WITH: Human Computer Collaboration for Data Annotation and Enrichment

HumL@WWW2018

Alexandros Chortaras, Anna Christaki, Nasos Drosopoulos, Eirini Kaldeli, Maria Ralli, Anastasia Sofou, Arne Stabenau, Giorgos Stamou, Vassilis Tzouvaras

Intelligent Systems Laboratory, National Technical University of Athens
Digital Era of Cultural Heritage

- Vast amounts of content are available through cultural institutions.
- Content is aggregated through cross domain hubs, such as Europeana, DPLA.
- Poor data and metadata quality.
- Content has limited accessibility and discoverability.

The main motivation of WITH was to utilize CH repositories in unison and promote the digital cultural content by **enhancing** its **accessibility** and **discoverability** and achieving **user engagement**.
Introducing WITH

http://withculture.eu/

WITH is a cultural ecosystem that:

- Exploits cultural heritage content
- Promotes human-computer collaboration
- Provides enhanced services for data/metadata management and enrichment
- Facilitates accessibility and discoverability of available cultural content
WITH User Engagement

Federated Search and the Content Management processes enable users to collect and organise content.

Metadata Enrichment and the Crowdsourcing processes enable users to advance content descriptions, using AI content analysis tools or human annotations.
WITH is a CH aggregation platform with focus on human-computer collaboration through user engagement.

WITH services are:

- content aggregation and management
- metadata enrichment through automatic annotations and crowdsourcing campaigns
Aggregation and Federated Search

WITH aggregates metadata from multiple sources and through APIs mashups stores them in its database using WITH data model.

It enables search with multiple metadata criteria (e.g. sources/ rights/media type/date).
WITH Data Model

- Compatible with Europeana Data Model (EDM)
- Includes extensions to ensure interoperability with various data models
- Supports various serializations JSON, XML, RDF
Content Management

Users can create interesting content views and presentations

- **Collections** group user collected items together.
- **Exhibitions** provide enhanced and more playful visualization features.
- **Spaces** provide cultural content organization in different thematic categories and views. Spaces enable CH organisations to promote their content and engage with other users.
Featured Collections

Parisienne de Photographie Collection #2
4076 Items - By Parisienne de Photographie
2016/11/17

Lithuanian Art Museum Collection #1
206 Items - By Lithuanian Art Museum Museum

Zelius Madsen Gunnar
342 Items - By polfoto polfoto
2016/11/17

United Archives Collection #1
4648 Items - By United Archives Uniarc
2016/11/16

United Archives Collection #2
3345 Items - By United Archives Uniarc
2016/11/16
WITH Metadata Enrichment Process

Additional metadata in form of Linked Data Resources (or IRIs) can be associated with WITH items or parts of them.

Enrichment can be accomplished in two ways:

- **Automatic** enrichment of metadata via image and text analysis methodologies
- **Manual** annotation using controlled vocabularies and thesauri, and via **crowdsourcing** initiatives

WITH annotations (additional metadata) associate a WITH item, or a part of it, with a Linked Data resource or other IRI.
WITH includes a thesauri manager to facilitate the creation, retrieval, management and interoperability of annotations.

**Thesauri manager** converts the imported vocabularies from their source format (e.g. SKOS thesauri, OWL ontologies, N-triples datasets) to a common model, stores them in the **WITH thesauri database** and indexes them for fast research and retrieval.

**Supported Linked data resources**
- Getty Art and Architecture Thesaurus **AAT**
- GEMET thesaurus
- MIMO
- **WordNet**
- Europeana Fashion Thesaurus,
- Europeana photoVocabulary
- **DBpedia**
- Geonames
WITH Annotation Model

WITH annotation model is based on W3C’s Web Annotation Model.

It consists of:

- **id**
- **list of annotators** (info about origins of annotation),
- **body** (Linked Data resource of IRI),
- **target** (WITH item, metadata field value or part of item),
- **list of scores** (users that have upvoted or downvoted the annotation).
Manual Annotation

- Users choose a resource from the underlying thesauri database.

- Assign terms from the thesauri to the item.

