

# 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

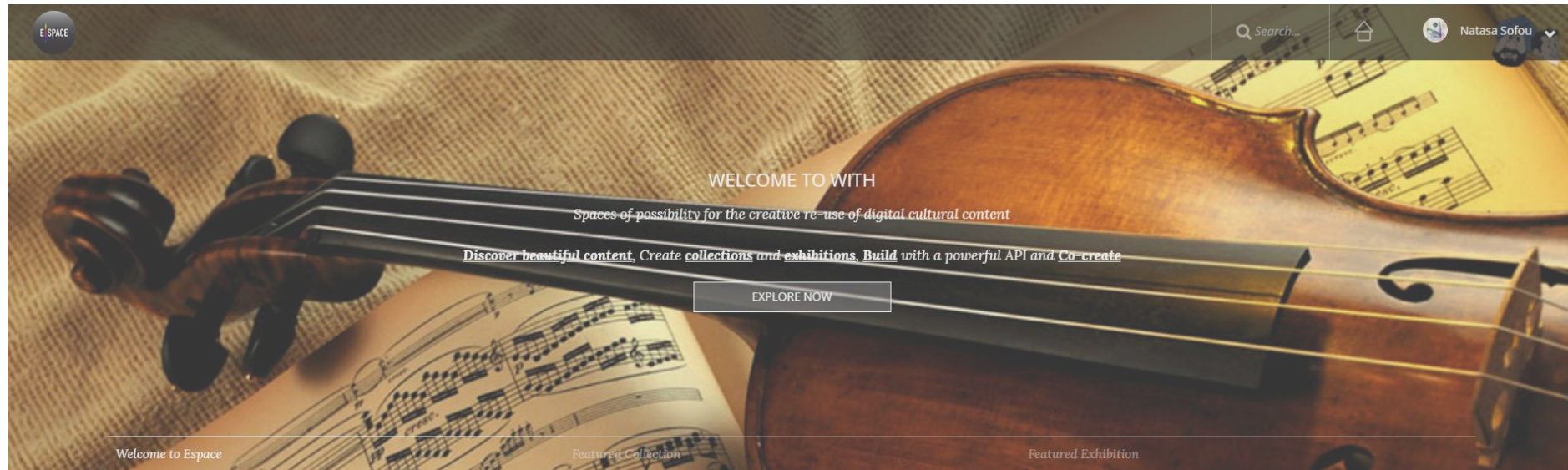
**Intelligent Systems Laboratory, National Technical University of Athens**

