessible through the MINT platform.

3 Metadata cleaning and normalisation module

The enrichment of aggregated metadata records is approached through three main processes. The first one is that of data cleaning, in this case the user will be able to cleanse the data. Typographical errors will be corrected and it will be possible to conform to specific conventions easily. The second methodology is data reconciliation, based on this users will be able to align their metadata to

[1] http://www.w3.org/2001/sw/wiki/SKOS_Play!
[2] http://openskos.org/

vocabularies provided by WP1. The last methodology which is complementary to data reconciliation is that of linking the metadata records to external resources which is automatic and is available after the metadata is ingested to Europeana servers.

Metadata enrichment can be applied to different steps of the aggregation/ingestion process, through a set of different functionalities that are either integrated into the ingestion process via the MINT platform or provided as complementary tools.

Within MINT, enrichment is approached through the cleaning and reconciliation of metadata records. In the case of cleaning and reconciliation, the use of MINT's XSLT mapping editor for correcting values using simple value mappings and mappings to the SKOS vocabularies (Figure 1 shows the SKOS thesaurus navigation panel in MINT). More complex cases can be handled using conditional mappings, string manipulation functions and constant value mappings. A more detailed description of the above functionalities and the way they can be applied to perform metadata enrichment can be found in MS25 and will also be reported in the upcoming report D5.3 *Sounds thesaurus and the metadata cleaning and normalisation module*.

As an additional and more advanced set of enrichment tools in terms of metadata cleaning and normalization, a **single item annotator** for EDM Sounds profile and a **group item annotator** that works based on dataset filtering and application of CRUD[3] operations on schema elements is available and provides advanced enrichment functionality. At item level, annotations can be applied on record level, and the selected dataset is both browsable and searchable.

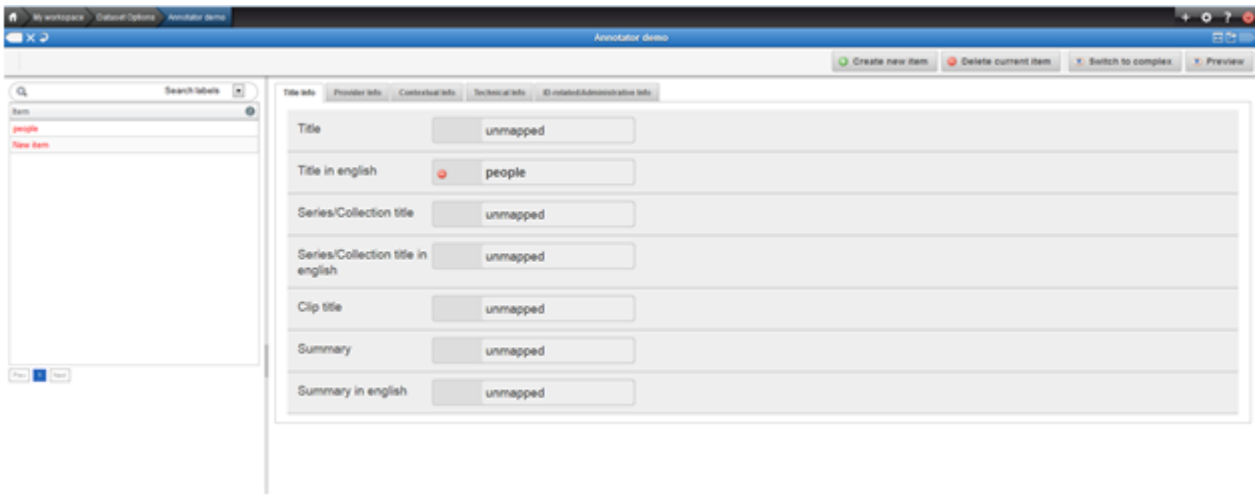

Figure 2 - the item annotator

At group level, a number of operations can be defined under the group edit module:

- The creation of an element on each selected record, which can be filled with a user defined value.
- The deletion of an element based on conditional criteria from all the selected records.
- Updating of a value from existing elements.