- Geotagging tool is offered as a manual annotation service.
Manual Annotation Example

Annotate

Select vocabularies

Art & Architecture Thesaurus
DBPedia Ontology
DBPedia Resources
EuscreenXL Thesaurus

Start typing a term

wn30: painting
http://wordnet-rdf.princeton.edu/wn30/10391513-n

trade

wn30: painting
http://wordnet-rdf.princeton.edu/wn30/01792557-v

graphic art

wn30: painting
http://wordnet-rdf.princeton.edu/wn30/07494363-n

fine arts art

wn30: painting
http://wordnet-rdf.princeton.edu/wn30/14322699-n

application

aat: eludoric painting
Automatic Annotation

Textual analysis: automatic identification of name entities (persons, locations, organisations) in descriptive metadata

- named entity recognition and disambiguation NERD (using DBpedia spotlight).
- dictionary lookup

Visual analysis: automatic visual annotation of images

- computer vision algorithms
- feature extraction
- deep neural net methods for detection and localization of faces, diverse set of common objects, generic image classification (using ImageNet DB and WordNet concepts)
Automatic Annotation Example

**Annotate**

**MANUAL**  **AUTOMATIC**  **GEO**

**TERM DETECTION**

**NAMED ENTITY RECOGNITION**

**IMAGE ANALYSIS**

Select the named entity recognition engines that will be used

DBPedia Spotlight
Annotator
WITH Named Entity Recognizer

Select the image analysis services that will be used

CultIVML Annotator

[Annotate]  [Close]
Crowdsourcing Data Annotation

WITH offers a crowdsourcing infrastructure that essentially complements any automatic enrichment.

- annotate
- validate
- up/downvote

Initiating a crowdsourcing campaign

- import /select cultural content
- make a content-thematic Space
- organise data into collections
- enrich their data where possible with automatic annotation tools
- specify the desired crowdsourcing features such as duration, target annotation number, desired annotation type (semantic tagging, image tagging, geotagging, etc.), vocabularies and thesauri to be used.
Campaign: Semantic Tagging of Music Recordings

TAG OUR ARCHIVE: SPOT MUSICAL INSTRUMENTS

Have an ear for musical instruments? Use your talent to tag all the instruments you hear in the music archives of Europeana Sounds! There are so many historical music recordings in our archives, that it is impossible to describe all the instruments by ourselves. This is why we kindly ask your help to tag our beloved music collections with all the instruments that you can hear in a song. The music in Europeana Sounds ranges from beautiful folk songs from Greece to Scandinavian improvised space rock, all the way to sounds coming from Tibetan monasteries.

Featured Collections

Music from BNF
3731 Items - By annac
2016/04/26

Music from ITMA
1382 Items - By Sounds Project Sounds
2016/06/24

Music from Denmark
591 Items - By annac
2016/04/26

Music from FMS
1232 Items - By Sounds Project Sounds
2016/06/44

Music from the Internet Archive
13427 Items - By Sounds Project Sounds
2016/06/10

Music from CNRS-REM
7365 Items - By Sounds Project Sounds
2016/03/19
Defining the Campaign Features

- Creation of Dedicated Space
- Organisation of music recordings into collections (13 collections - 36,791 items)
- User engagement through social media and special events
- Organization of dedicated crowdsourcing sessions

Crowdsourcing features:
  - Duration: 1 month
  - Type: semantic tagging
  - Vocabulary: MIMO Vocabulary
  - Goal: 30,000 tags
User Identified MIMO Tags

<table>
<thead>
<tr>
<th>Clarinet Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>clarin</td>
</tr>
<tr>
<td>mimo: Triple clarinet</td>
</tr>
<tr>
<td>mimo: Idioglott clarinet</td>
</tr>
<tr>
<td>mimo: Double clarinet</td>
</tr>
<tr>
<td>mimo: Basset clarinet</td>
</tr>
<tr>
<td>mimo: Contrabass clarinet</td>
</tr>
</tbody>
</table>

MIMO Thesaurus

<table>
<thead>
<tr>
<th>Greek</th>
<th>English</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>mimo</td>
<td>Violí piccolo</td>
<td>i 1 0</td>
</tr>
<tr>
<td>mimo</td>
<td>Pianola</td>
<td>i 0 1</td>
</tr>
<tr>
<td>mimo</td>
<td>Dwarspiano</td>
<td>i 1 0</td>
</tr>
<tr>
<td>mimo</td>
<td>Clarinet</td>
<td>i 1 0</td>
</tr>
<tr>
<td>mimo</td>
<td>Tsambouna</td>
<td>i 1 0</td>
</tr>
<tr>
<td>mimo</td>
<td>Qanun</td>
<td>i 0 1</td>
</tr>
<tr>
<td>mimo</td>
<td>Percussion instruments</td>
<td>i 1 0</td>
</tr>
<tr>
<td>mimo</td>
<td>Violine</td>
<td>i 1 0</td>
</tr>
<tr>
<td>mimo</td>
<td>Harp</td>
<td>i 0 0</td>
</tr>
<tr>
<td>mimo</td>
<td>Flute</td>
<td>i 1 0</td>
</tr>
</tbody>
</table>
Music Item Annotated with MIMO Tags