# 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



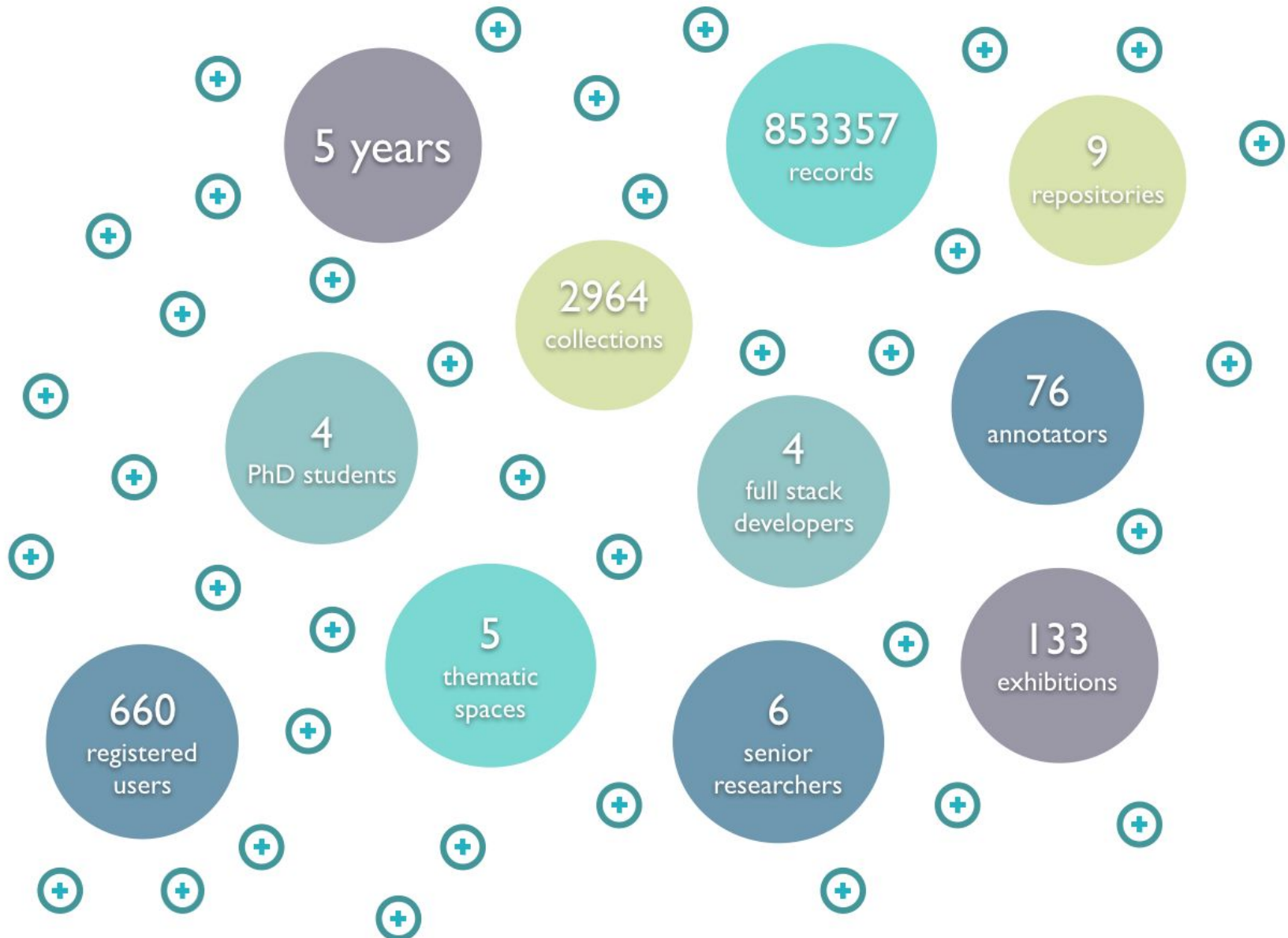
<http://withculture.eu/>

WITH is a cultural ecosystem that:

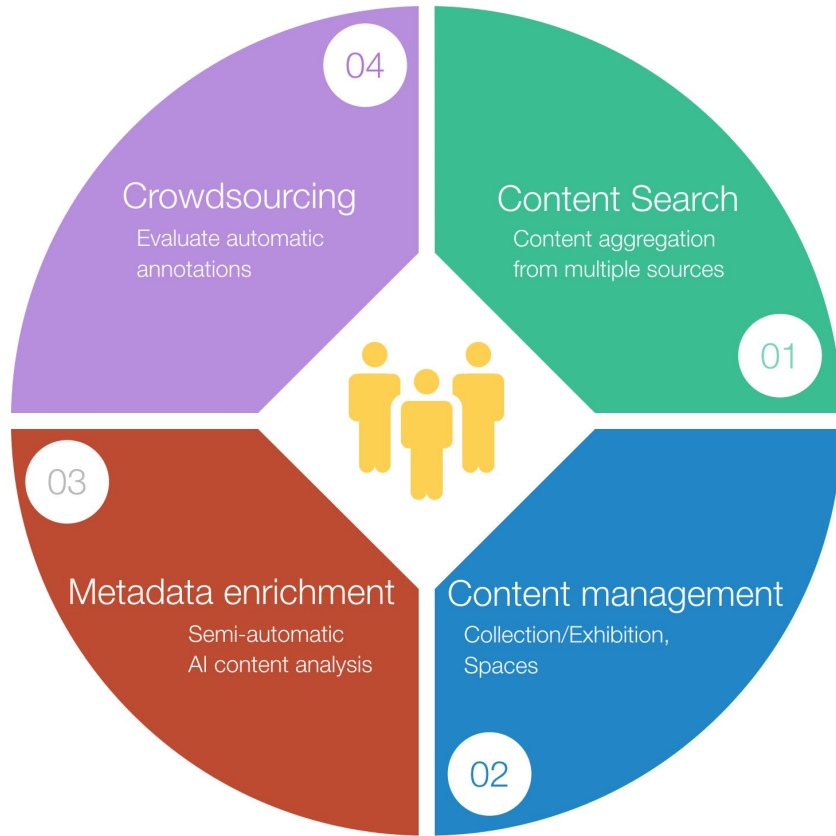
- Exploits cultural heritage content
- Promotes human-computer collaboration
- Provides enhanced services for data/metadata management and enrichment
- Facilitates accessibility and discoverability of available cultural content

