The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.gr

Haris Georgiadis PhD | Computer Scientist
National Documentation Centre

01.12.2017, Tallin

MTSR 2017

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΤΕΚΜΗΡΙΩΣΗΣ
NATIONAL DOCUMENTATION CENTRE

National Documentation Center

www.ekt.gr

{ SearchCulture.gr } { Semantics.gr }

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΤΕΚΜΗΡΙΩΣΗΣ
NATIONAL DOCUMENTATION CENTRE

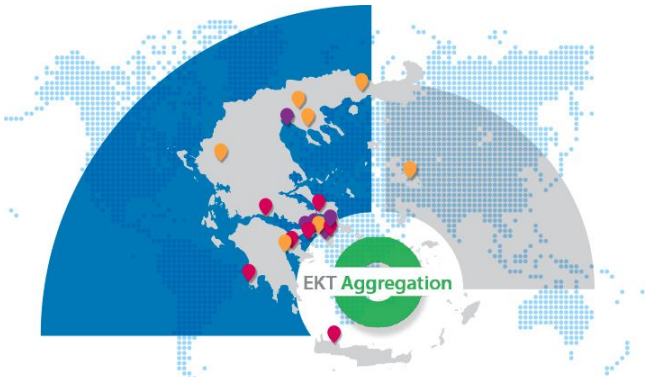
ΕΘΝΙΚΟ ΙΔΡΥΜΑ ΕΡΕΥΝΩΝ 1958

eContentEKT

digital content and services

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

- repositories EKT
- openABEKT | proprietary Integrated Library System
- ePublishing



- Large scale **aggregators** of Greek scientific and cultural digital content

SearchCulture.gr | national aggregator for cultural content

OpenArchives.gr | scientific content

SearchCulture.gr | Greek Cultural Heritage Aggregator

Aggregates digital content from repositories

66 collections from 52 institutions

museums • archives • ephorates of antiquities • municipalities • cultural foundations

430.000+ digital assets | 600.000+ still to come

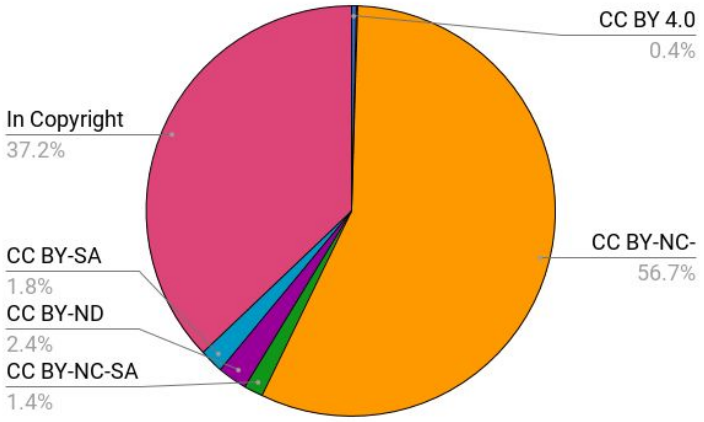
archaeological items • historical documents • folklore items • works of art • cartographic material • books • oral history