Radio Oranje 04-11-1944 - minister J. van de Broek over geldcirculatie
Netherlands Institute for Sound and Vision

Start typing an instrument

Instruments tagged
- mimo : Baritone
- mimo : Baritone bugle
- mimo : Dl
- mimo : Kiasir
- mimo : Guitars
- mimo : Guitar
- mimo : Tar
- mimo : Zang
- mimo : Zang
Inspiring Users with Gamification Features

Progress monitoring - goal achievement

Help us reach our goal

The objective is to have musical instruments identified in at least 30,000 recordings from our archives by the end of the crowdsourcing campaign (31st of January). The tags are all controlled and available in many languages.

5872 / 30000 Tags

Badges

Dynamic Leaderboard

<table>
<thead>
<tr>
<th>Tags</th>
<th>Leaderboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>annnc</td>
<td>2304</td>
</tr>
<tr>
<td>Clémence</td>
<td>838</td>
</tr>
<tr>
<td>maths28</td>
<td>514</td>
</tr>
<tr>
<td>rakhid</td>
<td>174</td>
</tr>
<tr>
<td>valia.vraka</td>
<td>110</td>
</tr>
<tr>
<td>tzervou</td>
<td>93</td>
</tr>
<tr>
<td>ropiat</td>
<td>87</td>
</tr>
<tr>
<td>Spyros B</td>
<td>71</td>
</tr>
<tr>
<td>vera.kriezi</td>
<td>69</td>
</tr>
<tr>
<td>eleni</td>
<td>58</td>
</tr>
</tbody>
</table>
## Campaign Statistics

<table>
<thead>
<tr>
<th><strong>Annotations</strong></th>
<th><strong>Duration</strong>: 1 month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotations added: 5872</td>
<td><strong>Annotators</strong>: 76</td>
</tr>
<tr>
<td>Tracks annotated: 2035</td>
<td></td>
</tr>
<tr>
<td>Number of different annotations: 63</td>
<td></td>
</tr>
<tr>
<td>Mean annotation frequency: 71.44</td>
<td></td>
</tr>
<tr>
<td>Median annotation frequency: 20.0</td>
<td></td>
</tr>
<tr>
<td>Max annotation frequency: 651</td>
<td></td>
</tr>
<tr>
<td>Min annotation frequency: 1*</td>
<td></td>
</tr>
</tbody>
</table>

*There are 12 annotations which appear only once in the dataset while 26 annotations appear less than 10 times.*

<table>
<thead>
<tr>
<th><strong>Annotations per Track</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean annotations per track: 2.28</td>
<td></td>
</tr>
<tr>
<td>Median annotations per track: 2.0</td>
<td></td>
</tr>
<tr>
<td>Max annotations per track: 24</td>
<td></td>
</tr>
</tbody>
</table>
Machine intelligence and human intelligence can cooperate and improve each other in a mutually rewarding way.

- Exploit the user obtained annotations for training/improving machine learning algorithms
- Use machine learning methods to validate user acquired labels
- Active learning methodologies for Musical instrument identification
- Design targeted Crowdsourcing campaign with specifically selected content that will serve as informative cases, which will improve performance of automated machine learning system (achieve better performance with less but informative samples)
Ongoing Work

WITH is an evolving ecosystem: new repositories are aggregated, new spaces are created and new features and services are constantly designed and aimed to be deployed.

Some of the features under development are:

Automatic Services:

- New **automatic annotation**s with visual analysis extraction methodologies for image metadata enrichment (e.g. aesthetic assessment of image content for photography enthusiasts and professionals)
- Automatic annotations of music recordings

Crowdsourcing features

- Fully automated crowdsourcing campaign creation
- Introduce advanced features like annotator profiles to assess their expertise
Thank you!