# WITH IN NUMBERS

---



# WITH User Engagement



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

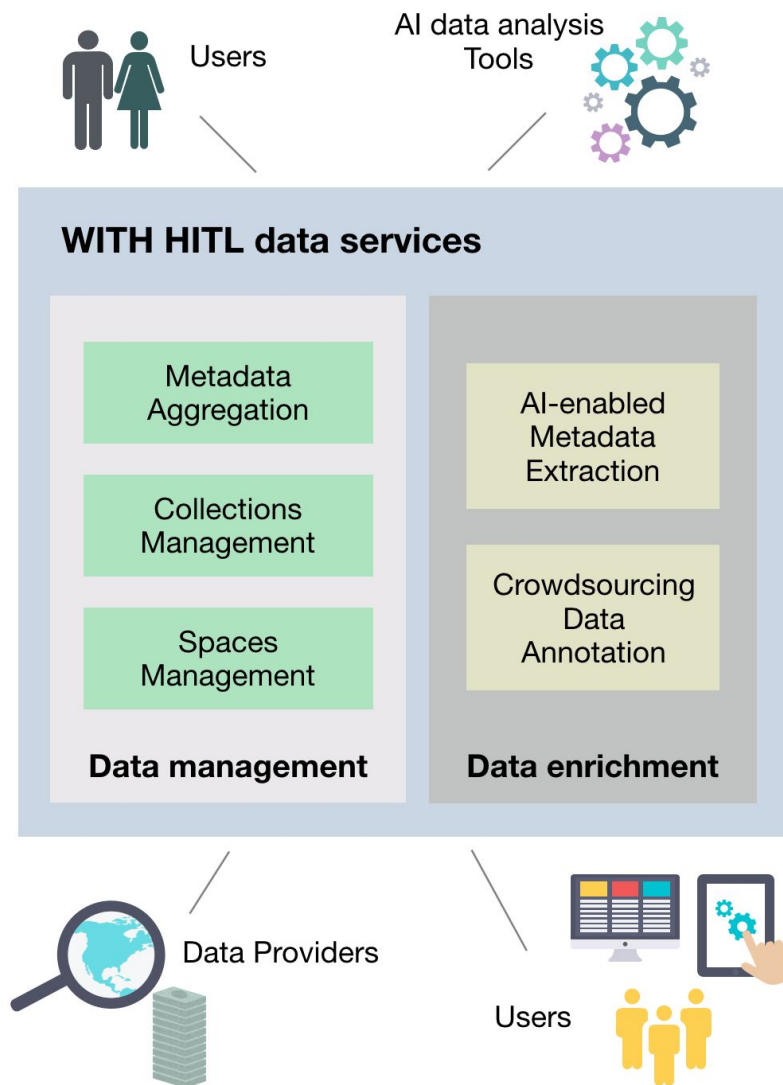
**Metadata Enrichment** and the **Crowdsourcing** processes enable users to advance content descriptions, using AI content analysis tools or human annotations.

# WITH Human Computer Collaboration Services

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**).



**RIJKS MUSEUM**

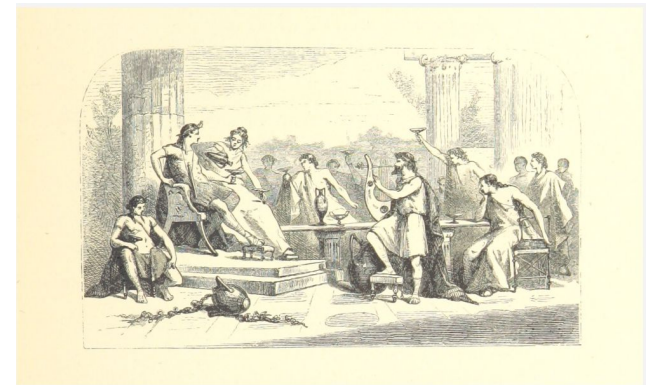
**historypin**



# WITH Data Model

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

```
"descriptiveData": {  
  "label": "Greek from Festival of Song",  
  "description": "This image has been taken from  
Festival  
of Song: a series of Evenings with the Poets",  
  "keywords": [  
    "Greek",  
    "kylix",  
    "lyre",  
    "symposium"  
  ],  
  "isShownAt": "http://www.europeana.eu/api/ANnuDzRpW",  
  "isShownBy":  
"http://farm8.staticflickr.com/7406.jpg",  
  "rdfType": "http://www.europeana.eu/schemas/  
edm/ProvidedCHO",  
  "country": "united kingdom",  
  "dclanguage": "English",  
  "dctype": "scanned image",  
  "dcrights": "Public Domain",  
  "dctermspatial": "New York, 1866",  
  "dcformat": "jpg"  
}
```