100.000+ digital assets published in

Europeana

700.000+ will be published soon

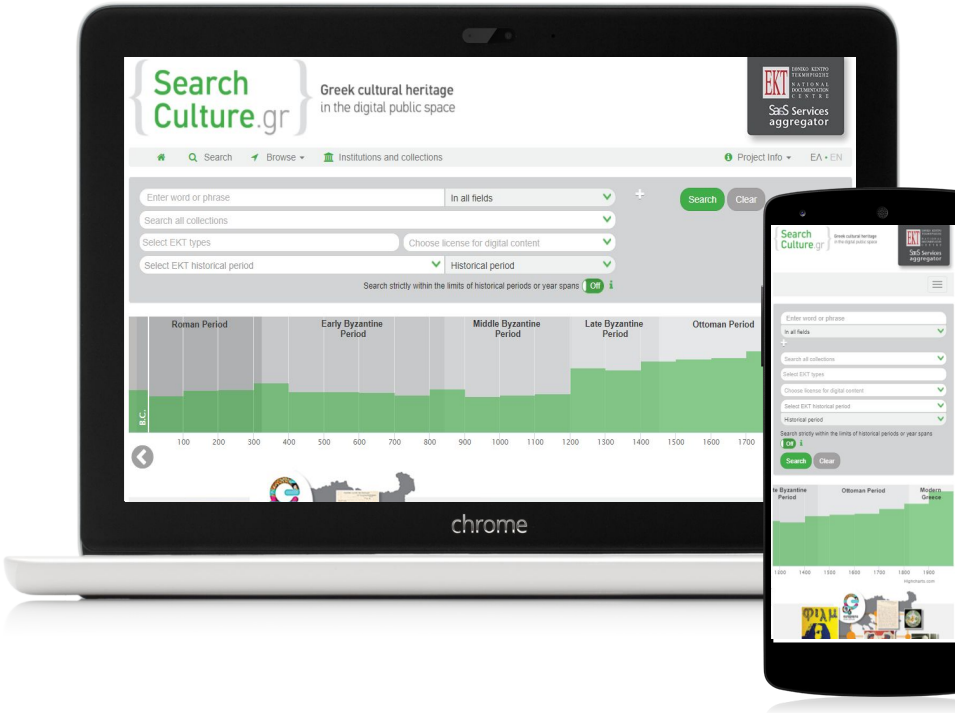
creative commons for more than 60% of the digital files



SearchCulture.gr | Greek Cultural Heritage Aggregator

The public portal www.searchculture.gr

- modern and effective search engine
- time-based search and filtering
- advanced search faceting
- advanced hierarchical browsing
- dissemination as Linked Data
- bilingual environment, search & browsing



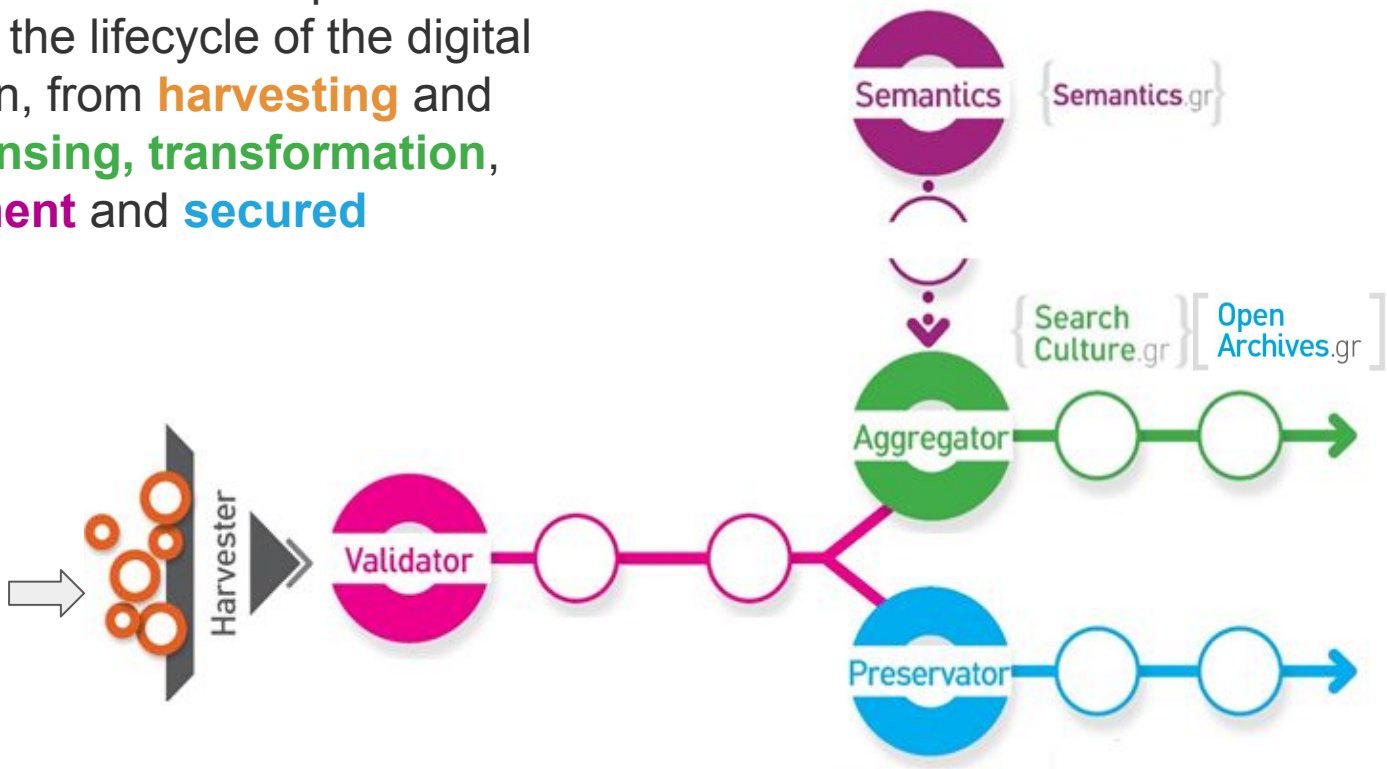
- rights licenses for digital assets
- internal data model based on EDM

