Semantic Enrichment Strategy for SearchCulture.gr

Haris Georgiadis
Agathi Papanoti
National Documentation Centre /EKT

Co-financed by the Connecting Europe Facility of the European Union

Danse de trois faunes et trois bacchantes, Hieronymus Hopfer, Bibliothèque municipale de Lyon , Public Domain

housekeeping

Any questions that you have you can add them in the **Q&A** section or to this document **https://bit.ly/30ZwZM2**

Use the **Chat** space to introduce yourself to the group

The recording and relevant documents will be made available on the **event page on Europeana Pro**
https://bit.ly/2zKsZUo or shared with you via **EventBrite**

Survey https://bit.ly/3drpAYq - Thank You!

speakers

Dr Haris Georgiadis

SearchCulture.gr and Semantics.gr Technical Lead
Head of e-Services Department,
National Documentation Centre / EKT
hgeorgiadis@ekt.gr

Agathi Papanoti

Data Ingestion Specialist
National Documentation Centre / EKT
apapano@ekt.gr

overview

- Goal and key challenges
- Pros and cons of two different enrichment approaches
- The SearchCulture.gr enrichment strategy
  - Type enrichment
  - Subject enrichment
  - Time normalization and enrichment
- Outcome: how it affected SearchCulture.gr

europeana

National Documentation Centre/ EKT

NATIONAL DOCUMENTATION CENTRE

eContent
Digital content & services

Digital content and services

We aggregate, collect, document, preserve and disseminate digital content & data, produced and used by the Greek scientific, research and cultural communities.

- **repositories EKT**| DSpace SaaS repository service
- **openABEKT** | SaaS Integrated Library System
- **ePublishing** | OA scholarly publishing suite
- **National Archive of PhD Theses**
- **SearchCulture.gr** | National Aggregator for CH
- **OpenArchives.gr** | Scientific Aggregator

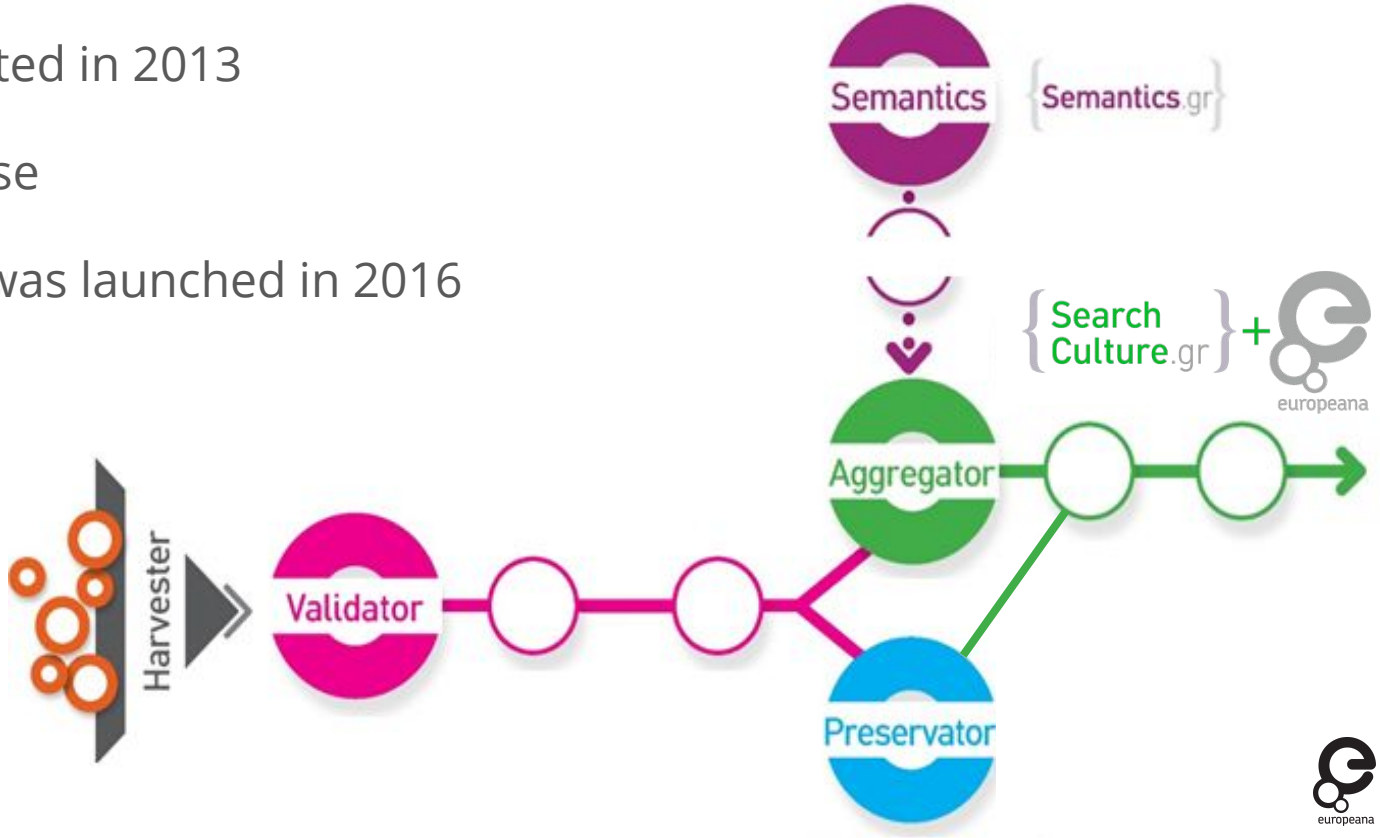
europeana

EKT aggregation infrastructure

Development started in 2013

Developed in-house

SearchCulture.gr was launched in 2016



SearchCulture.gr | Greek National Aggregator

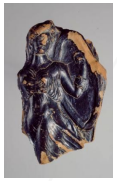
- 85 collections from 66 institutions
- museums I archives I ephorates of antiquities I municipalities I cultural foundations
- 680.000+ digital assets
- archaeological items I historical documents I folklore items I works of art I maps I oral history
- 320.000+ digital assets published in Europeana

europeana

SearchCulture.gr is a cross-domain aggregator

vase

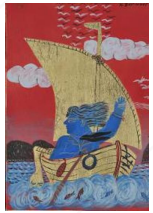
arm



ostracon

coin

jewellery



painting

photo

video

Bronze Age

3000 BC

2000 BC

Middle Bronze Age

Geometric Period

1000 BC

1 AD

Roman Period

Byzantine Period

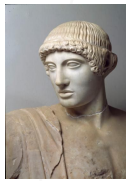
1000

2000

Modern Greece

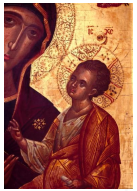
figurine

disk



sculpture

inscription



icon

manuscript

newspaper

audio

What were we trying to achieve?

- **browsing** by item types, subjects, time spans and historical periods
- **search** with time criteria; year ranges or historical periods
- **search** by choosing types and/or subjects
- **facets** on types, subjects, year ranges and historical periods
- submit combined **queries** like
  - icons from the late byzantine period
  - photographs from the WW2
  - paintings depicting the 1821 Revolution
  - manuscripts dating from 1850 to 1860
  - sculptures dated strictly within the Greek middle classical period

europeana

Metadata heterogeneity

Key challenges for large scale cultural aggregators

europeana

Problems (1)

- Too detailed vs too generic documentation (e.g. “amphora” | “archeological object”)
- Different languages
- Synonyms
- Plural vs singular (e.g. “photograph” | “photographs”)
- Small vs capital letters (e.g. “BOOK” | ”Book” | ”book”)
- Typos and spelling errors (e.g. “archaeology” | “archeology”)
- Lack of URI references to controlled vocabularies or thesauri
- Too specific or event-based subjects which makes disambiguation difficult without context
- Item types used as subjects

Original Item Types

Original Subjects

Problems (2)

- Historical periods vs. chronologies
- Descriptive expressions in chronologies
- Different date/time formats
- Different granularity of time (dates, years, centuries)
- Misuse of dc:date, dcterms:created, dcterms:temporal and dcterms:issued

Original Time Values

europeana

Pick your enrichment strategy

fully automated strategy **vs.** “exhaustive classification” strategy

how

- Use **existing large vocabularies** (Getty AAT, wikidata, LCSH)
- **fully automated** mapping on distinct values or directly on records (string similarity matching, machine learning)

pros + cons

✓ Cheap: fully automated process
✓ Searchability is improved at some extend

? Are they covering target languages and/or country specific aspects (e.g. event-based local history concepts)
✘ Large ratio of invalid enrichments
✘ The enrichment coverage remains low
⇨ Not appropriate for browsing
⇨ Discoverability and visual presentation are not improved

how

- **Create more compact vocabularies** and link them to popular ones such as Getty AAT or LCSH
  - **Greek labels and Greek-aspect**
- **Semi-automated mapping** on distinct values: automated mapping (string similarity matching, machine learning) combined with a **curation** phase (manual mappings)
- Enrich per collection

pros + cons

✓ Searchability is improved
✓ The enrichment coverage is very high
✓ Curation improves quality: lower ratio of invalid enrichments
⇨ Appropriate for browsing
⇨ Discoverability is improved drastically

✘ More expensive: labour-intensive process

europeana

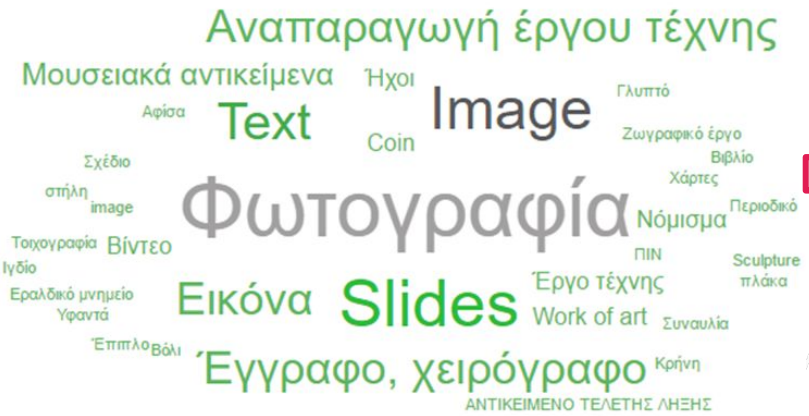
Enrichment process

Our semantic enrichment method

- **Semantics.gr**: a platform developed by EKT where institutions can create, establish and publish LOD vocabularies, taxonomies, thesauri and authority files (SKOS but also other schemas as well: MADS/RDF, EDM Contextual Classes)
  - Vocabulary of 139 item types
  - Unesco Thesaurus - EKT version (1400 concepts) + a Vocabulary with 500 thematic tags
  - Vocabulary of 94 Greek historical periods
- **Enrichment tool** of **Semantics.gr**: a tool for setting *enrichment mapping rules* from distinct metadata values to vocabulary terms
  - Mappings per collection
  - Semi-automatic (automatic suggestions and curation) and self-improving
  - Mappings can be based on one or more metadata fields
- **Time normalization tool**: a tool for setting parametric normalization patterns of time values

europeana

Type enrichment | Our goal



Each record is enriched with a new field:
EKT Type: from our vocabulary of types

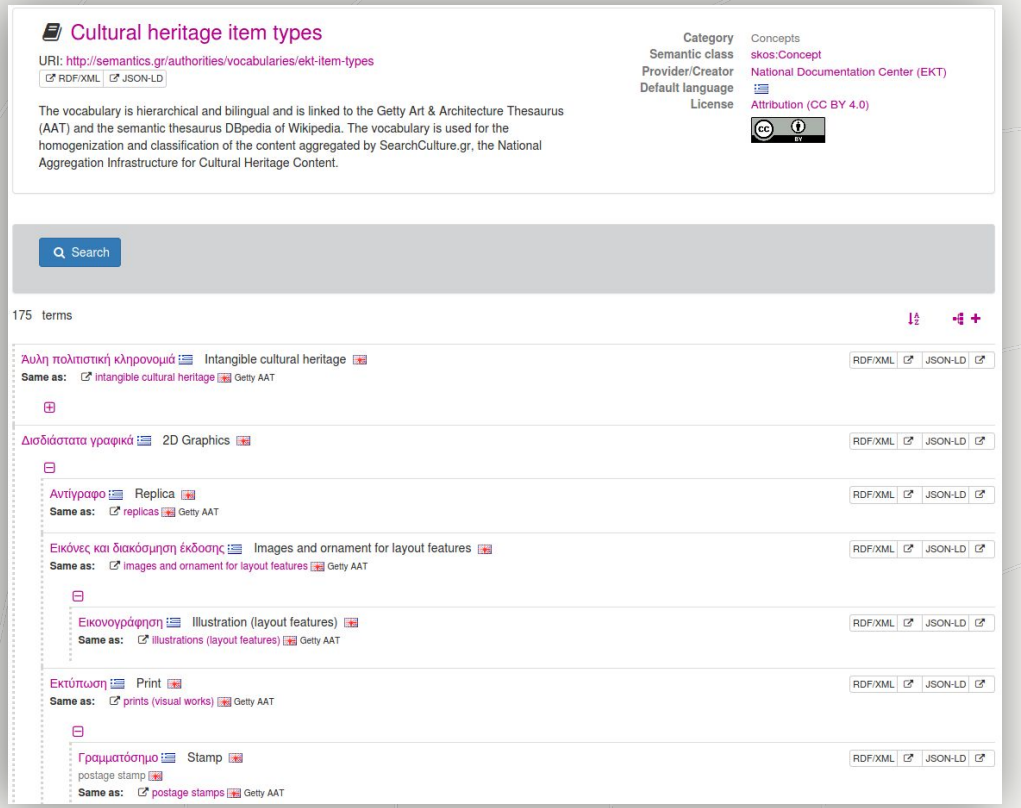
europeana

Type enrichment | The vocabulary ‘EKT Types’

https://www.semantics.gr/authorities/vocabularies/ekt-item-types?language=en

Cultural heritage item types

- 139 item types
- Hierarchical
- Bilingual (Greek and English)
- Linked to Getty AAT
- Schema: skos:Concept



