



Project Acronym: Europeana v2
Grant Agreement number: 270902
Project Title: Europeana Version 2

MS40: Report on innovative applications, second iteration

Revision	Second draft
Date of submission	15-12-2013
Author(s)	Maarten Brinkerink, Netherlands Institute for Sound and Vision
Dissemination Level	Public

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
1	29-10-2013	Maarten Brinkerink	Netherlands Institute for Sound and Vision	First draft, based on previous plans, e-mail exchanges and and teleconferences.
2	26-11-2013	Maarten Brinkerink	Netherlands Institute for Sound and Vision	Second draft, based on input from WP7.
3	15-12-2013	Maarten Brinkerink	Netherlands Institute for Sound and Vision	Third version, based on input from the Europeana Foundation and WP7.
4	16-12-2013	Antoine Isaac	Europeana	Final read-through and edit.

Statement of originality:

This milestone contains original unpublished work except where clearly indicated otherwise.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

Introduction

This milestone reports on the plans for further development of the innovative applications, as part of Task 7.2. This is a plan that builds upon the work reported on in MS39¹, as part of Subtask 7.1.1., to select a number of key applications that were beneficial to Europeana and the wider European cultural heritage domain. A first cycle of further development of these innovative applications has already happened, following the Functional specifications for social semantic functions that were part of D7.5². This Milestone presents the plans for further development, based on the lessons learned during the previous development cycle and feedback from the EuropeanaTech network.

¹ <http://pro.europeana.eu/documents/866067/48d5f16d-9383-4b13-afa0-8235c463f03c>

² http://pro.europeana.eu/group/europeana-v2.0/documents?p_p_auth=3YLSHY3i&p_p_id=20&p_p_lifecycle=1&p_p_state=exclusive&p_p_mode=view&p_p_col_id=column-2&p_p_col_count=1&_20_struts_action=%2Fdocument_library%2Fget_file&_20_groupId=866067&_20_folderId=1660973&_20_name=50007

Waisda?

Waisda?³ allows anyone who is interested to use the white label open source version of the software to easily create their own video tagging game for crowdsourcing time-related video-annotations. The platform provides all the standard functionality of the game and allows for customization of the interface and game mechanisms.

In their article ‘Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges’⁴ Johan Oomen and Lora Aroyo describe how “GLAMs and their users are now beginning to inhabit the same, shared information space”. As a result they identify potential for user involvement in a range of tasks that were previously only executed within institutions. They would classify the task that Waisda? enables to crowdsource as “Classification”, or “Gathering descriptive metadata related to objects in a collection”. Waisda? can function as a tool to involve internet users in the gathering of descriptive metadata for video content.⁵ This metadata has a level of granularity – because of the exact time-relation - that most institutions could never support with their own resources for documentation. As a result Waisda? supports the creation of metadata that would otherwise most probably be non-existent.

This chapter describes the further development of the Waisda? prototype⁶ that was set up as part of Task 7.2, following the Functional specifications for social semantic functions, which were part of D7.5. In the previous round of development Waisda? has been enhanced with easy integration of Europeana records. This allows anyone (including Europeana Data Providers) to easily create their own instance of an online tagging game for video content with the open source video annotation software provided by Waisda?, based on a selection of Europeana records (using the Europeana API).

The further development of Waisda? within the context of Task 7.2 consists of four different types of further development:

1. Connecting Waisda? to OpenSKOS instances, allowing for a generic method to utilize controlled-vocabularies that are available in a SKOS format for the matching mechanisms of the game.
2. Extending Waisda? with a harvestable export mechanism, allowing Europeana – and potentially also other third parties – to reuse the user generated metadata that results from Waisda?
3. Creating an updated Waisda? prototype for a multitude of video collections that are available through Europeana, to allow the Europeana Network to investigate the possibilities that Waisda? offers them.
4. Updating the open source code repository of Waisda?, allowing third parties to also benefit from the abovementioned further development.

³ <https://github.com/beeldengeluid/waisda>

⁴ <http://www.cs.vu.nl/~marieke/OomenAroyoCT2011.pdf>

⁵ Sound and Vision used the Waisda? platform for three production versions, in collaboration with public broadcasters in the Netherlands, and was able to gather over a million ‘social tags’.

⁶ <http://waisda.tuxic.nl:8080/>

OpenSKOS Connector

Tag entries gathered with Waisda? are already matched against several controlled vocabularies. The outcome of this matching is recorded in the field “dictionary” in the “TagEntry” table. Adding dictionaries with words is done by filling the table “DictionaryEntry” of the MySQL database with records. This can be done by importing an SQL or CSV dump of a vocabulary or using a script that generates insert statements. Each tag entry is tied to a dictionary, which allows the scoring engine to differentiate between different dictionaries. For example, if you have entries for dictionaries “people” and “placenames”, the scoring engine could choose to award more points to place names than to people's names. The current Waisda? prototype for Europeana utilizes vocabulary terms from the GTAA⁷ controlled vocabulary from The Netherlands Institute for Sound and Vision.