feasible thanks to or radically improved by **exhaustive semantic enrichment** on types, chronologies and historical periods

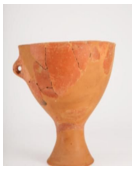
EKT aggregation infrastructure

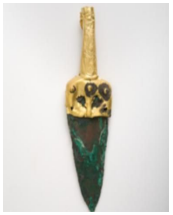
EKT has developed an aggregation infrastructure that consists of five platforms and systems that cover the lifecycle of the digital content aggregation, from **harvesting** and **validation**, to **cleansing, transformation**, **semantic enrichment** and **secured preservation**.

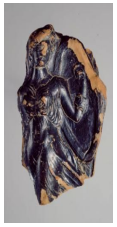
museums
archives
ephorates of antiquities
municipalities
cultural foundations
universities
scientific institutions
research centres
libraries
public organizations

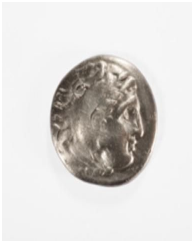


Item types and temporal information | key metadata


vase

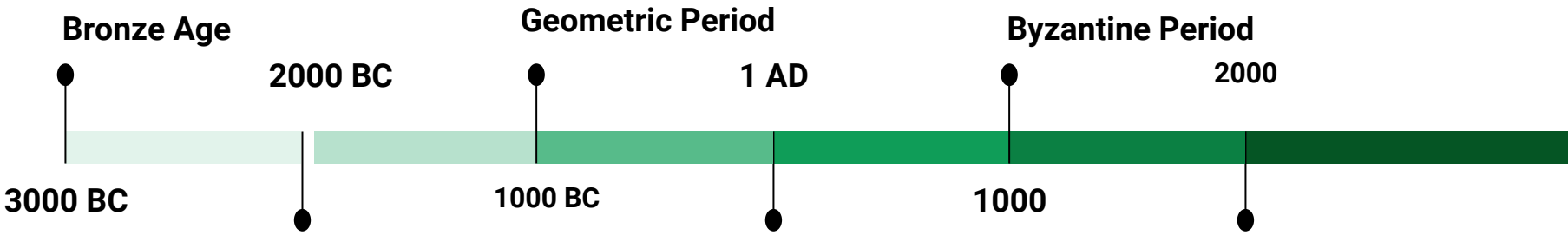

weapon

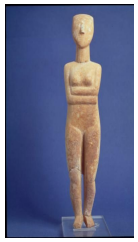

ostracon


coin


jewellery


painting




figurine


disk


sculpture


inscription


icon

Search Culture.gr

Item types and temporal information | key metadata

When it comes to cultural and historical content, keyword-based searching is far from sufficient. Users expect to be able to:

- search the content with time criteria; year ranges or historical periods
- explore the content by browsing through historical periods and a timeline
- explore the content by browsing through types
- filter search results on types, year ranges and historical periods
- submit combined queries like
  - **icons** from the **late byzantine period**
  - **manuscripts** dated from **1850 to 1910**
  - **sculptures** dated strictly within the **middle classical period** of Greece

Huge challenge for large scale cultural aggregators due to the **heterogeneity of metadata**

Search Culture.gr

The heterogeneity of types

Heterogeneity in dc:type metadata field

- **Representation-related**
  ex. different languages, synonyms, plural / singular numbers, different case styles
- **Documentation-related**
  ex. use of very broad
  or very narrow (specialized) terms

The heterogeneity of temporal values

Heterogeneity in dc:date, dcterms:temporal, dcterms:issued metadata fields

- **Use of period label** values: as problematic as types
- **Use of chronological values** range from strict date format standards to descriptions that approach natural language