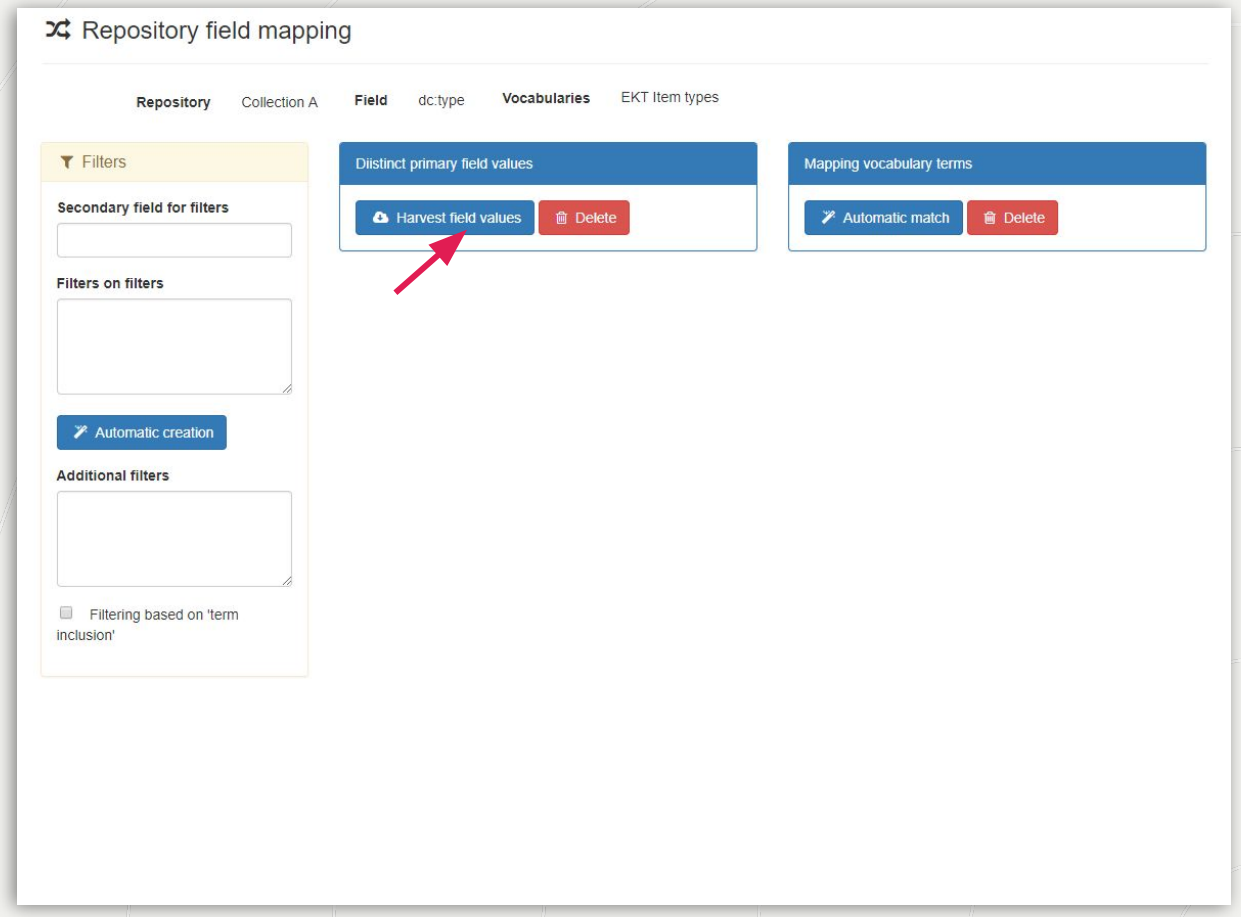
Type enrichment | Steps per collection

1. Insert the collection in the **acceptance portal**
2. Inspect the documentation quality of types
    - is **dc:type** sufficient to set mappings on? (primary field)
    - should we use **OAI setSpecs** instead? (primary field)
    - should we use values of **dc:subject** as filters? (secondary field)
    - should we use keywords in **dc:title** as filters? (secondary field)
3. Create **mapping rules** in the **Enrichment tool** of Semantics.gr
4. Re-index the collection in the **acceptance portal and check**
5. Insert (or re-index) and publish the collection in SearchCulture.gr

europeana

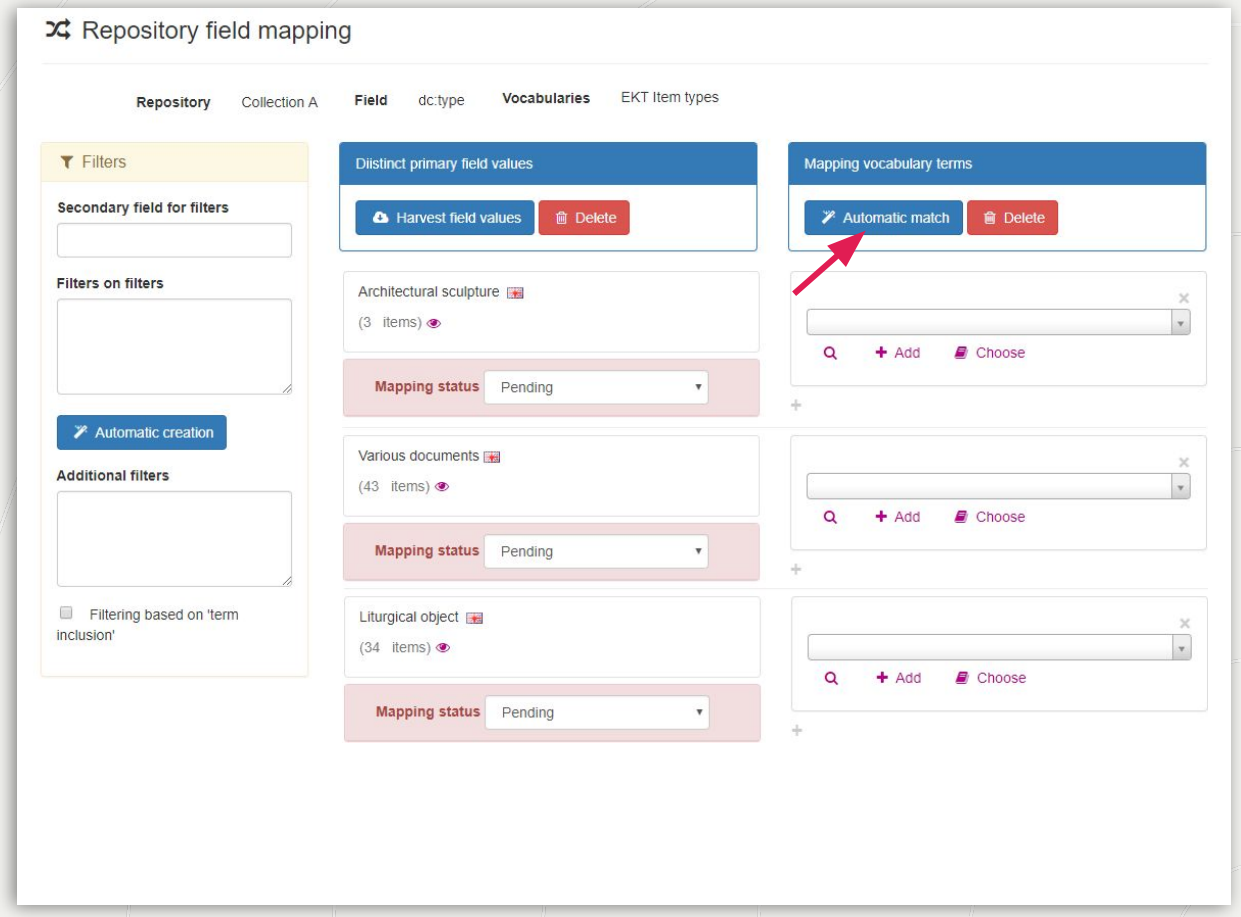
Type enrichment | mappings on dc:type values (1)

- **Good documentation of dc:type**
- **Primary field:** dc:type
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



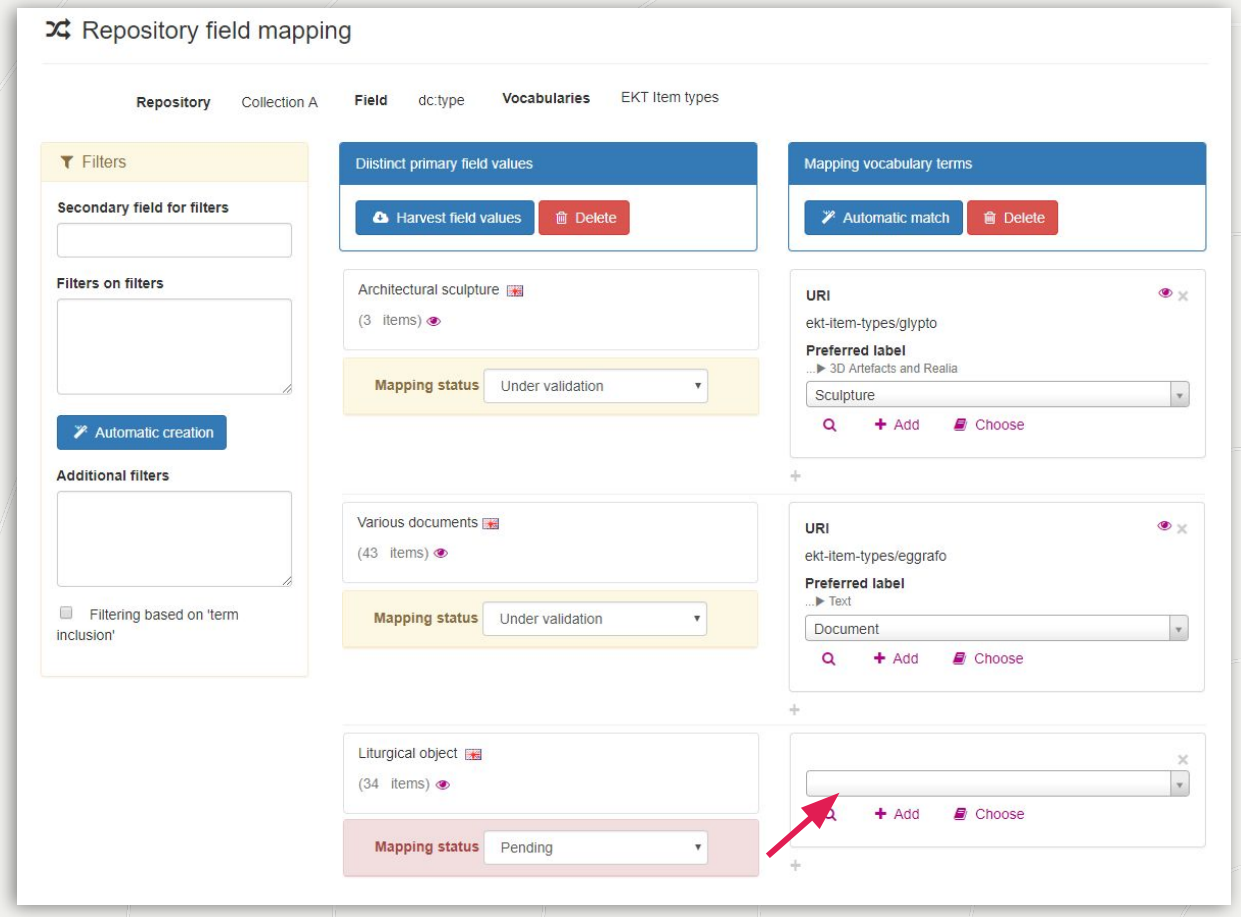
Type enrichment | mappings on dc:type values (2)

- **Good documentation of dc:type**
- **Primary field:** dc:type
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



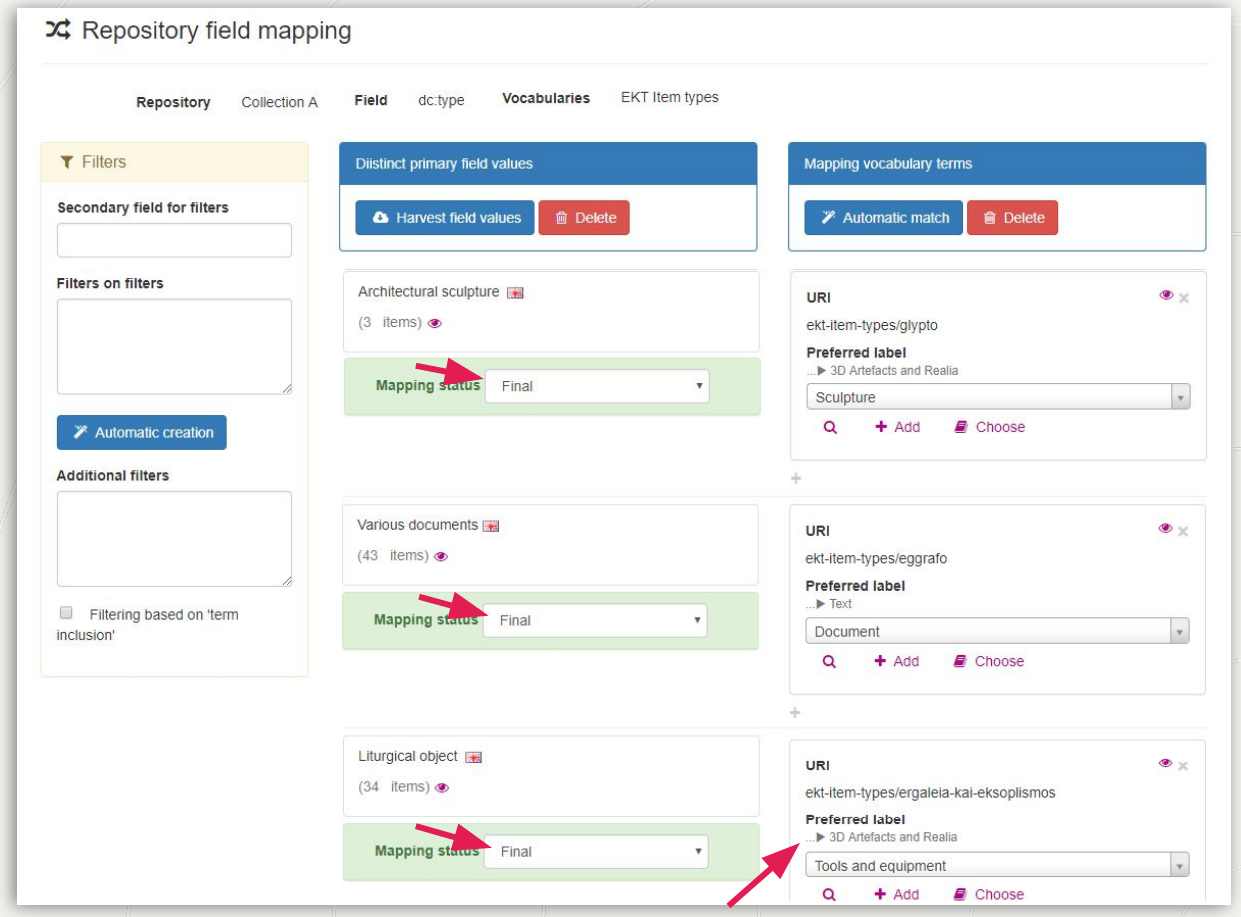
Type enrichment | mappings on dc:type values (3)

- **Good documentation of dc:type**
- **Primary field:** dc:type
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



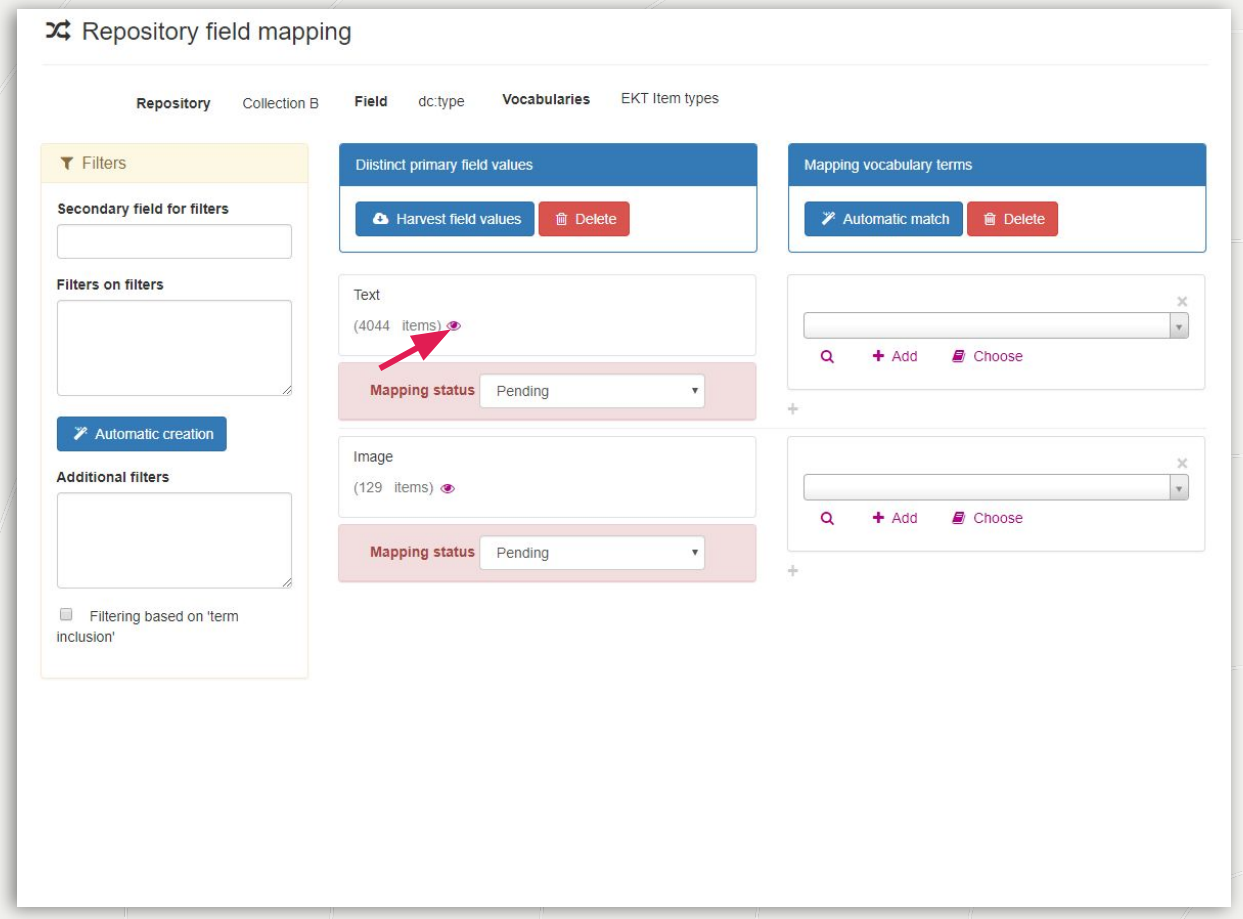
Type enrichment | mappings on dc:type values (4)