As previously mentioned in D7.5, the aim is to enhance the Waisda? prototype for Europeana with a generic method to utilize various controlled vocabularies that are available as SKOS data (Simple Knowledge Organization System)⁸, that are suited to the specific needs of a data provider, collection, target audience, etc. To enable such a generic method, the Waisda? prototype for Europeana will be extended with a OpenSKOS connector. OpenSKOS⁹ is a CATCHPlus¹⁰ project. It provides an API on top of SKOS data. It functions as a repository to host and organize SKOS vocabularies. The RESTful webservice¹¹ that OpenSKOS provides will be connected to the Waisda? game mechanisms. This will allow for a generic method to utilize a SKOS vocabulary within the game, once it is loaded into an OpenSKOS instance. In practice it will mean that Waisda? will be extended with functionality that will call the RESTful webservice of a specified OpenSKOS repository, to see whether it matches and relevant controlled-vocabulary terms. The result of this call will be stored in the database of the game, for future reuse purposes.

User Generated Metadata Exporter

As a video tagging game platform for crowdsourcing time-related video-annotations, Waisda? is aimed at gathering large amounts of user generated metadata. This data consists of descriptive tags about the visual and/or audio content of a video. These ‘social tags’ are time-related and enriched different types of information. For instance the database stores, who entered the tag, whether these tags were confirmed by other players, and whether these tags matched with one or more predefined controlled-vocabularies.

To enable reuse of the social tags by Europeana, the integration of the application with Europeana mentioned in MS39, the Waisda? prototype for Europeana will be extended with a harvestable export mechanism, allowing third parties to extract the user generated metadata from the database of the game. Records of the video items that run in the game will be made harvestable via the Open Archives Initiative Protocol for Metadata Harvesting¹², that already forms the basis of the Europeana aggregation workflow. User generated metadata that was gathered within the game for these video records will be modeled according to the Open Annotation data model¹³. A first version of this export mechanism for the user generated metadata will focus on exporting the terms that match controlled-vocabularies (including time stamps and URIs). Together with WP6 a workflow will be determined for the actual harvesting of these video records, enriched with user generated

⁷ <http://openskos.org/api/collections/beng:gtaa.html>

⁸ W3C maintains a list of SKOS vocabularies (<http://www.w3.org/2001/sw/wiki/SKOS/Datasets>). Increasingly GLAMs (Galleries, Libraries, Museums and Archives) are also making their controlled vocabularies available in a linked open data format. See for instance <http://www.heritagedata.org/blog/vocabularies-provided/> and <http://www.getty.edu/research/tools/vocabularies/lod/index.html>.

⁹ <http://openskos.org/>

¹⁰ <http://www.catchplus.nl/>

¹¹ <http://openskos.org/api>

¹² <http://www.openarchives.org/pmh/>

¹³ <http://www.openannotation.org/spec/core/>

metadata. This can either be through the usual harvesting workflow for Europeana, or based on the annotations extension of the Europeana API, mentioned in MS39.

Updated Prototype

To showcase the possibilities of the abovementioned enhancements of the Waisda? platform to the EuropeanaTech community, the Europeana Network and beyond, an updated Waisda? prototype will be set up for Europeana. This prototype – or demonstrator – will incorporate a multitude of video collections from different Europeana Data providers. During the previous round of development Waisda? has been enhanced with easy integration of Europeana records, based on the Europeana API. Currently the Europeana dataset includes over 3,000 suitable video records, from data providers like the European Film Gateway¹⁴, Linked Heritage and the Digitale Collectie¹⁵. In a joined effort with Community Coordinator Joris Pekel from the Europeana Foundation, suitable data providers will be identified and approached, to see if they are interested to add their collections to the prototype and – if so – align further steps.

NB: this will also help determining in a more precise way, what is the level of applicability of Waisda? for the Europeana collections. Prior research efforts have shown that the number of collections with suitable audiovisual content is very limited—hence the efforts to develop new content re-use frameworks, such as the one worked on in the EuropeanaAwareness and EuropeanaCreative project.

Code Repository Update

Lastly, the open source code repository of the Waisda? video tagging platform will be updated with all the abovementioned enhancements of the code base. This will allow third parties, from the Europeana Network and beyond, to also benefit from the further development. The software will be released under an open source GPL license at <https://github.com/beeldengeluid/waisda> (including documentation).