Our semantic enrichment and homogenization scheme

- **Semantics.gr**: a platform developed by EKT where institutions can create, establish and publish vocabularies, taxonomies, thesauri and authority files
- **Enrichment tool** of **Semantics.gr**: a tool for setting *enrichment mapping rules (**EMRs**)* from metadata values to vocabulary terms

- **Time normalization tool** of the **aggregator platform**: a tool for setting parametric normalization patterns of time values

Semantics.gr | Platform for creating vocabularies & thesauri

The portal www.semantics.gr

A pilot platform where EKT and other institutions create, establish & link their own semantic vocabularies and thesauri

concepts | time periods | places | agents

Parametric schema modeling

- create parametric **owl properties**
- group owl properties in **owl classes**
- successfully modelled:
  skos:Concept | edm:TimeSpan | edm:Place | edm:Agent

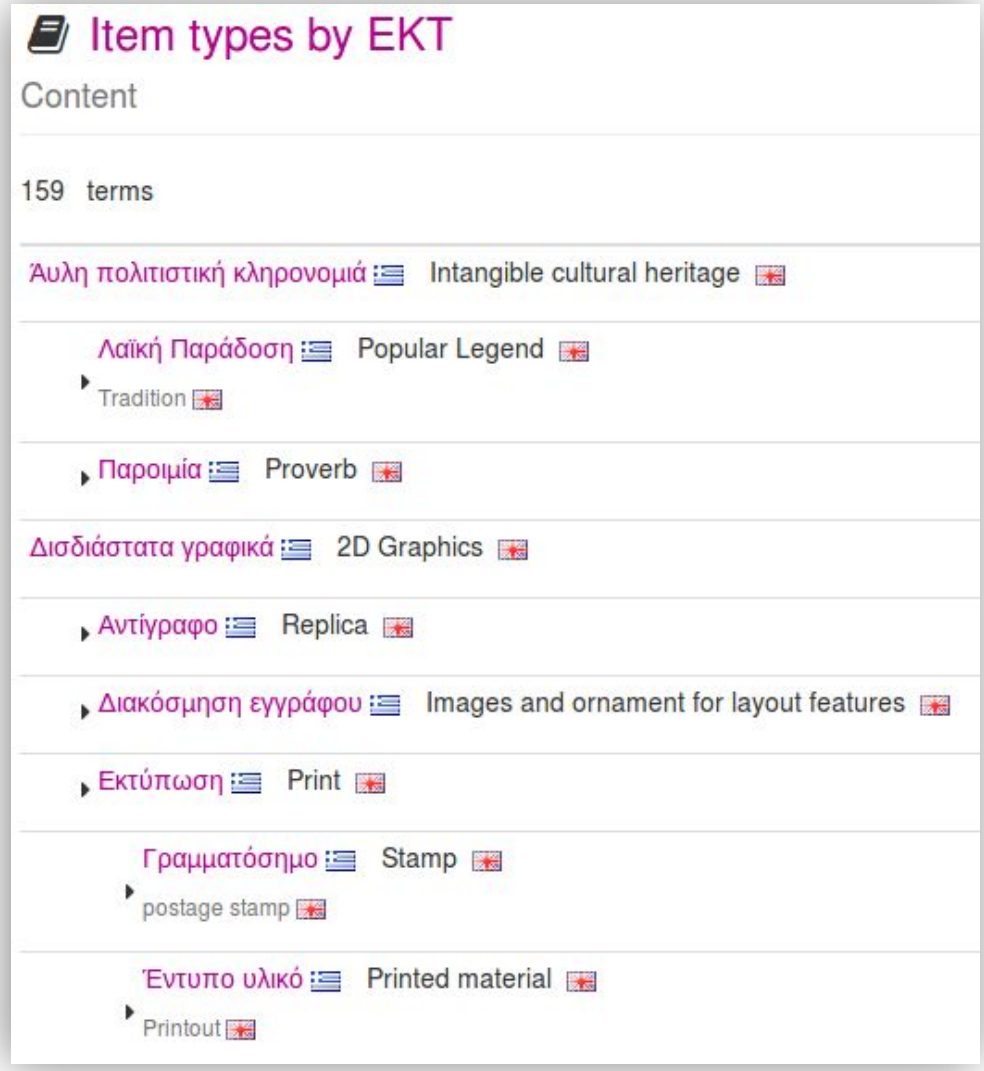
Enrichment tool of Semantics.gr

- *Enrichment Mapping Rules (**EMRs**)* from metadata field values to vocabulary terms per collection/repository
- Metadata field : **primary field** (ex. dc:type)
- **Automatic mapping suggestion** enhanced by a **self-improving mechanism**
- Use of a **secondary metadata field** if necessary
  - secondary metadata field **values as filters** (ex. dc:subject)
  - or **key-terms** inside **descriptive fields** as filters (ex. dc:title)
  - curated **complex expressions on filters** in order to create finer and more precise rules and avoid false positives
- A mapping rule set assigned to a repository/collection is available through a REST API in json format

Type enrichment | The vocabulary of EKT Types

Semantics.gr

- 159 terms
- Hierarchical
- Bilingual (Greek and English)
- Links to Getty AAT
- Schema: skos:Concept



Type enrichment | Steps per collection

Semantics.gr

1. Insert the collection in **acceptance portal**
2. Inspect the documentation quality of types