- **Good documentation of dc:type**
- **Primary field:** dc:type
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



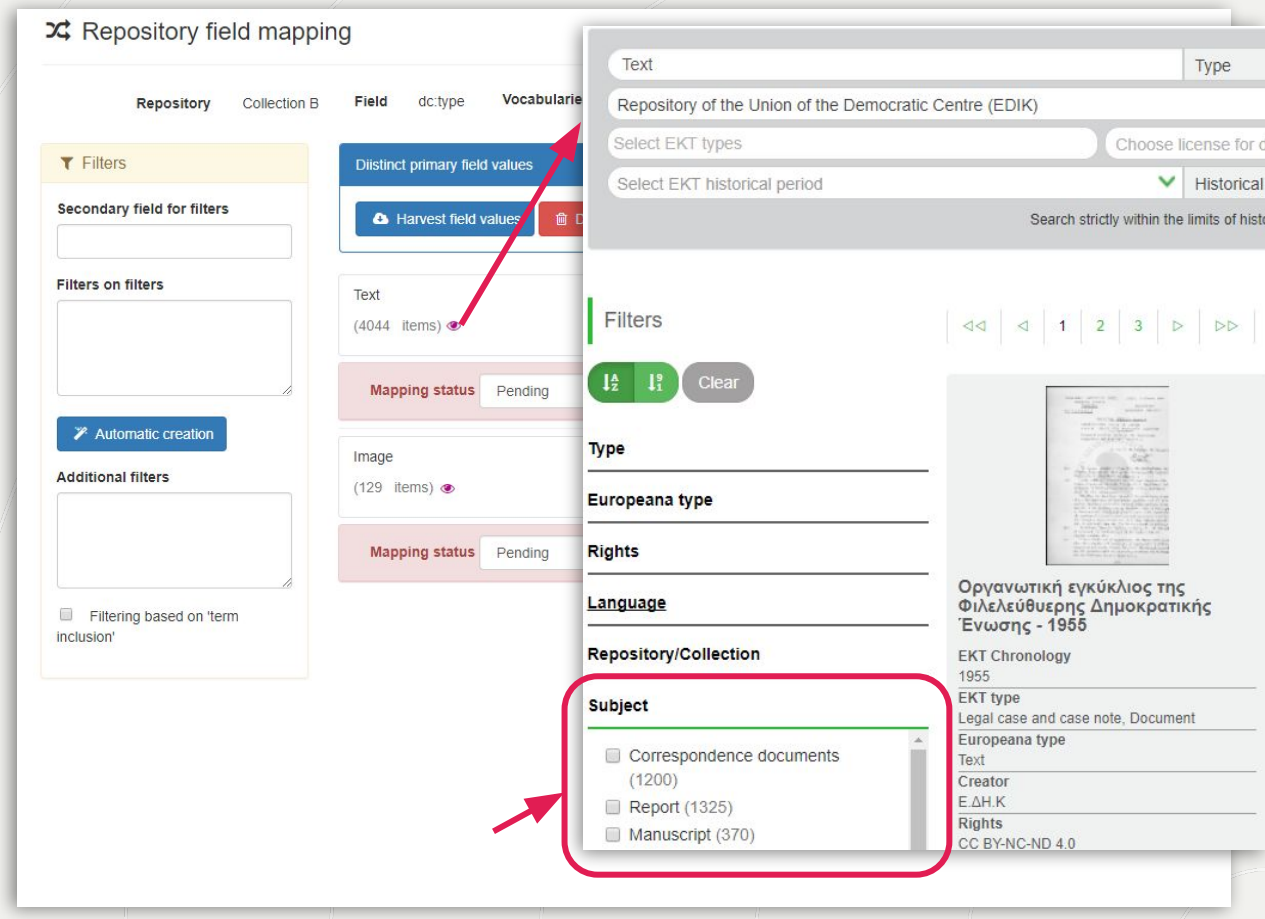
Type enrichment | mappings on dc:type & dc:subject values (1)

- **Poor documentation of dc:type, useful dc:subject**
- **Primary field**: dc:type
- **Secondary field**: dc:subject for filters
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



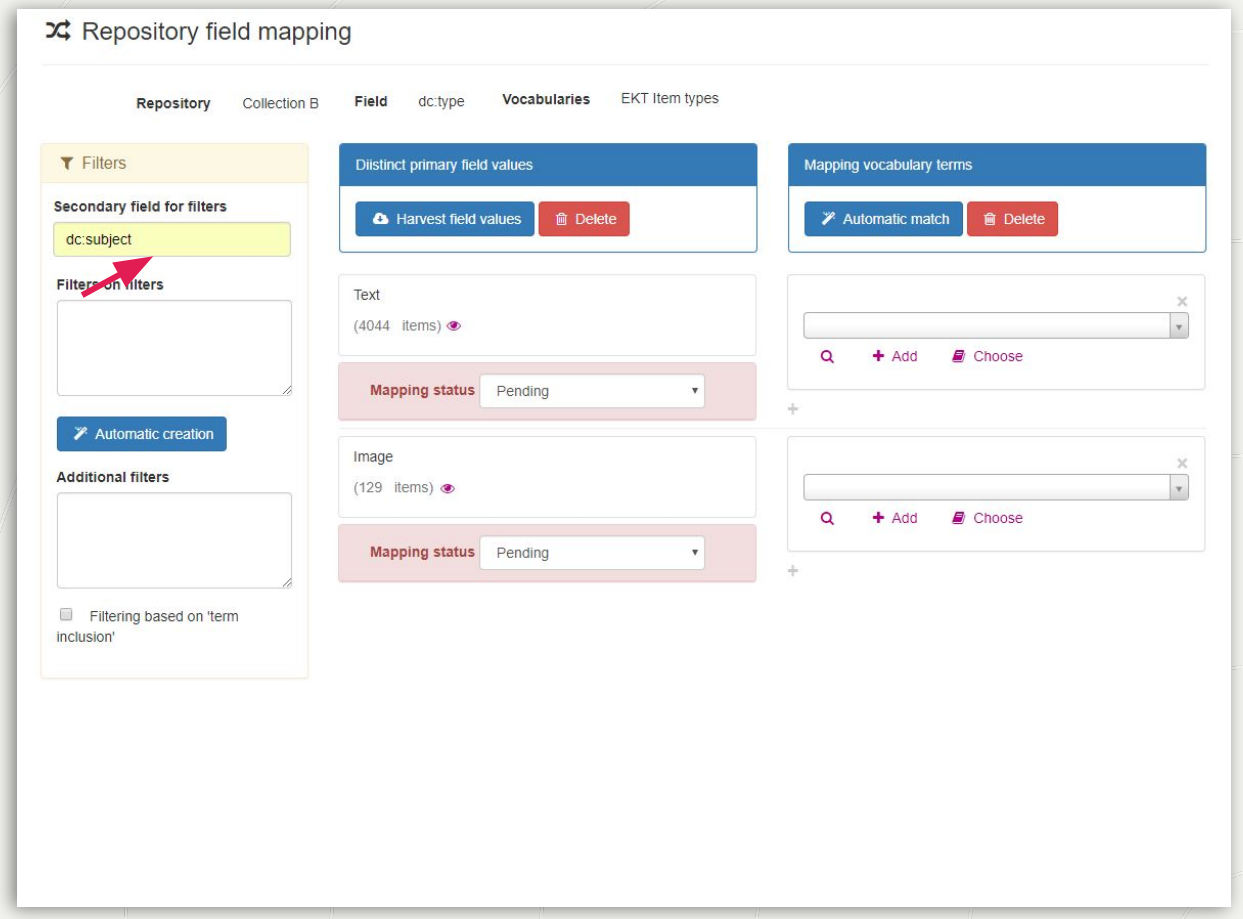
Type enrichment | mappings on dc:type & dc:subject values (2)

- **Poor documentation of dc:type, useful dc:subject**
- **Primary field**: dc:type
- **Secondary field**: dc:subject for filters
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



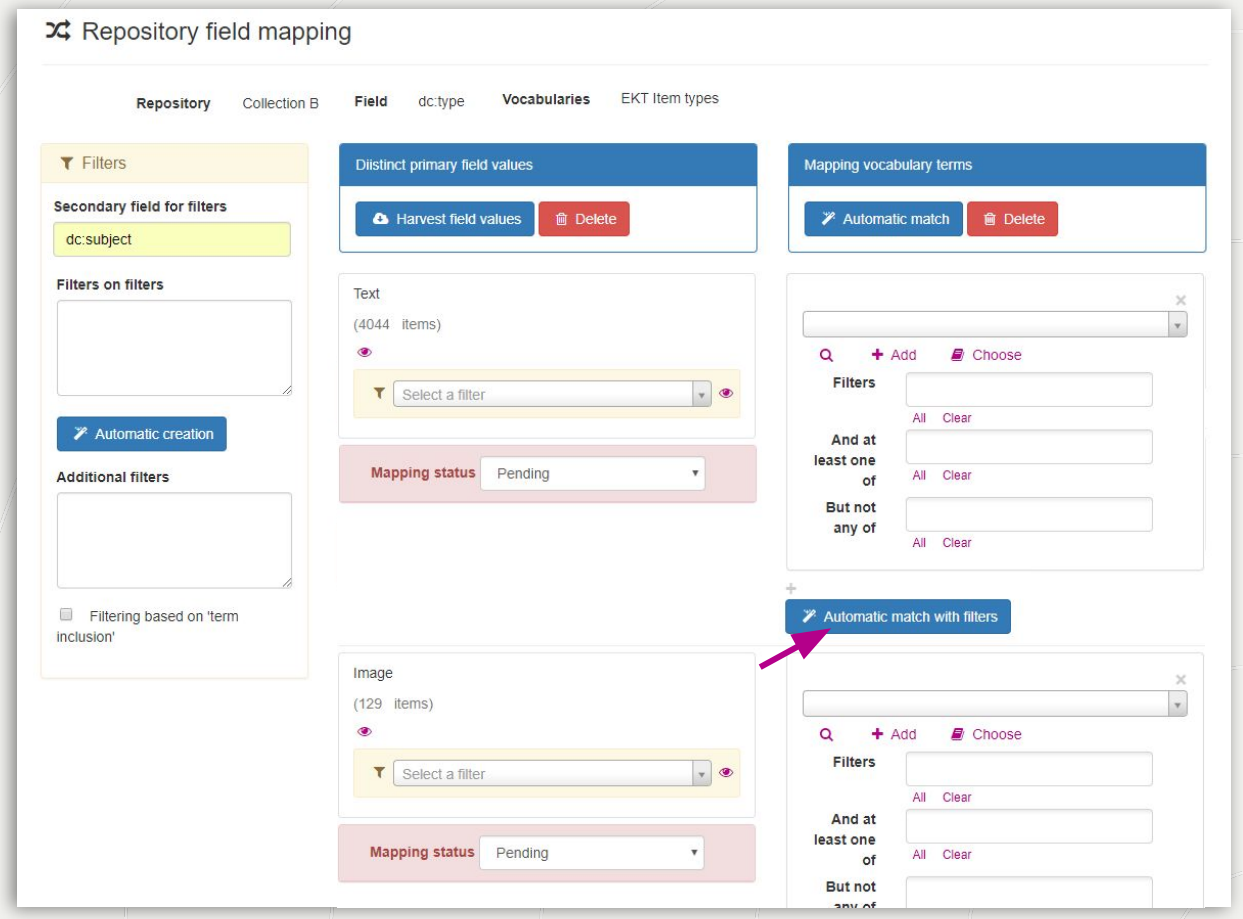
Type enrichment | mappings on dc:type & dc:subject values (3)

- **Poor documentation of dc:type, useful dc:subject**
- **Primary field**: dc:type
- **Secondary field**: dc:subject for filters
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



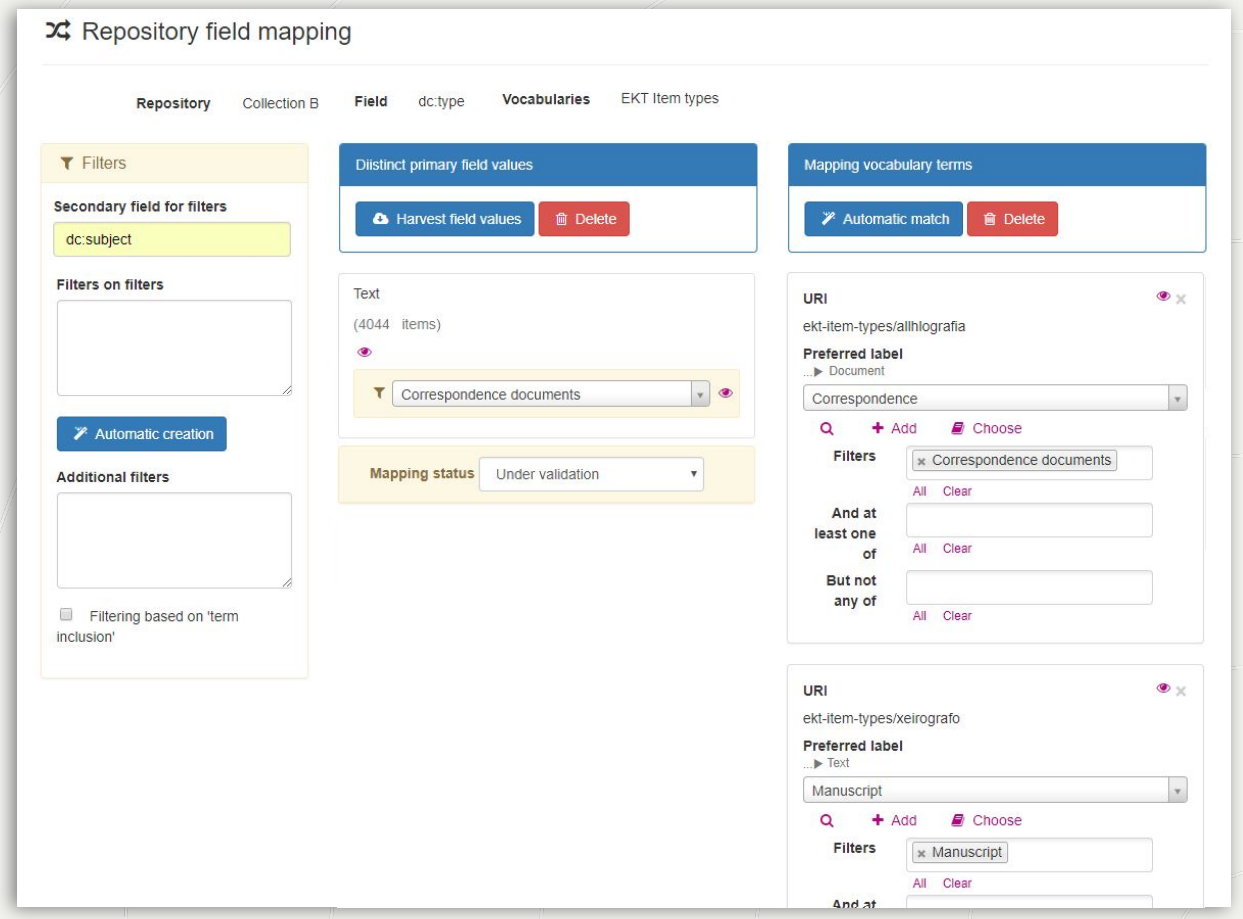
Type enrichment | mappings on dc:type & dc:subject values (4)

- **Poor documentation of dc:type, useful dc:subject**
- **Primary field**: dc:type
- **Secondary field**: dc:subject for filters
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