[3] http://en.wikipedia.org/wiki/Create,_read,_update_and_delete

A screenshot of the group annotator and the corresponding actions log is illustrated in **Error! Reference source not found.** 3. A detailed list of all the functionalities and actions that can be applied with group edit module can be found in MS25 *Sounds SKOS ontology and normalization and cleaning module beta*, and they will also be reported in the upcoming report for D5.3 *Sounds thesaurus and the metadata cleaning and normalisation module*.

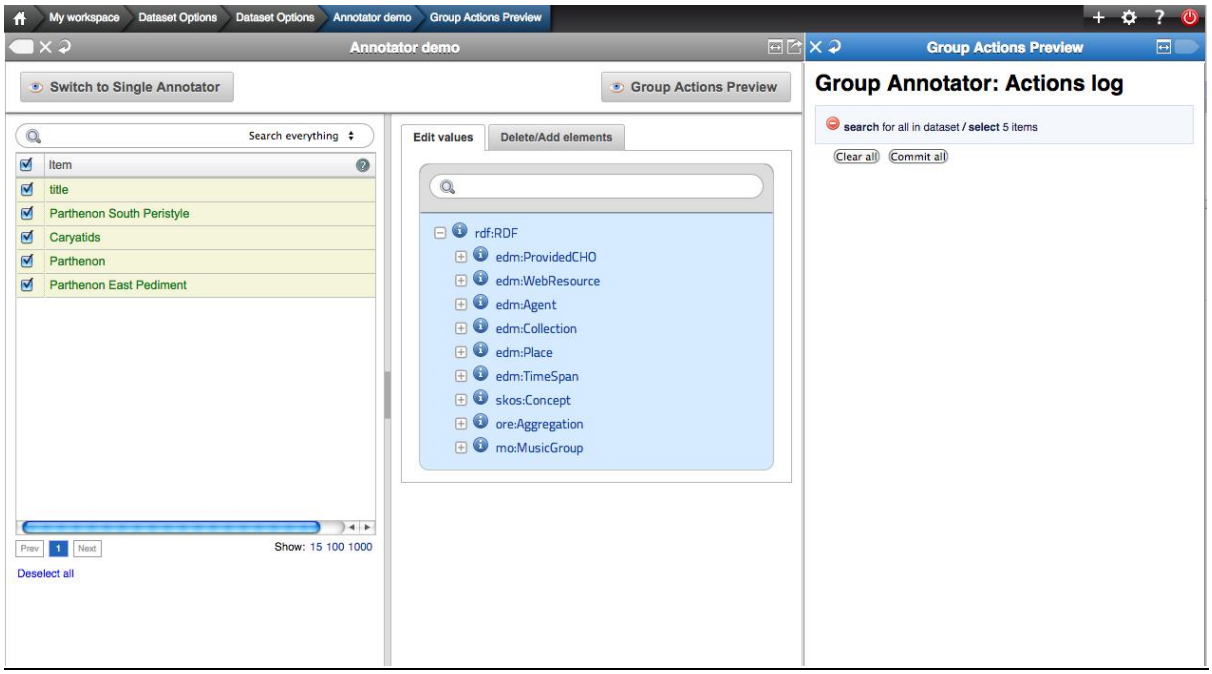

Figure 3 - group annotator view

4 Technical information

4.1 Implementation details

Cleaning and normalization modules (single item and group item annotator) are parts of the MINT platform and can be accessed through the MINT interface.

As reported previously, all parts of the backend software are written in Java and are executed in a 1.6+ JVM. The platform is developed using JAVA, JSP, HTML and JavaScript. PostgreSQL is used as an object-relational database with Hibernate as the data persistence framework, and mongoDB as a document-oriented database. MINT uses other open source development frameworks and libraries according to specific deployments and customizations. MINT source code versions are released under a free software license (GNU Affero GPL).

4.2 Installation and usage

In order to install MINT the following software is needed:

- JAVA v1.6+ and Tomcat 6

- Postgres 8.4
- Maven, an Apache build manager for java projects.