¹⁴ In the context of the PrestoPRIME project, an experiment with utilizing Waisda? For content from the European Film Gateway was already deployed: <http://prestoprime.cs.vu.nl/efg>

¹⁵ Based on the following query: http://europeana.eu/portal/search.html?query=provider_aggregation_edm_isShownBy%3A*.mp4

PyBossa

PyBossa¹⁶ is an open source platform for creating and running crowdsourcing applications that utilize online assistance in performing tasks that require human cognition, knowledge or intelligence such as image classification, transcription, geocoding and more.

This chapter describes the further development of the PyBossa prototype¹⁷ that was set up as part of Task 7.2, following the Functional specifications for social semantic functions, which were part of D7.5. In the previous round of development PyBossa has been tailored into a generic platform for creating online tagging games for still images that lets users perform classification tasks, based on a selection of Europeana records (using the Europeana API). To illustrate the usage of PyBossa as an online tagging game for still images, based on Europeana records, two tasks were developed; one for correcting metadata and one for correcting enrichments.

The further development of PyBossa within the context of Task 7.2 consists of three different types of further development:

1. Extending PyBossa with a third crowdsourcing task for still images, which allows users to assess images similarity by linking documents and determining their relationship.
2. Extending PyBossa with an export mechanism, allowing Europeana – and potentially also other third parties – to reuse the user generated metadata that results from PyBossa
3. Creating an updated PyBossa prototype for a multitude of image collections that are available through Europeana, to allow the Europeana Network to investigate the possibilities that PyBossa offers them.
4. Updating the open source code repository of PyBossa?, allowing third parties to also benefit from the abovementioned further development.

NB: problems related to finding a suitable developer for PyBossa imply that some prioritization of the above items may happen, in order to focus on the most important ones. We will especially keep in mind the objective of proving the general concepts that the prototype is aimed to demonstrate (especially for the Europeana network), and meeting the functional requirements defined by D7.5.

¹⁶ <http://dev.pybossa.com>

¹⁷ <http://culttag8.herokuapp.com>

Linking Documents Task

Editors who set up an image tagging instance with PyBossa are able to compile tasks that consist of a collection in Europeana, then select appropriate vocabularies (that can be used to describe the objects) for it and select specific tasks. Currently the tasks “Correcting metadata” and “Correcting enrichments” have been implemented (as described in D7.5). This will be extended with the task “Linking documents”.

The task to link documents will be based on the standard IR-similarity that the Europeana API transmits to present two images that are similar to each other to the user. The user will then need to judge the two images and be presented with the question underneath the two images “Are these two images similar to each other?” and the possibility to answer with “yes/no” (via a radio button); and then underneath it, it should ask “Can you specify what type of relationship these two images have?” and have a field where the user then can type in the specification of the type of relationship. As suggested in D7.5 this specification will be based on the following typology:

- Same objects/duplicates
- Part of another Cultural Heritage Object (CHO)
- Views of the same CHO
- Derivative work
- Part of thematic cluster
- Collections

The first two of the abovementioned task correspond with what Oomen and Aroyo call “Correction and Transcription Tasks” or “Inviting users to correct and/or transcribe outputs of digitisation processes”, as they support the involvement of users in the quality assessment and correction of metadata and/or enrichments. The third task to link documents corresponds with what Oomen and Aroyo call “Contextualisation”, because the users are asked to add “contextual knowledge to object” by linking them to other object and specifying their relation. In all these cases the user contributions are leveraged to enhance the metadata quality, a strategic goal for the Europeana foundation¹⁸.

User Generated Metadata Exporter

Similar to what will be done for the updated Waisda? prototype, PyBossa will be extended with a function to export the user-generated metadata. We will study whether the Open Annotation data model can be also be used here, possibly after some extension. This will be carried out in line with the development on image annotation in the EuropeanaCreative project.

Updated Prototype

To showcase the possibilities of the abovementioned enhancements of the PyBossa platform to the EuropeanaTech community, the Europeana Network and beyond, an updated PyBossa prototype will be set up for Europeana. This prototype – or demonstrator – will incorporate a multitude of video collections from different Europeana Data providers. The prototype and the possibility to participate as a data provided will be promoted through EuropeanTech and a blogpost on Europeana Professional.

Code Repository Update

Lastly, the open source code repository of the PyBossa generic platform for online image tagging games will be updated with all the abovementioned enhancements of the code base.

¹⁸ for example, The Europeana Network has launched a task force on metadata quality in the weeks preceding the delivery of this report.

MS40: Report on innovative applications, second iteration

This will allow third parties, from the Europeana Network and beyond, to also benefit from the further development. The software will be released under an open source license on Github including documentation).