Greek from "Festival of Song: a series of Evenings with the Poets. Prepared by the author of "Salad for the Solitary" ... (F. Saunders). With ... pictures by Members of the National Academy of Design. Engraved by Bobbett and Hooper"  
The British Library

INFO RELATED ANNOTATIONS

IMAGE

Record Source

Europeana

Keywords

FI AF

kylix

Language

EN



# 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.



PHOTOCONSORTIUM - INTERNATIONAL CONSORTIUM FOR PHOTOGRAPHIC HERITAGE

*Join us and discover how the camera has, from the very beginning of photographic history, captured the world from its most beautiful angles as well as its most dramatic days.*

EXPLORE NOW

Featured Collections

[View all collections](#)



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



Lthuanian 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.

# Thesauri manager and Linked Data Resources

**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

COLLECT ITEM



Αθηναϊκό μπαλκόνι, 1955  
nationalgallery

INFO RELATED COMMENTS SHARE ANNOTATIONS

Description  
Λάδι σε μουσαμά, 115 x 146 εκ.

Provider  
nationalgallery

Rights  
Creative\_SA

Item url  
<http://withculture.eu/#/item/575692304c747965a122b814>

Record Source

## Annotate

MANUAL AUTOMATIC GEO

Start typing a term

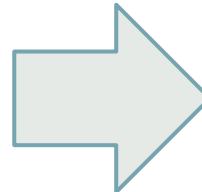
Select vocabularies

Art & Architecture Thesaurus

DBpedia Ontology

DBpedia Resources

EuscreenXL Thesaurus



painting

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

trade

wn30: painting  
<http://wordnet-rdf.princeton.edu/wn30/01792567-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

Annotate Close

Annotate

---

MANUAL **AUTOMATIC** GEO

---

TERM  
DETECTION

NAMED  
ENTITY  
RECOGNITION

**IMAGE  
ANALYSIS**

*Select the image analysis services  
that will be used*

CultIVML Annotator

Annotate Close

# Crowdsourcing Data Annotation

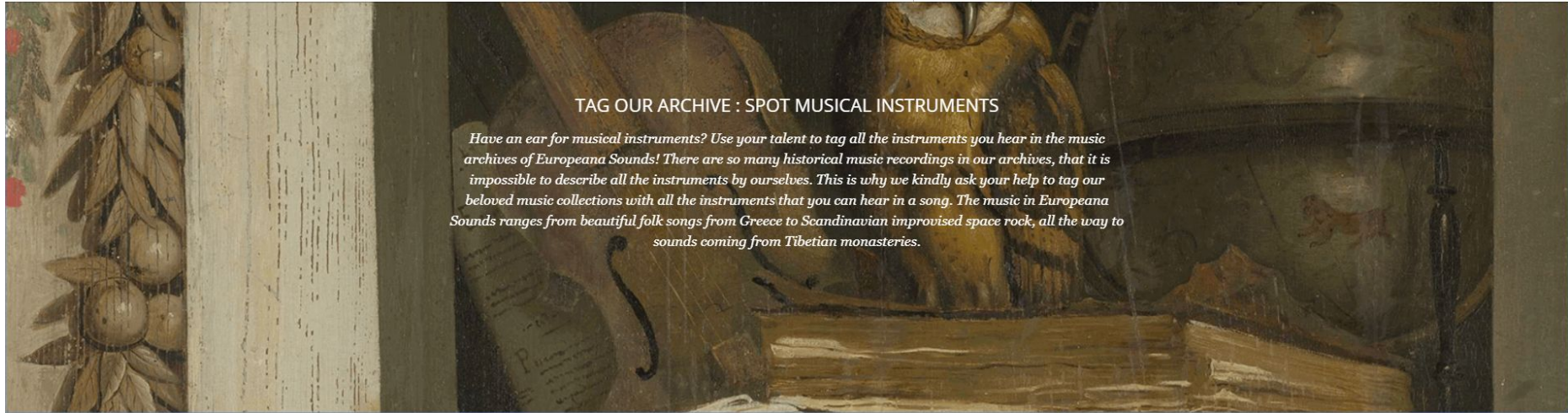
WITH offers a crowdsourcing infrastructure that essentially complements any automatic enrichment.