   - is **dc:type** sufficient to set EMRs on? (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 **EMRs** in the **Enrichment tool** of Semantics.gr
4. Re-index the collection in **acceptance portal and check**
5. Insert (or re-index) and publish the collection in SearchCulture.gr

Type enrichment | Our goal

Semantics.gr

Each record is enriched with a new field:

EKT Type: from the vocabulary of types

Type enrichment | EMRs on dc:type values

Semantics.gr

EKT Types Vocabulary

➔ http://scs.gr/sculpture
skos:prefLabel "Sculpture"@en | "Γλυπτό"@el
➔ http://scs.gr/figurine keywords: "statuette"
skos:prefLabel "Figurine"@en | "Ειδώλιο"@el
➔ http://scs.gr/Jewellery
skos:prefLabel "Jewellery"@en | "Κόσμημα"@el
➔ http://scs.gr/vessel
skos:prefLabel "Vessel"@en | "Σκεύος"@el
➔ http://scs.gr/vase
skos:prefLabel "Vase"@en | "Αγγείο"@el

dc:type value	Entry from vocabulary	
sculpture art *(120 items)*	http://scs.gr/sculpture	auto
greek vases *(230 items)*	http://scs.gr/vase	auto
statuette *(15 items)*	http://scs.gr/figurine	manual

Type enrichment | EMRs on dc:type & dc:subject values

Semantics.gr

Vocabulary
EKT Types

➔ http://scs.gr/sculpture
skos:prefLabel "Sculpture"@en | "Γλυπτό"@el keywords: "statue"
➔ http://scs.gr/figurine
skos:prefLabel "Figurine"@en | "Ειδώλιο"@el keywords: "statuette"
➔ http://scs.gr/Jewellery
skos:prefLabel "Jewellery"@en | "Κόσμημα"@el keywords: "earring"
➔ http://scs.gr/vessel
skos:prefLabel "Vessel"@en | "Σκεύος"@el
➔ http://scs.gr/vase
skos:prefLabel "Vase"@en | "Αγγείο"@el keywords: "amphora", "oenochoe"

dc:type values	Filters: dc:subject	Entry from vocabulary *V1*	
ceramic objects *(101 items)*	amphora (↗) , vase (↗), statuette (↗), figurine (↗), …	http://scs.gr/vase	auto
		if filter in ["vase", "amphora"]	auto
		http://scs.gr/figurine	auto
		if filter in ["statuette", "figurine"]	auto
exhibits *(55 items)*	earing (↗), amphora (↗), …	http://scs.gr/Jewellery	auto
		if filter in ["earing"]	auto
		http://scs.gr/vase	auto
		if filter in ["amphora"]	auto
		but **NOT** in ["earing"]	manual