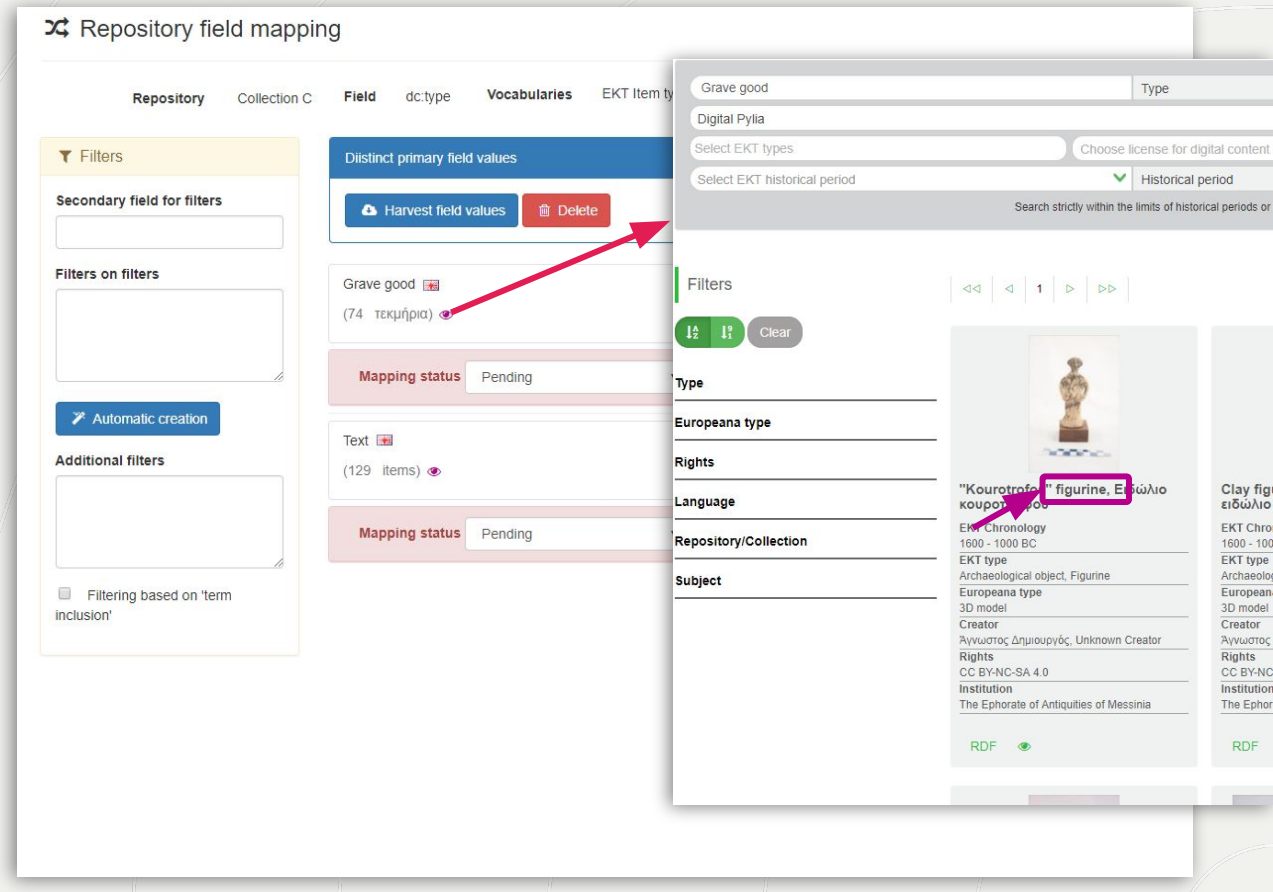
Type enrichment | mappings on dc:type & dc:subject values (5)

- **Poor documentation of dc:type, useful dc:subject**
- **Primary field**: dc:type
- **Secondary field**: dc:subject for filters
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



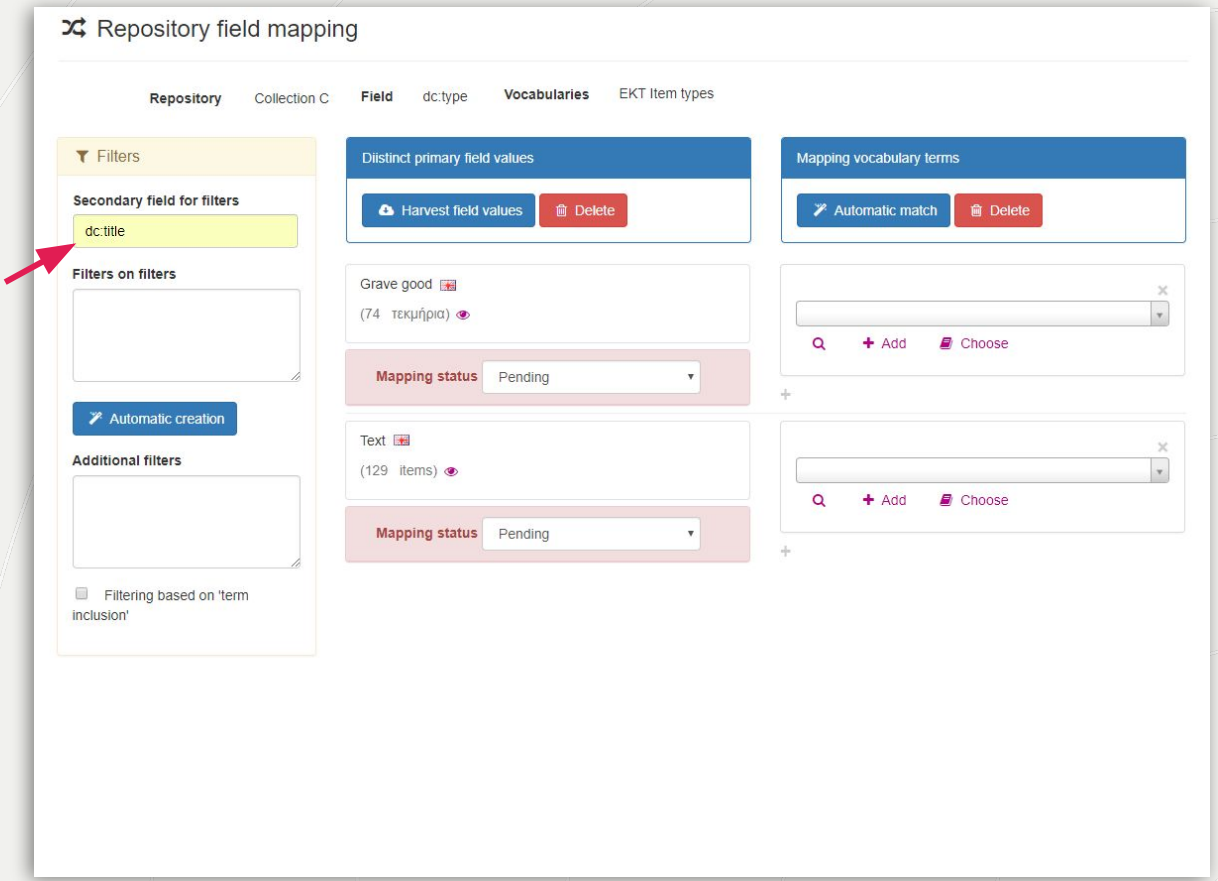
Type enrichment | mappings on dc:type values & dc:title keywords (1)

- **Poor documentation of dc:type, useful dc:title**
- **Primary field:** dc:type
- **Secondary field:** dc:title Keywords for filters
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



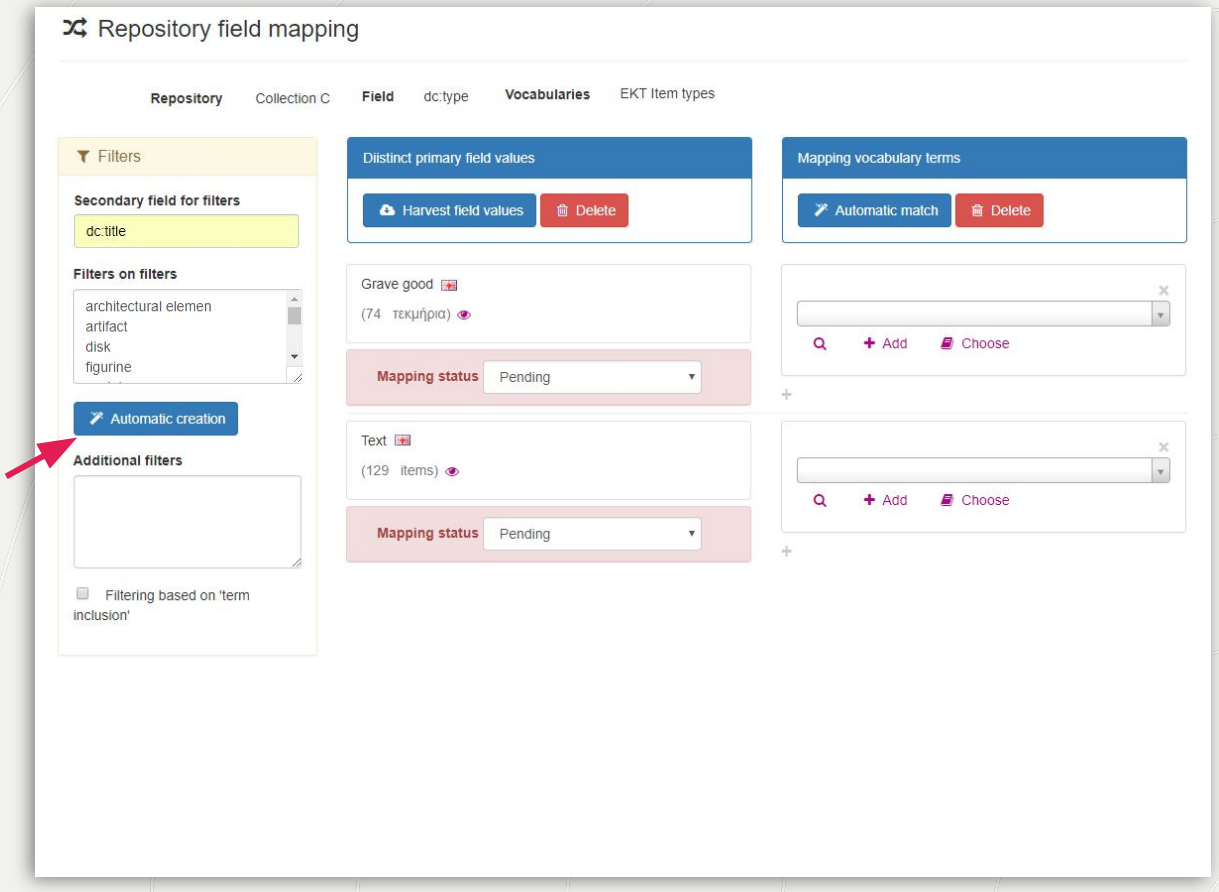
Type enrichment | mappings on dc:type values & dc:title keywords (2)

- **Poor documentation of dc:type, useful dc:title**
- **Primary field:** dc:type
- **Secondary field:** dc:title
  Keywords for filters
- Automatic suggestions of mappings
  & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



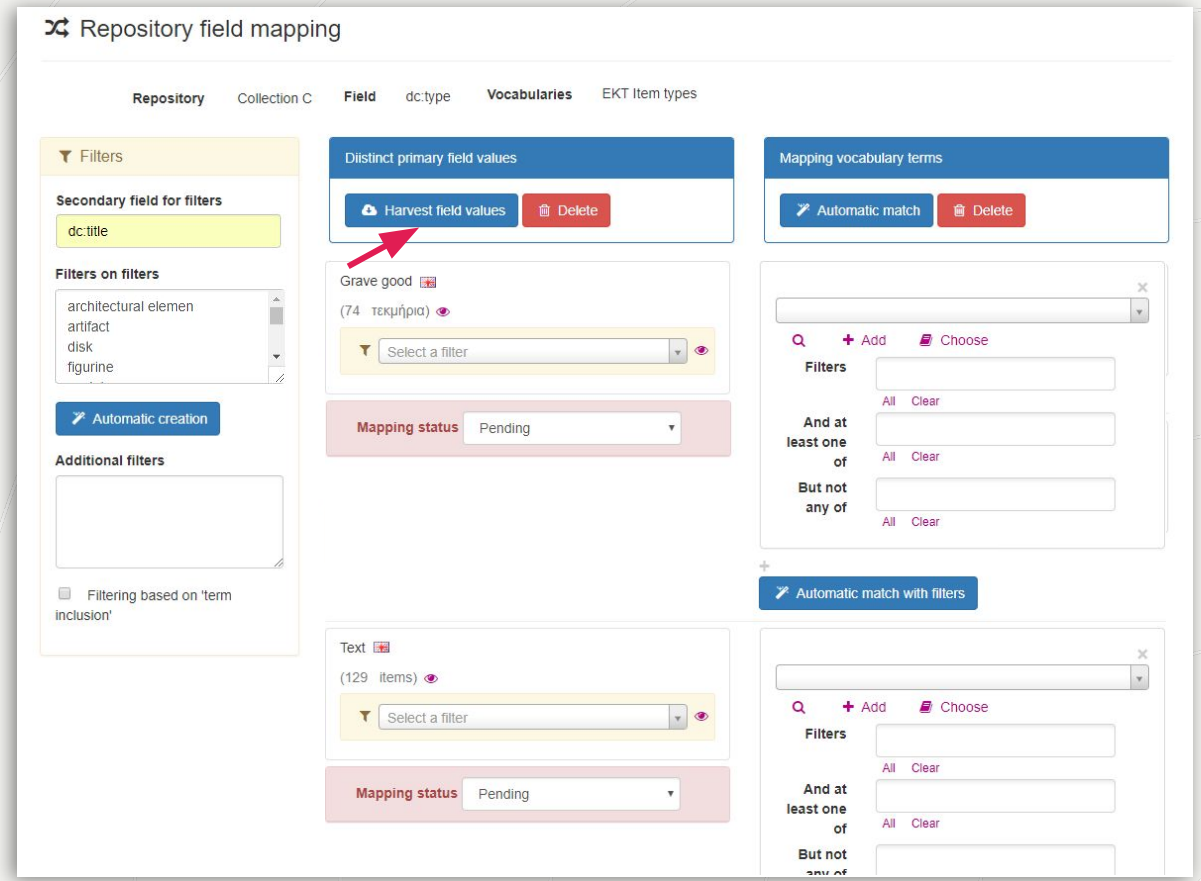
Type enrichment | mappings on dc:type values & dc:title keywords (3)

- **Poor documentation of dc:type, useful dc:title**
- **Primary field:** dc:type
- **Secondary field:** dc:title
  Keywords for filters
- Automatic suggestions of mappings
  & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