- **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



## Featured Collections

[View all collections](#)



Music from BNF  
3731 Items - By *anna c*  
2016/04/26



Music from ITMA  
1582 Items - By Sounds Project Sounds  
2016/06/24



Music from Denmark  
591 Items - By *anna c*  
2016/04/26



Music from FMS  
1232 Items - By Sounds Project Sounds  
2016/06/14



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



Music from CNRS-CREM  
7365 Items - By Sounds Project Sounds  
2016/03/19

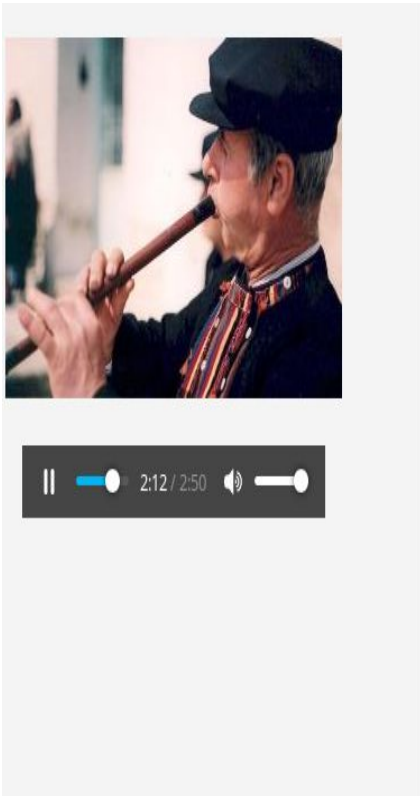
# Defining the Campaign Features

- 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: 30000 tags

# User Identified MIMO Tags



## Καλαματιανός

Music Library of Greece of The Friends of Music Society

INFO RELATED COMMENTS SHARE ANNOTATIONS

clarin	Select Vocabularies ▾
mimo: Triple clarinet	Clarinet
mimo: Idioglot clarinet	Clarinet
mimo: Double clarinet	Clarinet
mimo: Basset clarinet	Clarinet
mimo: Contrabass clarinet	Clarinet

## Καρσιλαμάς

Music Library of Greece of The Friends of Music Society

INFO RELATED COMMENTS SHARE ANNOTATIONS

### MIMO Thesaurus

🔗 mimo : Violí piccolo	i	👍 1	👎 0
🔗 mimo : Pianola	i	👍 0	👎 1
🔗 mimo : Dwarspiano	i	👍 1	👎 0
🔗 mimo : Clarinet	i	👍 1	👎 0
🔗 mimo : Tsambouna	i	👍 1	👎 0
🔗 mimo : Qanun	i	👍 0	👎 1
🔗 mimo : Percussion Instruments	i	👍 1	👎 0
🔗 mimo : Violine	i	👍 1	👎 0
🔗 mimo : Harp	i	👍 0	👎 0
🔗 mimo : Flute	i	👍 1	👎 0