Installation instructions can be found at:

http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/Mint_Installation_instructions

The new web service of Item and group annotator can be reached through MINT at:

http://mint-projects.image.ntua.gr/sounds/

by selecting: **My workspace> Dataset Options>Annotator Demo** and then choosing between single item and group item annotator, as illustrated in Figure 4 below.

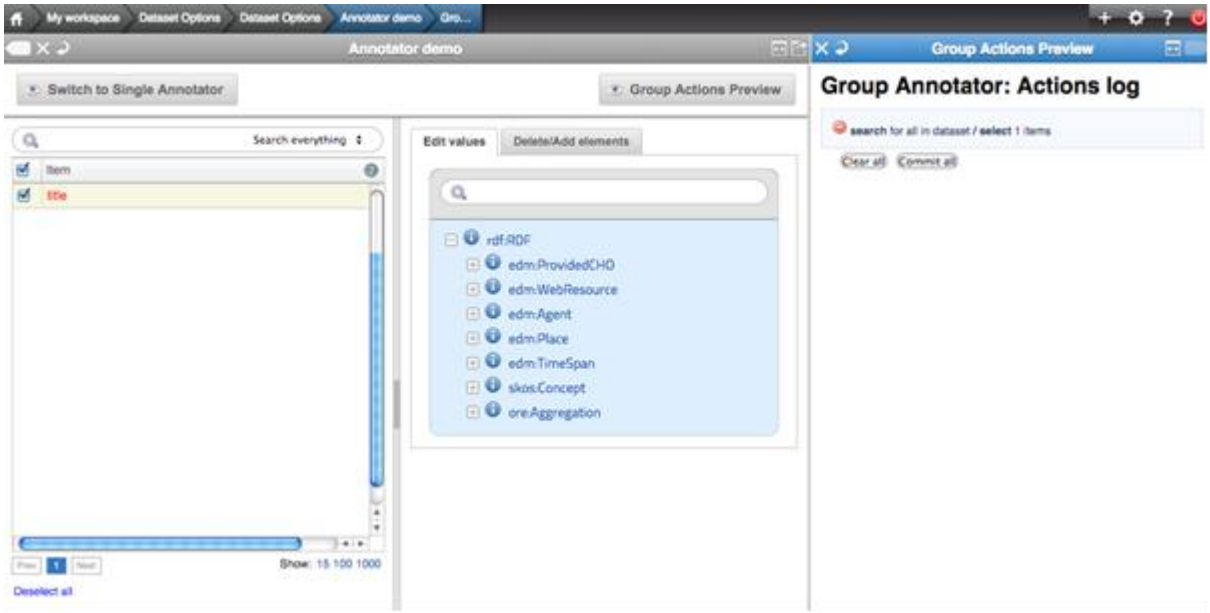

Figure 4 - **single/group item annotator access through MINT**

5 Conclusion

Metadata is enriched in many ways and in various phases of the aggregation workflow. There are enrichment phases after the metadata is ingested into the Europeana servers. WP5 is responsible for the enrichment that happens during the aggregation workflow. A module, part of MINT, has been developed for manually enriching metadata. This new web service enables providers to clean their metadata and normalize it using selected SKOS thesauri. Providers also have the option to browse their contributed items in the EDM profile, filter them based on specific elements, values and search criteria and group edit the resulted item set in order to perform a set of predefined actions such as; 1) data cleaning, correcting typographic errors 2) data reconciliation; aligning metadata elements with the sounds thesaurus.

This milestone, a sequel to MS25 *Sounds SKOS ontology and normalisation and cleaning module beta*, reports briefly on the delivered module, providing some general functionality information along with implementation, access and usage details. A more complete report will be delivered in Month 19 as D5.3 *Sounds thesaurus and the metadata cleaning and normalisation module*.

Appendix A: Terminology

A project glossary is provided at: http://pro.europeana.eu/web/guest/glossary.

Additional terms are defined below:

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
PMB	Project Management Board
TEL	The European Library
UAP	User Advisory Panel
WP	Work Package