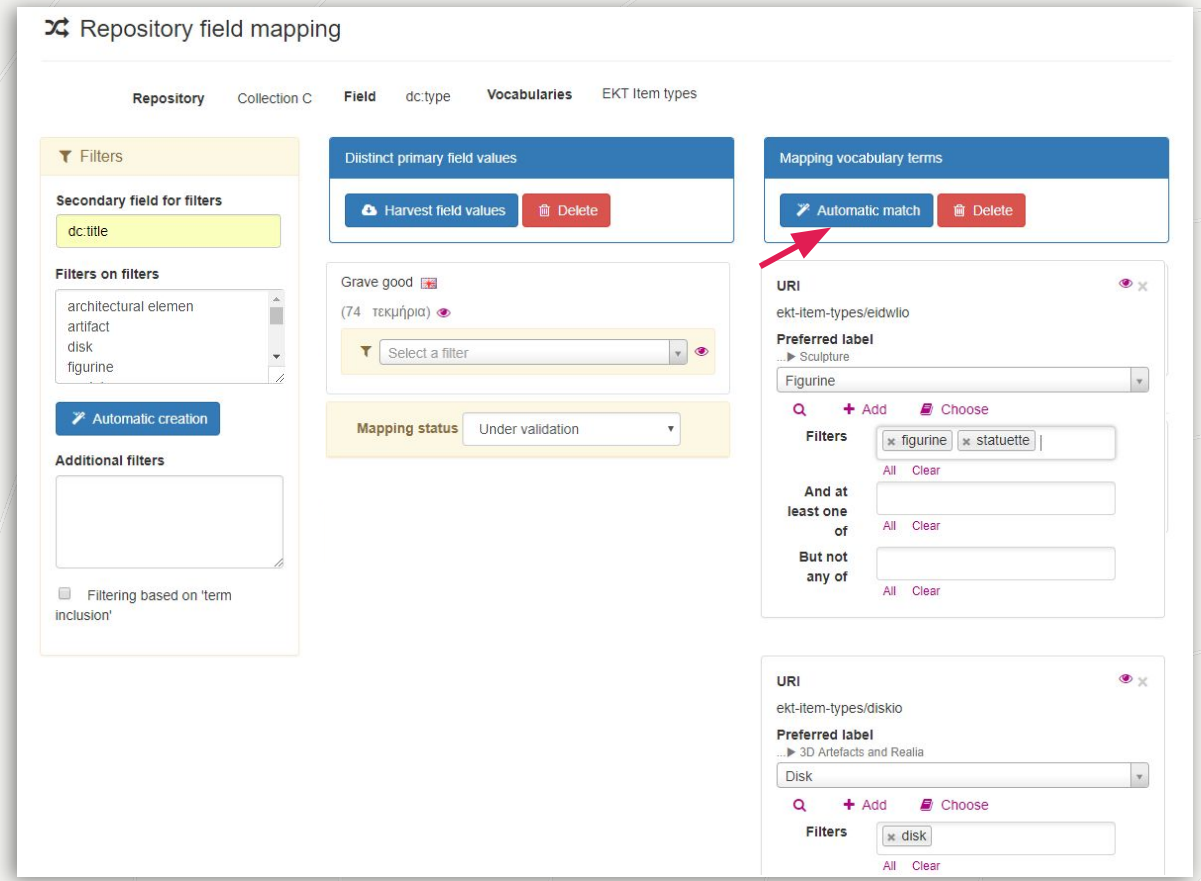
Type enrichment | mappings on dc:type values & dc:title keywords (4)

- **Poor documentation of dc:type, useful dc:title**
- **Primary field:** dc:type
- **Secondary field:** dc:title Keywords for filters
- Automatic suggestions of mappings & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



Type enrichment | mappings on dc:type values & dc:title keywords (5)

- **Poor documentation of dc:type, useful dc:title**
- **Primary field:** dc:type
- **Secondary field:** dc:title
  Keywords for filters
- Automatic suggestions of mappings
  & manual mappings
- The tool ‘remembers’ manual mappings for better future suggestions



Subject enrichment | Our goal

ΕΛ • EN

Each record is enriched with a new field
EKT Subject with terms from **2 vocabularies:**
A vocabulary based on UNESCO thesaurus
A complementary vocabulary with thematic tags

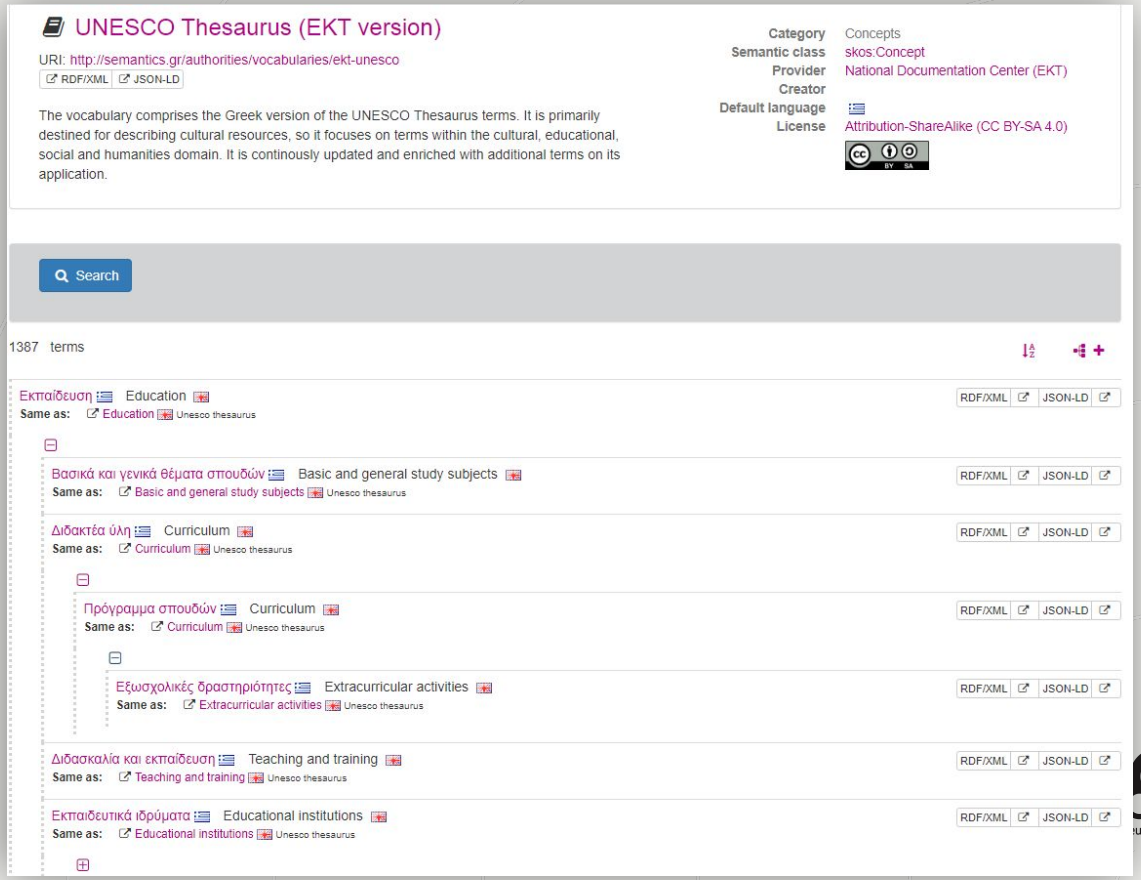
europeana

Subject enrichment | 2 vocabularies are used

http://semantics.gr/authorities/vocabularies/ekt-unesco?lang=en

UNESCO Thesaurus (EKT version)

- **1368** terms
- Hierarchical
- Bilingual (Greek and English)
- Linked to UNESCO Thesaurus
- Schema: skos:Concept



Subject enrichment | 2 vocabularies are used

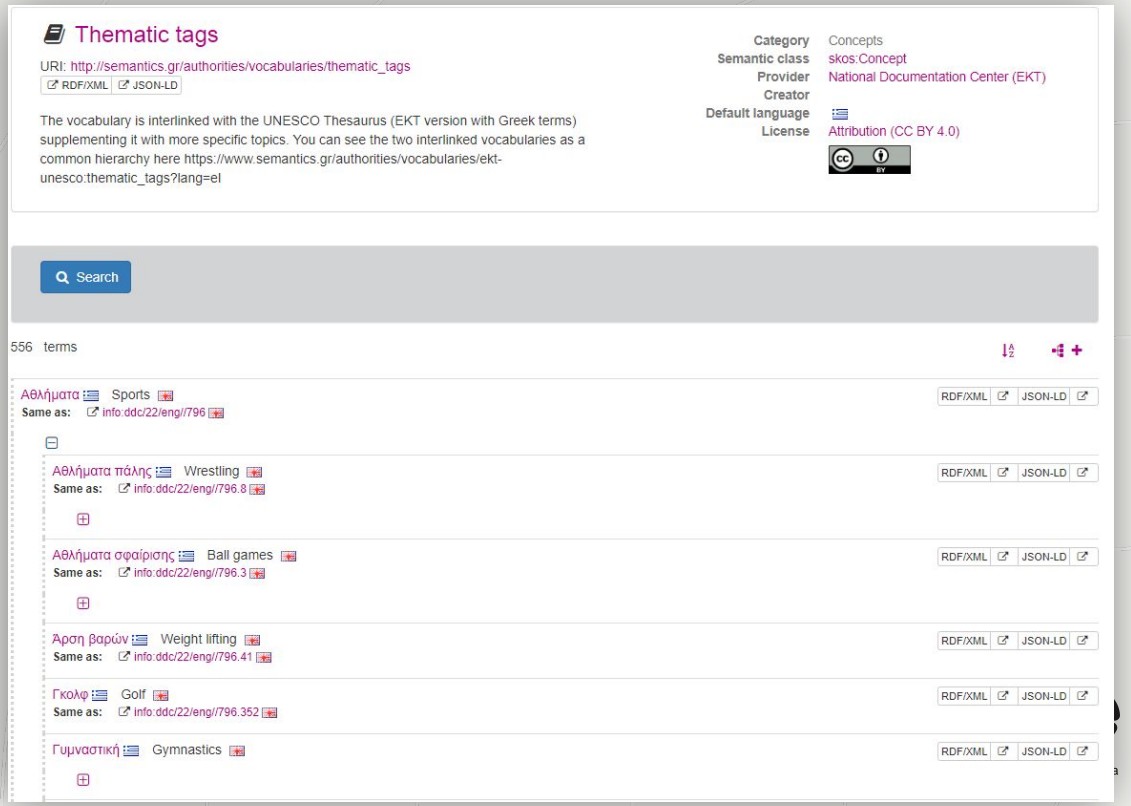
https://www.semantics.gr/authorities/vocabularies/thematic_tags?lang=en

Thematic Tags (by EKT)

- 553 terms
- Hierarchical
- Bilingual (Greek and English)
- Schema: skos:Concept
- Linked to

  UNESCO Thesaurus (EKT version)

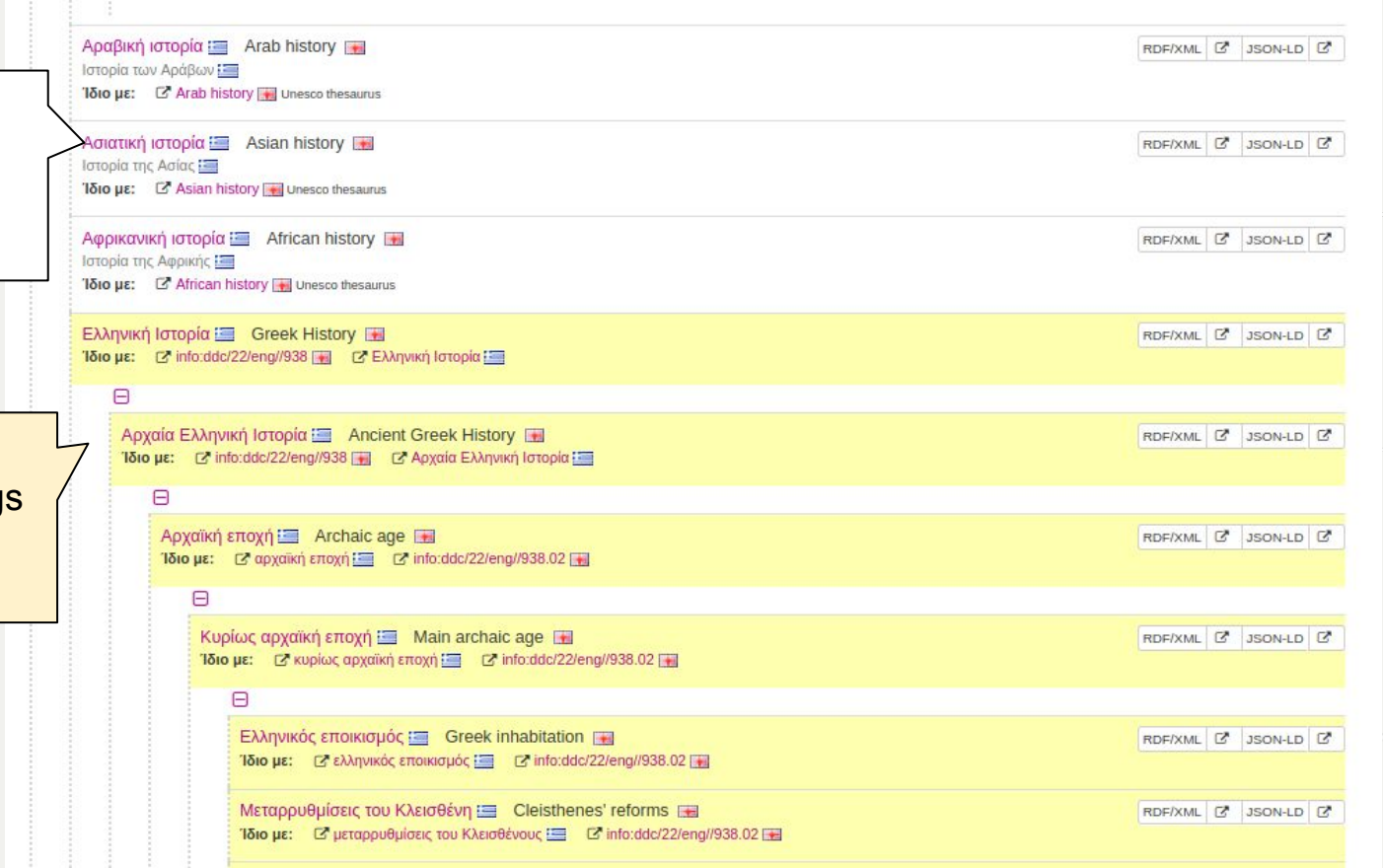
  via **skos:broaderMatch**



Subject enrichment | 2 vocabularies are used

UNESCO Thesaurus (EKT version)

Thematic Tags (by EKT)



europeana

Temporal enrichment | Our goal

Each record is enriched with 2 new fields

- **EKT Chronology**: year/year range (e.g. -31/324)
- **EKT Historical Period**: from a vocabulary of historical periods (e.g. Roman Period)

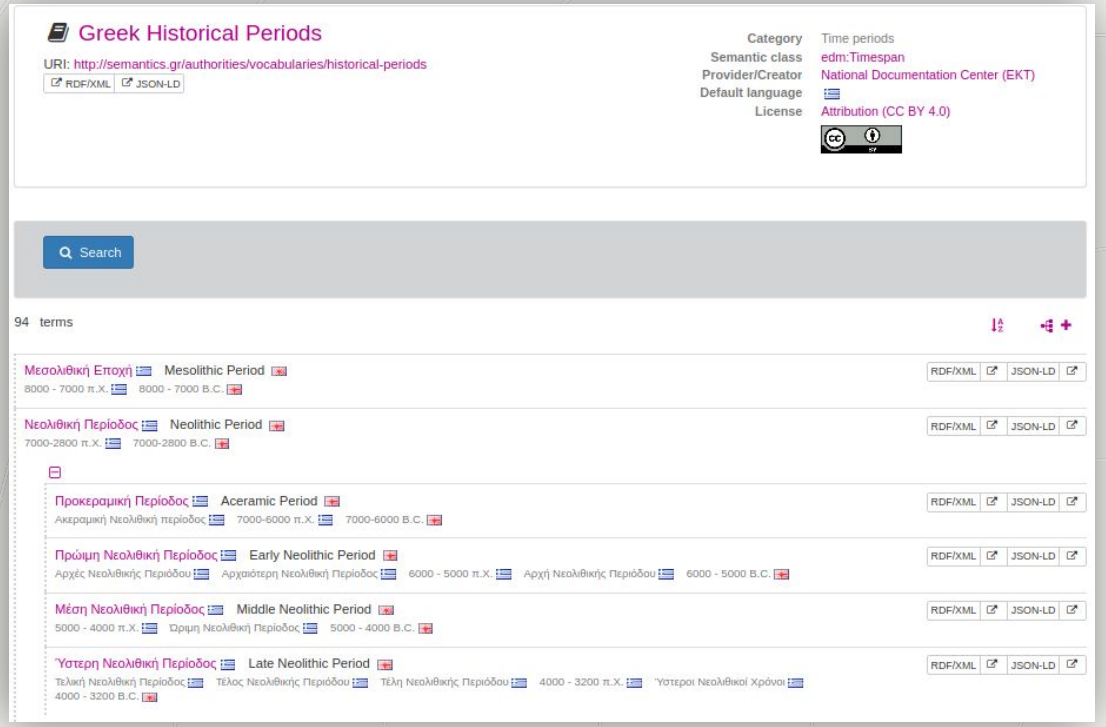
europeana

Temporal enrichment | The vocabulary of EKT Historical Periods

https://www.semantics.gr/authorities/vocabularies/historical-periods?lang=en

Greek Historical Periods

- **94** periods
- Hierarchical
- Bilingual (Greek and English)
- Schema: edm:TimeSpan
  - year ranges: edm:begin, edm:end
- **Absolute periods:** cover the entirety of hellenic territory
- **Relative periods:** have a strict local scope (e.g. minoan, cycladic and helladic periods)