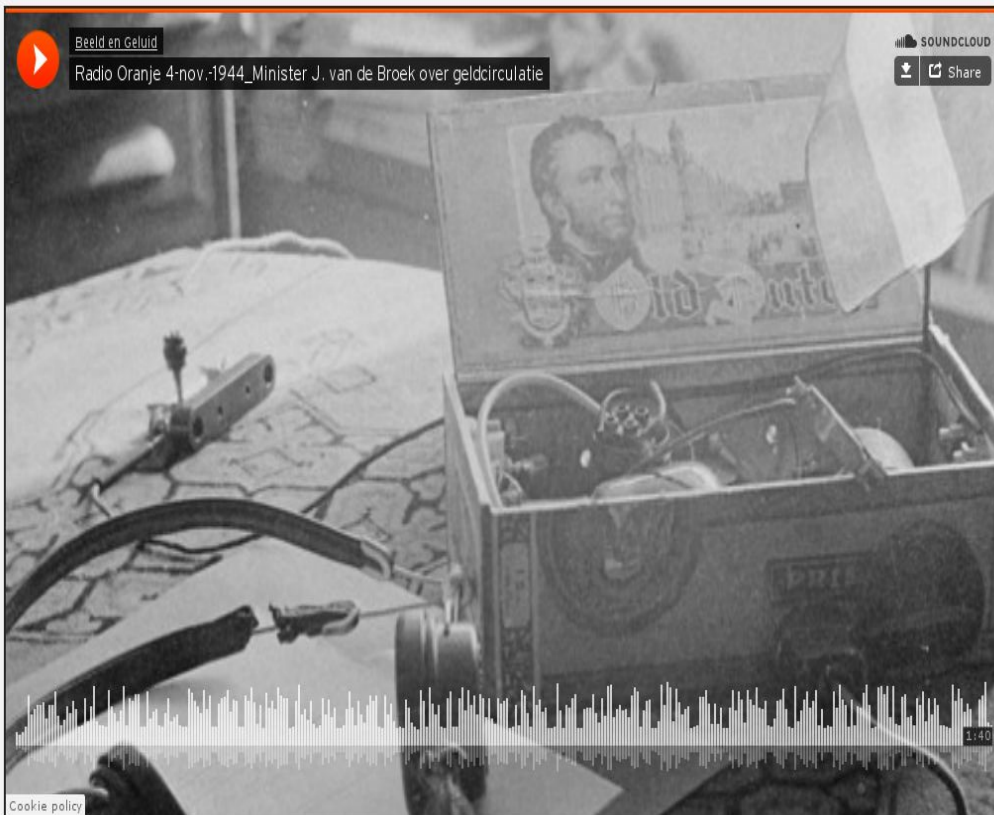
# Music Item Annotated with MIMO Tags

Beeld en Geluid

Radio Oranje 4-nov.-1944\_Minister J. van de Broek over geldcirculatie

SOUNDCLOUD

Share



1:40

Cookie policy

## Radio Oranje 04-11-1944 - minister J. van de Broek over geldcirculatie

NEXT ITEM

Netherlands Institute for Sound and Vision

INFO RELATED COMMENTS SHARE ANNOTATIONS

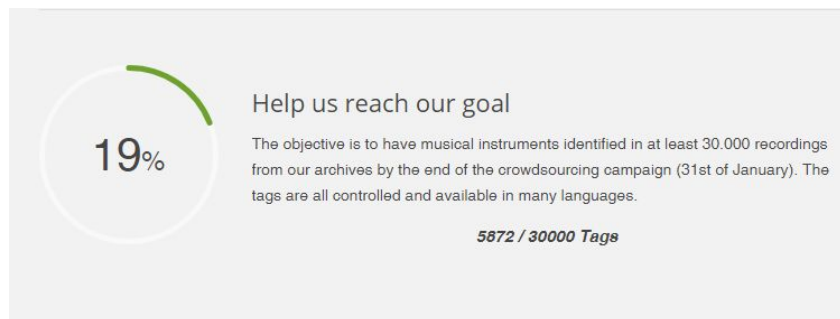
Start typing an instrument

### Instruments tagged

- |                       |   |   |   |
|-----------------------|---|---|---|
| mimo : Baritone       | i | 0 | 0 |
| mimo : Baritone bugle | i | 0 | 0 |
| mimo : Di             | i | 0 | 0 |
| mimo : Kisir          | i | 0 | 0 |
| mimo : Guitars        | i | 0 | 0 |
| mimo : Guitar         | i | 0 | 0 |
| mimo : Tar            | i | 0 | 0 |
| mimo : Tar            | i | 0 | 0 |
| mimo : Zang           | i | 0 | 0 |
| mimo : Zang           | i | 0 | 0 |

# Inspiring Users with Gamification Features












## Progress monitoring - goal achievement



## Badges



## Dynamic Leaderboard

TAGS Leaderboard	
 annac	 2304
 Clémence	838
 maths28	514
 rakhiD	174
 valia.vraka	110
 tzervou	93
 ropiai	87
 Spyros B	71
 vera.kriezi	69
 eleni	58

# Campaign Statistics

## Annotations

Annotations added: 5872

Tracks annotated: 2035

Number of different annotations: 63

Mean annotation frequency: 71.44

Median annotation frequency: 20.0

Max annotation frequency: 651

Min annotation frequency: 1\*

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

**Duration:** 1 month

**Annotators:** 76

## Annotations per Track

Mean annotations per track: 2.28

Median annotations per track: 2.0

Max annotations per track: 24



# Closing the Loop

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 annotations** 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!