Exhaustive temporal homogenization | Two approaches

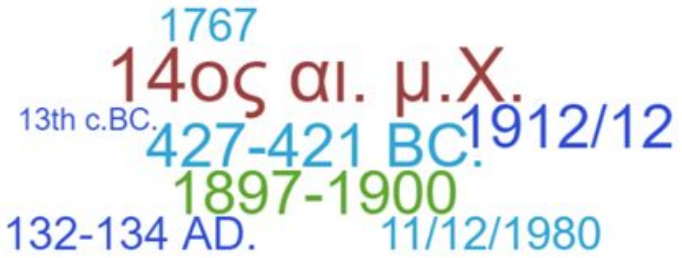
Historical period-driven enrichment

Original period label →
EKT Historical Period → EKT Chronology

according to **mapping rules** set in the **enrichment tool** of **Semantics.gr**

year ranges in voc terms

Chronology-driven enrichment



Original time value →
EKT Chronology → EKT Historical Period

according to the **chronological patterns** set in the **time normalization tool**

year ranges in voc terms

europeana

Historical period-driven enrichment

Metadata records

dcterms:temporal	EKT historical period Step 1: Mapping rules - primary field	EKT chronology Step 2: extract year span
Post-Byzantine Period	Ottoman Period	1453/1821
Second War	World War II	1940/1944
Middle - Late Hellenistic Years	Middle Hellenistic Period - Late Hellenistic Period	-220/-31

Mapping rules on **dcterms:temporal** for **collection** in **Semantics.gr** **(enrichment tool)**

Vocabulary of EKT historical periods

europeana

Exhaustive temporal homogenization | Two approaches

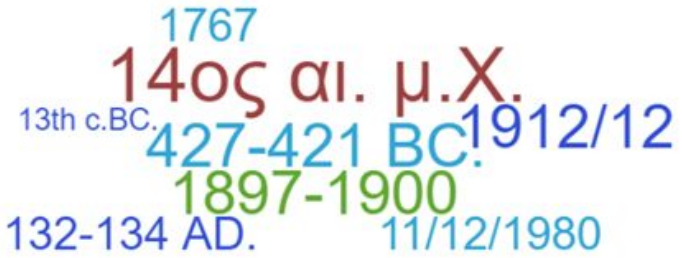
Historical period-driven enrichment

Original period label →
EKT Historical Period → EKT Chronology

according to **mapping rules** set in the **enrichment tool** of **Semantics.gr**

year ranges in voc terms

Chronology-driven enrichment



Original time value →
EKT Chronology → EKT Historical Period

according to the **chronological patterns** set in the **time normalization tool**

year ranges in voc terms

europeana

A tool for time normalization

- based on **regular expression** processing
- 4 classes of chronological patterns, each with different extraction algorithm
  - century range | century | year/date range | year/date
- Predefined and custom parametric placeholders are used inside patterns
  - **placeholder for "BC"**: "BC", "BCE", "π.Χ."
  - **placeholder for "first half of"**: "first half of", "1st half of" , "first A", "πρώτο μισό του"
  - **placeholder for "century"** : "century", "c.", "cent.", "αιώνας", "αι."
- completely parametric and extensible
- unlike mapping rules which are created per collection, once you create a pool of chronological patterns, you almost done for all collections
- we created 30 patterns to cover the diversity of **SeachCulture.gr** collections

europeana

A tool for time normalization | our pool of 30 patterns

Chronological Pattern Class	**# of patterns**	**Examples**
century range	5	2nd half of 5th c. BC until 4th c. BC→ **-450/-301** 5th c. bc (2nd half of) - 4th c. bc → **-450/-301**
century	7	early 18th century → **1700/1730** first half of 5th c. BC → **-500/-451** τέλος 4ου αι.π.χ. → **-330/-301**
date/year range	8	1342/48 → **1342/1348** 1342 - 1654 → **1342/1654** 579 - 570 π.Χ. → **-579/-570**
date/year	11	526 BC → **-526** 11/01/1980 → **1980** May the 1st 1870 →**1870**

europeana

A tool for time normalization | a pattern example

Pattern name: early Xth century

Pattern class: **century**

Pattern: \[?(**#century_identifier**)(.*\s)?(\d{1,4})\s?**#s0**?(\s)?(**#bc_ad**(\s*)?)?(\s**#s1**\.?)

Extraction pointers: 4

Predefined placeholders:

- #century_identifier
  early: “early”, “first quarter of”, “beginning of”, “αρχές” ...
  late: “late”, “end of”, “end of”, “τελος”, ...
- #bc_ad
  BC: “bce”, “bc”, “b.c.e.”, “b.c.”, “b.c”, “π.χ.”, “π.χ”, “πχ”
  AD: “μ.χ.”, “μ.χ”, “μχ”, “ad”, “a.d.”, “a.d”, “ce”, “c.e.”, “c.e”

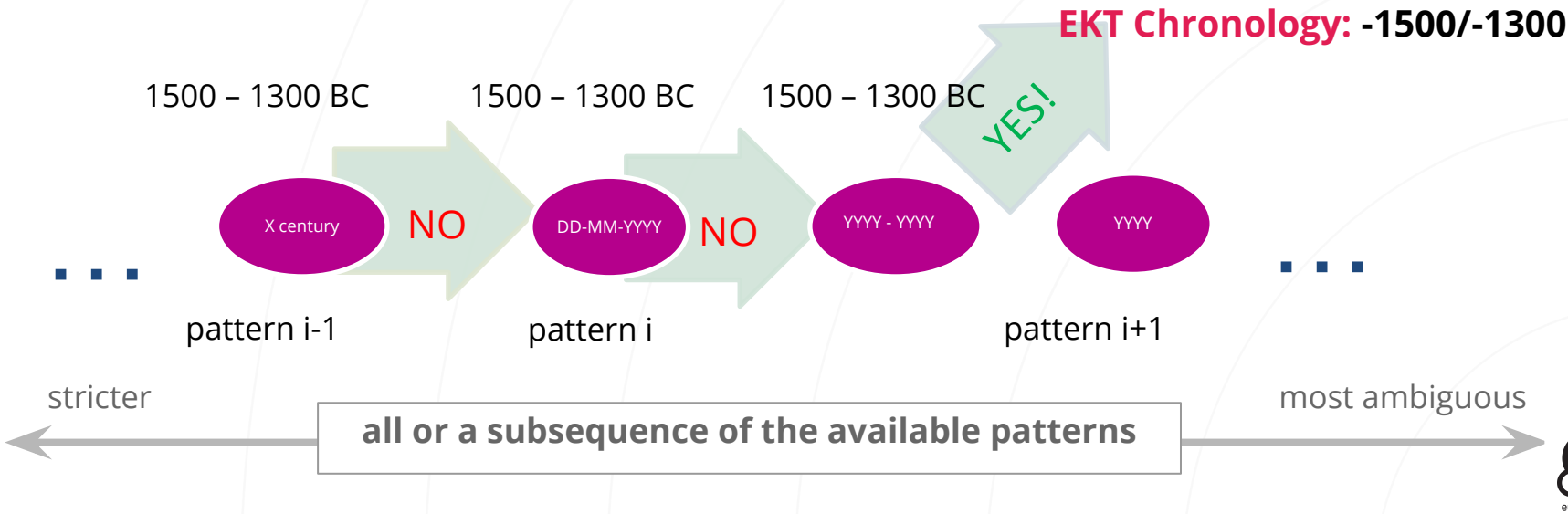
Custom placeholders:

- #s0: ος|ος|ου|st|nd|rd|th
- #s1: αιώνας|αι|α|century|cent|c

Examples:
early 6th c. BCE → -600/-571
early 6th c. AD → 500/530
first quarter of the **2nd c. AD** → 100/130
end of the **12th cent.** → 1171/1200
αρχές 5ου αι. π.Χ. → -500/-471

A tool for time normalization | how does it work?

- patterns have a natural order: from the stricter to the most ambiguous
- when a chronological value is to be normalized, it passes through all the chronological patterns sequentially, until the first match is found.



Chronology-driven enrichment

Metadata records

dc:date	EKT chronology Step 1: normalize chronologies	EKT historical period Step 2: enrich with corresponding ***absolute*** period
Late 5th century	471/500	Early Byzantine Period
7th c. B.C - mid 6th c. BC	-700/-551	Early Archaic *to* Middle Archaic Period
03/11/1980	1980	Regime change

Subsequence of chronological patterns
(time normalization tool)

Vocabulary of
EKT historical periods
Index from year range to terms

europeana

Lessons learned (1)

Choose the target vocabularies

- Very big vocabularies do not facilitate browsing
- Small vocabularies do not facilitate automated mappings based on string-similarity matching
- Consider the option of **creating your own vocabulary** especially when one is not available in your language or they don't cover country-specific subjects (e.g. historical events). Link it!

Enrichment based on mappings

- Mappings on **distinct** metadata values not per record!
- Fully automatically generated mappings are not as to the point as **automatic mapping suggestions combined with curation**
- Facilitate the curator's work by giving her tools that facilitate better understanding of the content during the enrichment process
- **Self-improving mapping suggestion mechanism** : manual mappings are precious if used for future automatic mapping suggestions
- Enrich **each collection separately** - take into account the general subject/material type/particularities of the collection during mappings
- Examine **other metadata fields** that may help (e.g. for type enrichment, apart from dc:type, dc:subject may help)
- Do not forget examining the **OAI SetSpecs**! (e.g. for type or subject enrichment, oai sets may give better results than dc:type or dc:subject)

europeana

Lessons learned (2)

After the enrichment

- Don’t hesitate to enrich/update your vocabularies if needed during processing a collection.
- But if you do so, you may have to go back and update existing mappings for some collections!

Time enrichment

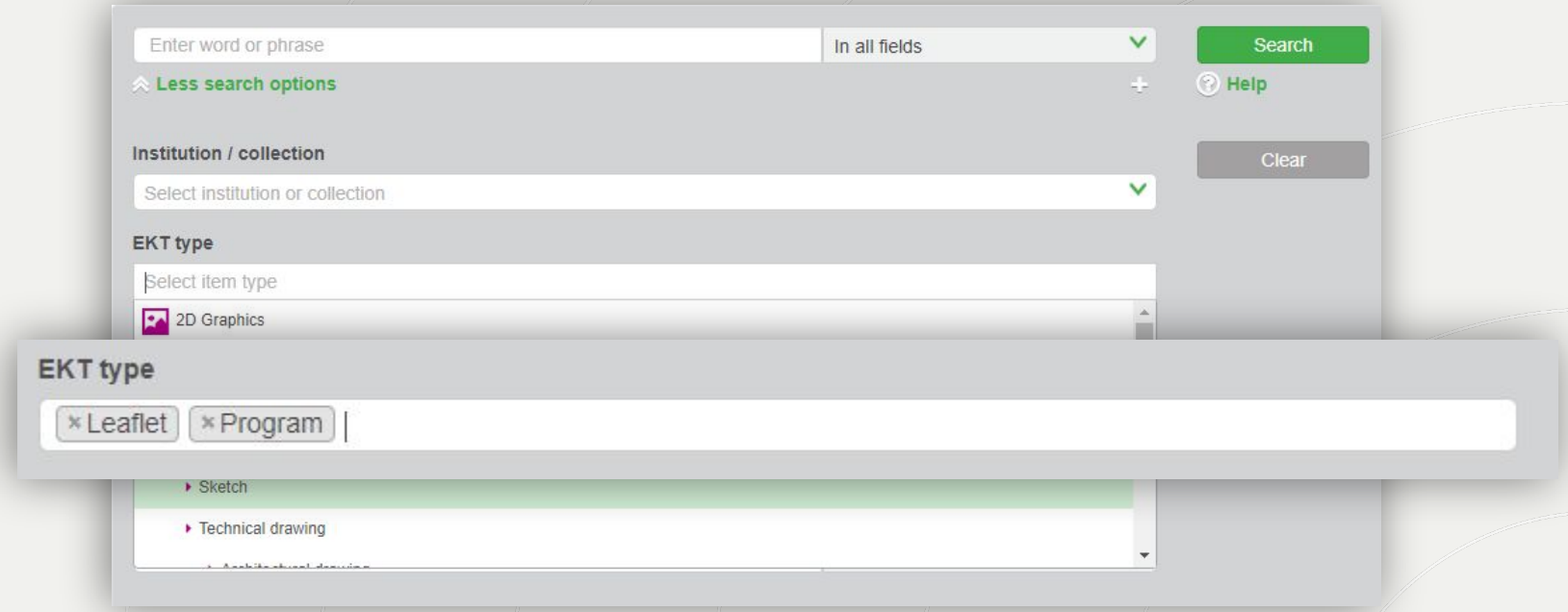
- Time enrichment is really useful and worthy when it leads to **time-based search and browsing**
- **Historical labels vs. chronological values**: Use time **uniformly**; they both should lead to **equivalent** time-based search and browsing
- Historical labels
    - **Mappings** to a Vocabulary of Historical Periods that includes the year intervals as well
    - Use the Historical Periods to **assign the respective year intervals as well!**
- Chronological values:
    - **Time normalization** is needed (years or year intervals is what you need)
    - Regular expressions that take into account **time language-specific descriptors** worked fine
    - Use normalized years/year intervals to **assign the Historical Periods as well!**

europeana

Results

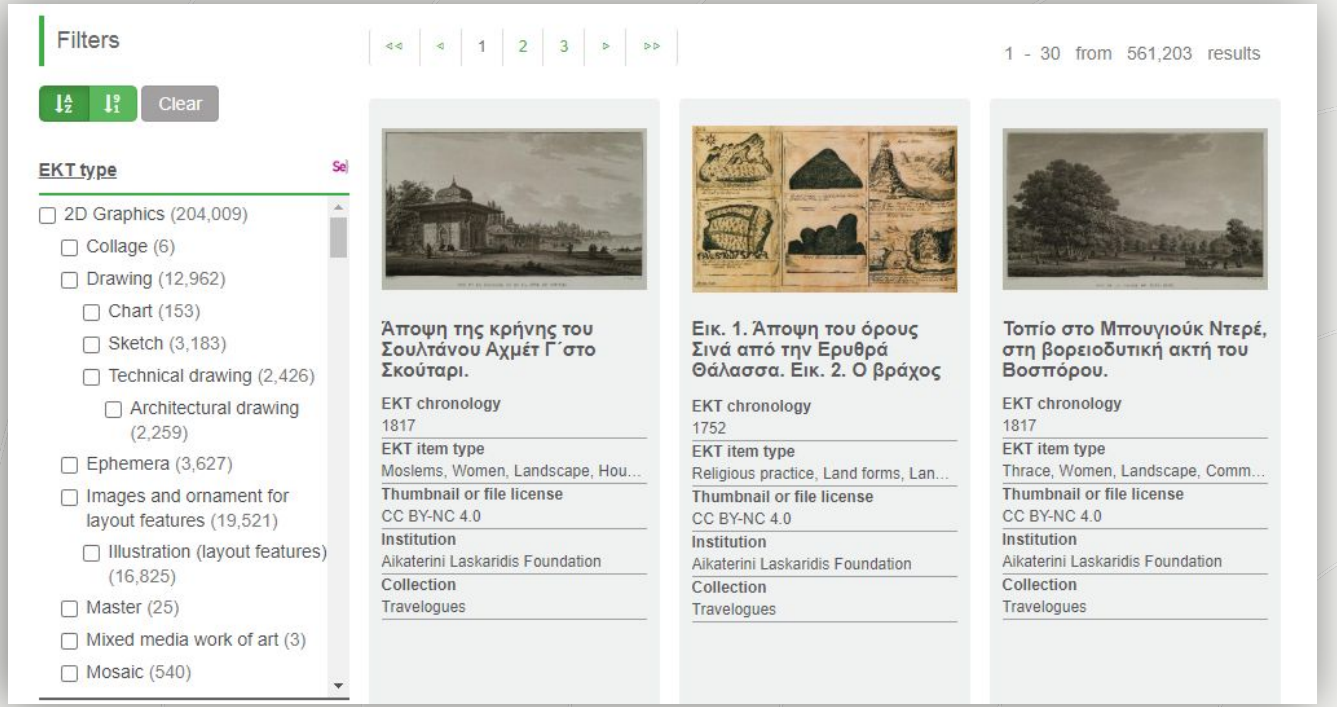
SearchCulture.gr in 2016 | 100,000 items • 20,000 in Europeana

Search by Types



SearchCulture.gr in 2016 | 100,000 items • 20,000 in Europeana

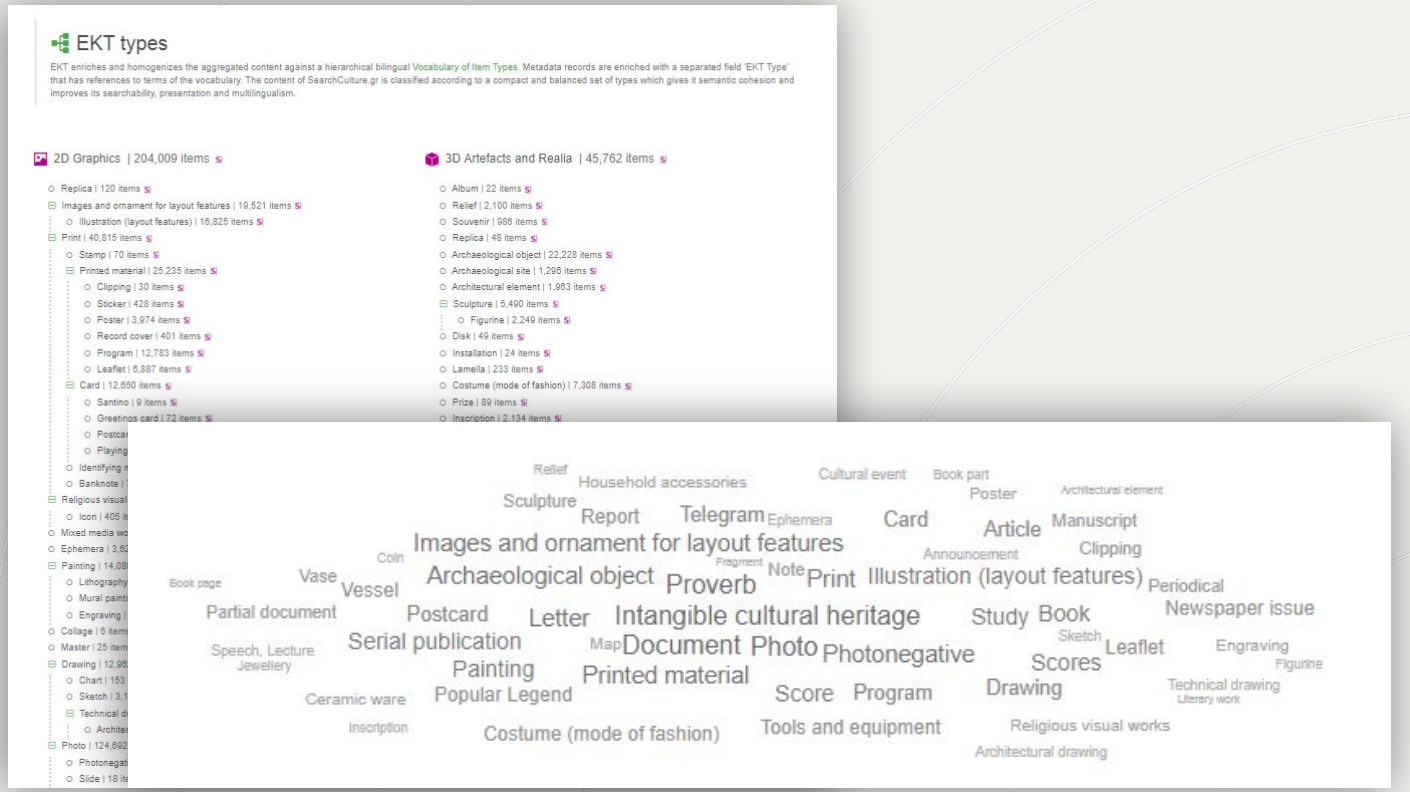
Facet by Types



europeana

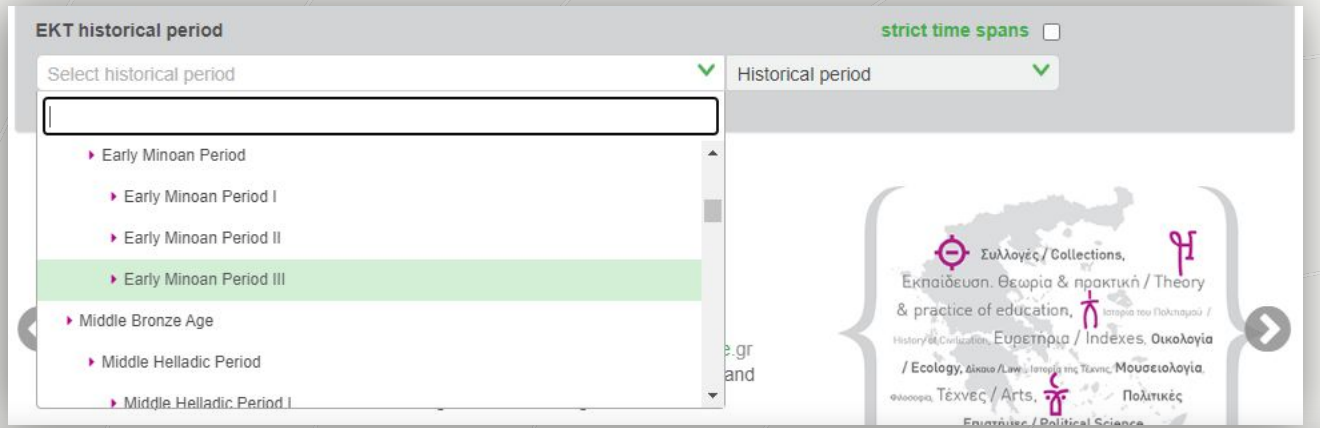
SearchCulture.gr in 2016 | 100,000 items • 20,000 in Europeana

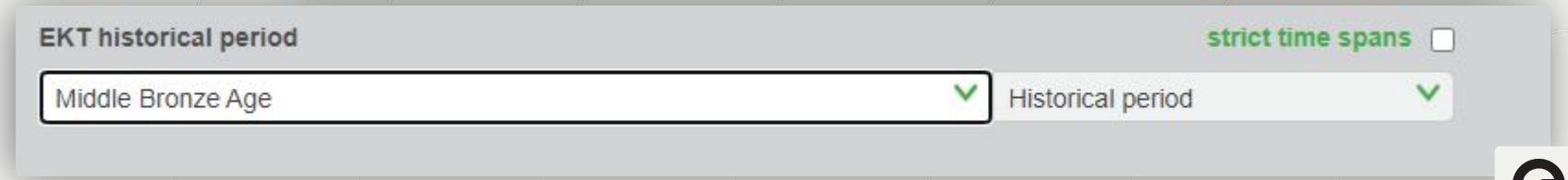
Browse by Types



SearchCulture.gr in 2017 | 160,000 items • 60,000 in Europeana

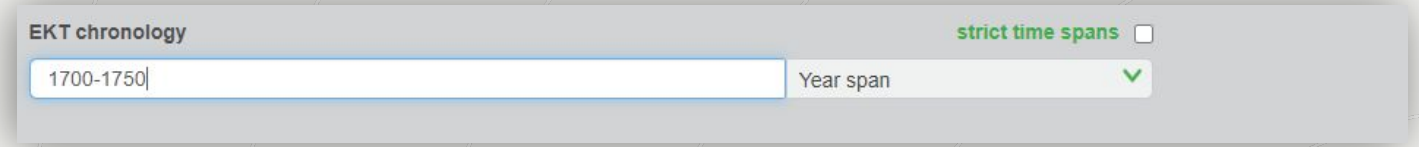
Search by Historical periods





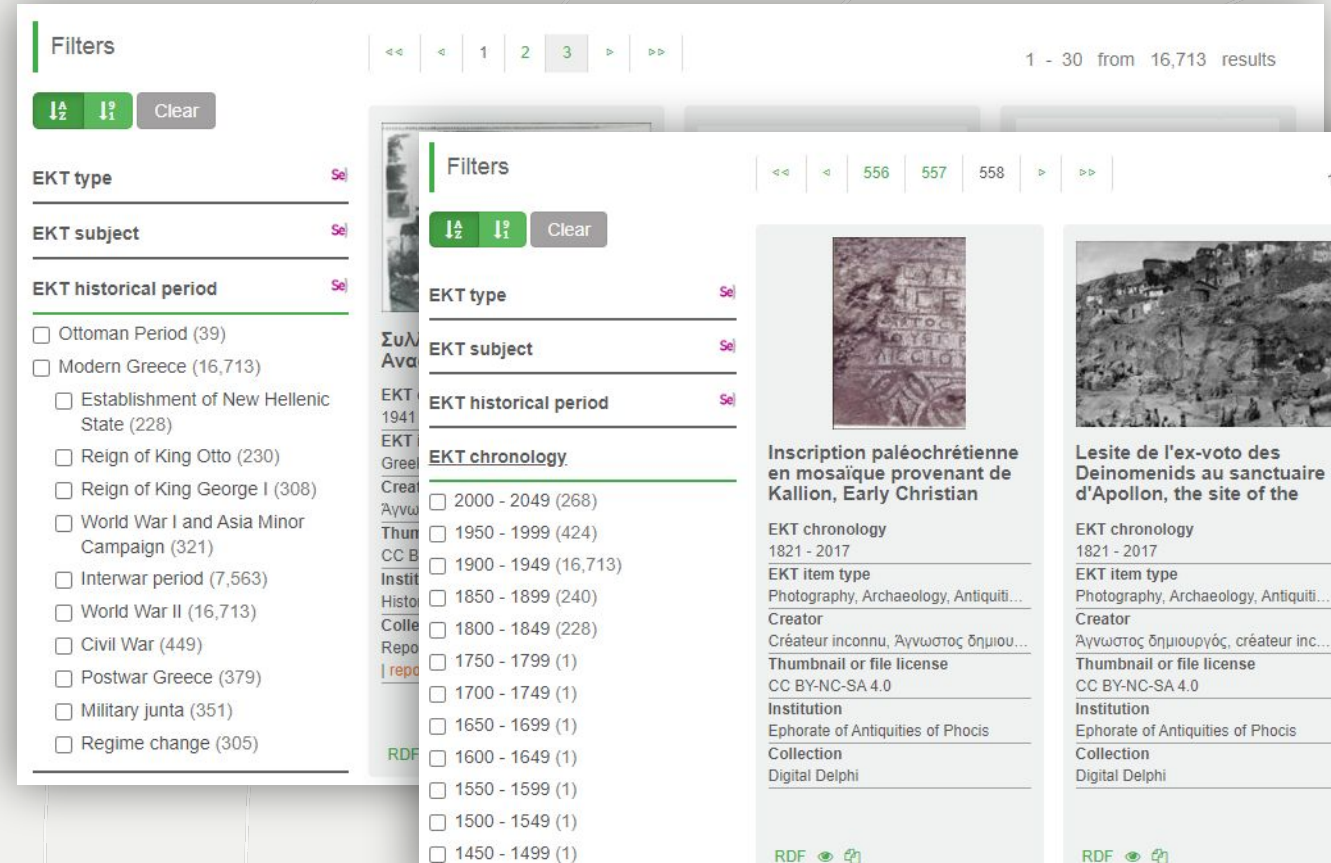
SearchCulture.gr in 2017 | 160,000 items • 60,000 in Europeana

Search by Year ranges



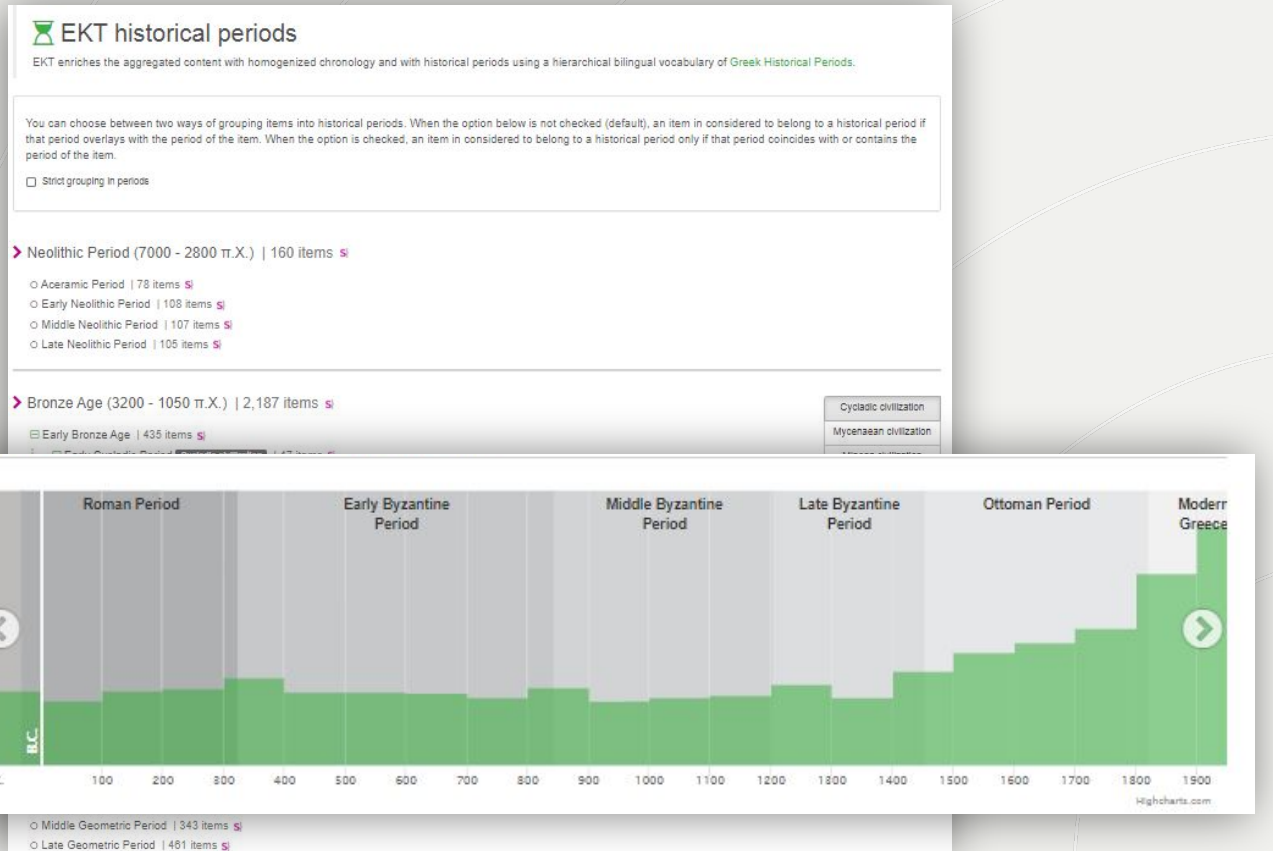
SearchCulture.gr in 2017 | 160,000 items • 60,000 in Europeana

Historical periods and time spans facets



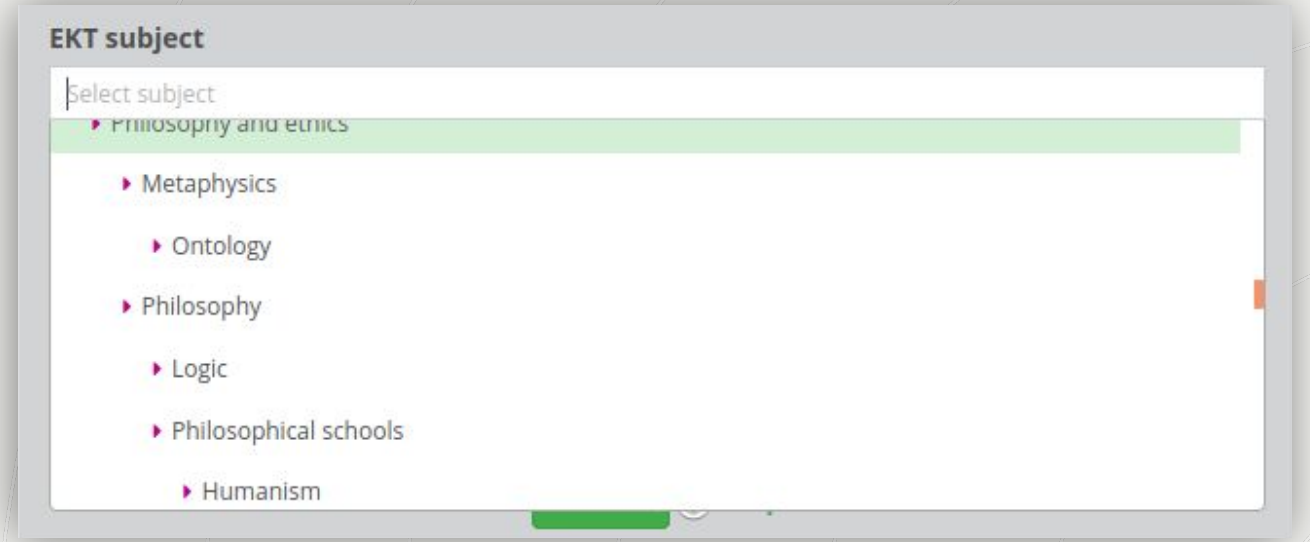
SearchCulture.gr in 2017 | 160,000 items • 60,000 in Europeana

Browse by historical periods



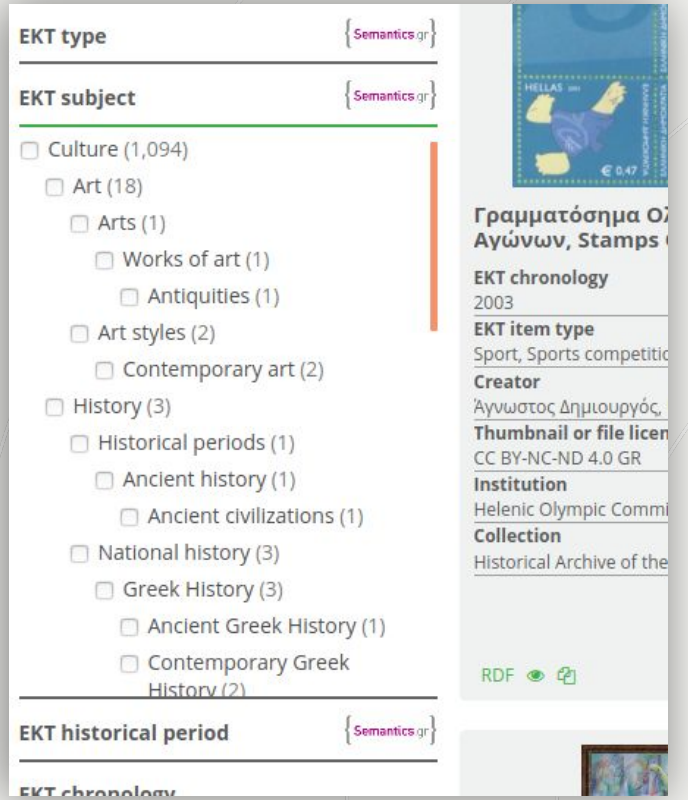
SearchCulture.gr in 2020 | 680,000 items • 320,000 in Europeana

Search by subjects



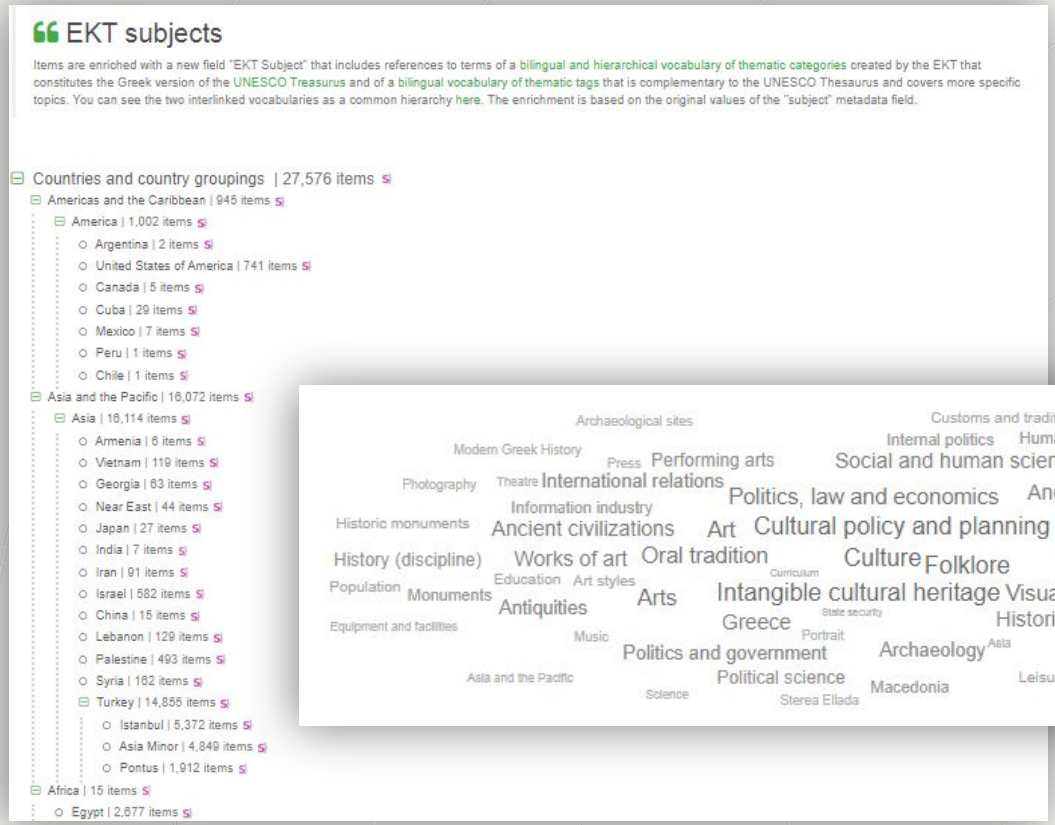
SearchCulture.gr in 2020 | 680,000 items • 320,000 in Europeana

Facet by subjects



SearchCulture.gr in 2020 | 680,000 items • 320,000 in Europeana

Browse by subjects



SearchCulture.gr in 2020 | 680,000 items • 320,000 in Europeana

Creating a query for a thematic exhibition

Example: creating a thematic exhibition titled “Archaeology”

Query: use combination of EKT fields

EKT Type: Archeological item, ...

EKT Chronology: [<1500]

EKT Historical Period: [< Middle Ages]

EKT Subject: Archeology, Ancient Greece, ...

Collection: Digital Delphi, Archaeological Museum of Chania, ...

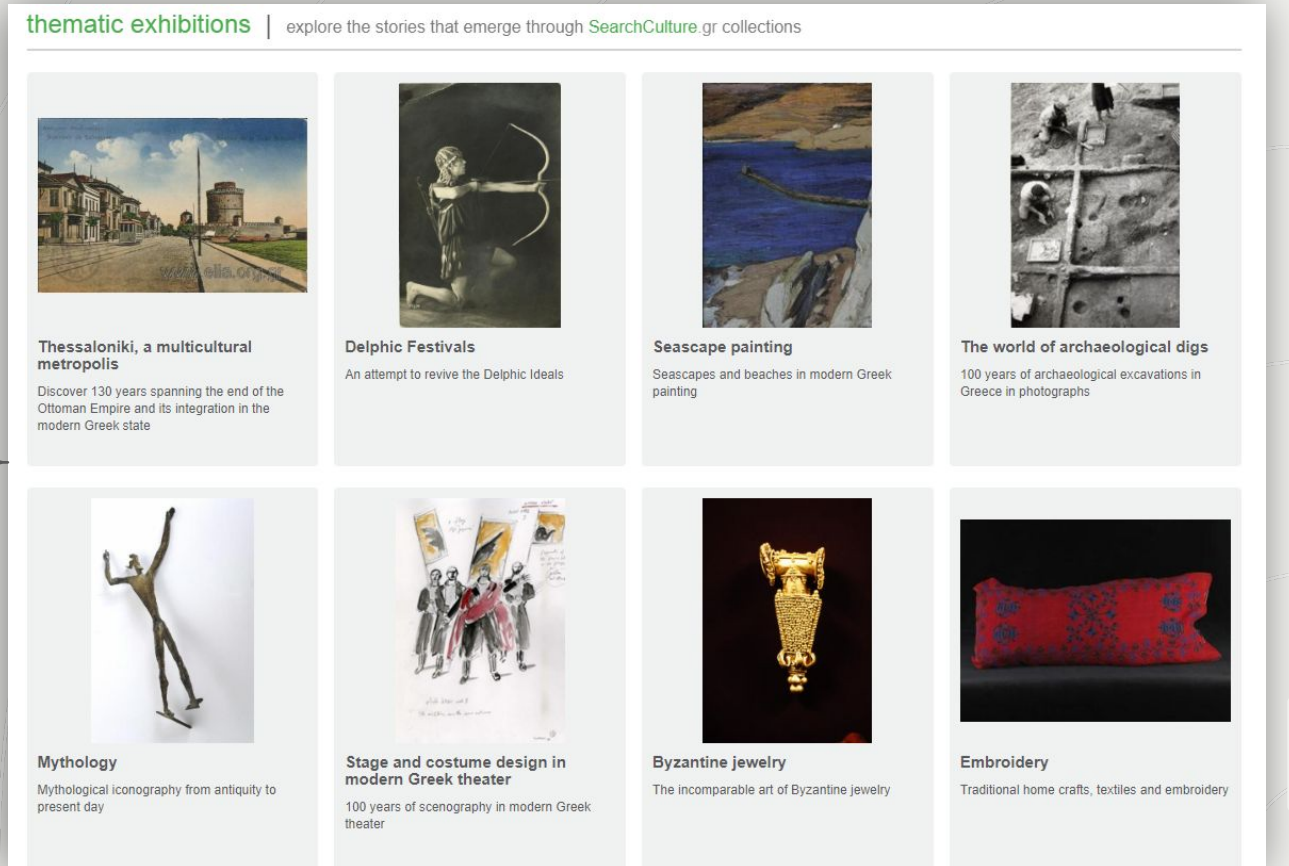
europeana

SearchCulture.gr in 2020 | 680,000 items • 320,000 in Europeana

Thematic exhibitions

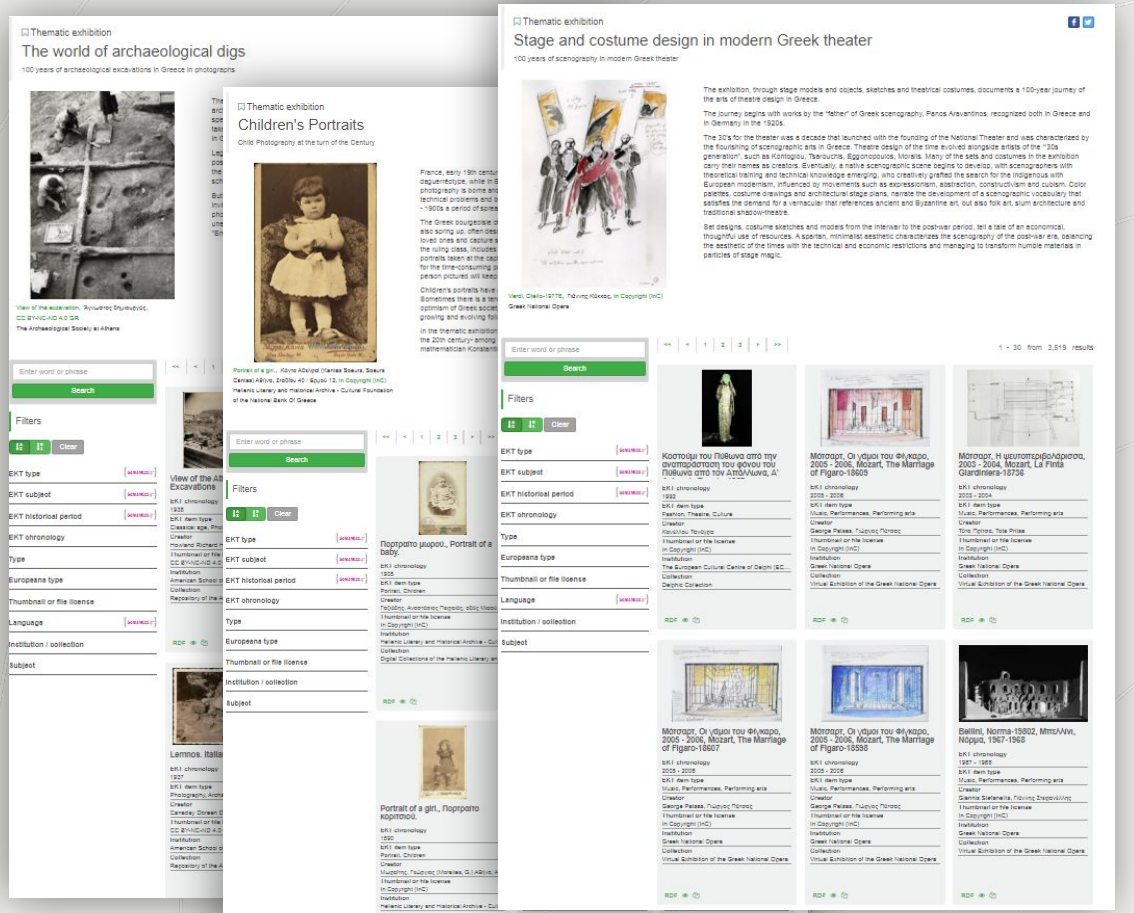
Covering a variety of EKT Types, Subjects, Historical Periods or Chronologies, from several collections.

Goal: to showcase variety and span



SearchCulture.gr in 2020 | 680,000 items • 320,000 in Europeana

Thematic exhibitions



Plans for future developments

- Enrichment of spatial metadata with Geonames (or wikidata)
- Enrichment on dc:creator, dc:contributor and dc:subject using an authority file of agents

europeana

Further reading

- Haris Georgiadis, Agathi Papanoti, Maria Paschou, Alexandra Roubani, Despoina Chardouveli, Evi Sachini: Using type and temporal semantic enrichment to boost content discoverability and multilingualism in the Greek cultural aggregator SearchCulture.gr. Int. J. Metadata, Semantics and Ontologies, Vol. 13, No. 1, pp.75–92, 2018 (preprint)
- Haris Georgiadis, Agathi Papanoti, et al: The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.gr, In Proc. MTSR 2017
- Haris Georgiadis, Agathi Papanoti, et al: Semantics.gr: A self-improving service to repositories and aggregators for massively enriching their content. In Proc. of the DHC Workshop of MTSR 2016
- Ioanna Ourania Stathopoulou, Haris Georgiadis, Vangelis Banos, et al: An Open Cultural Digital Content Infrastructure, In Proc. DL'2014

europeana

Links

- https://www.searchculture.gr/?language=en
- https://www.semantics.gr/authorities/?language=en
- Vocabularies in semantics.gr:
  - Cultural heritage item types
    https://www.semantics.gr/authorities/vocabularies/ekt-item-types?language=en
  - UNESCO Thesaurus (EKT version)
    http://semantics.gr/authorities/vocabularies/ekt-unesco?lang=en
  - Thematic tags
    http://semantics.gr/authorities/vocabularies/thematic_tags
  - Greek Historical Periods
    https://www.semantics.gr/authorities/vocabularies/historical-periods?lang=en

europeana

Thank you!!!

Dr Haris Georgiadis, Computer Scientist, Head of e-Services Unit
E-infrastructure and Information Systems Department
NATIONAL DOCUMENTATION CENTRE (EKT)

Agathi Papanoti, Archaeologist, Data Ingestion Specialist
Documentation Unit, Scientific and Cultural Digital Library Department
NATIONAL DOCUMENTATION CENTRE (EKT)

searchculture@ekt.gr

europeana

CC BY SA

europeana.eu

@EuropeanaEU

Co-financed by the Connecting Europe Facility of the European Union

Europa [Material cartográfico] : Nach den vorzüglichsten Hülfsnitteln, Götze, Johann August Ferdinand, 1773-1819 Biblioteca Digital de Madrid Spain, Public domain