Type enrichment | EMRs on dc:type values & dc:title

Semantics.gr

Vocabulary
EKT Types

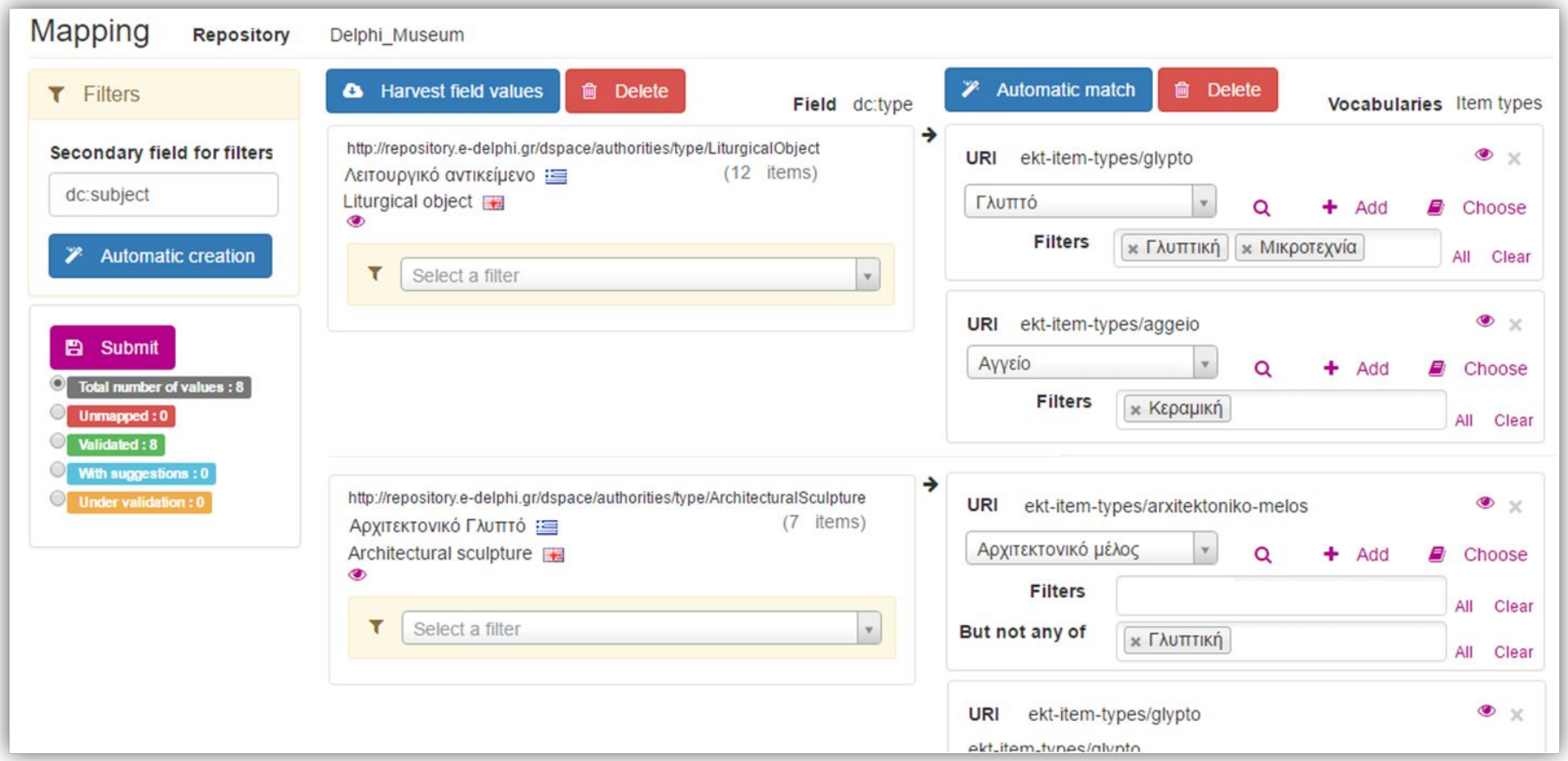
➔ http://scs.gr/sculpture
skos:prefLabel "Sculpture"@en | "Γλυπτό"@el keywords: "statue"
➔ http://scs.gr/figurine
skos:prefLabel "Figurine"@en | "Ειδώλιο"@el keywords: "statuette"
➔ http://scs.gr/Jewellery
skos:prefLabel "Jewellery"@en | "Κόσμημα"@el keywords: "earring"
➔ http://scs.gr/vessel
skos:prefLabel "Vessel"@en | "Σκεύος"@el
➔ http://scs.gr/vase
skos:prefLabel "Vase"@en | "Αγγείο"@el keywords: "amphora", "oenochoe"

Example of dc:title value: "An amphora from the Mycenaean period"

dc:type values	**Filters** (terms found in **dc:title** values)	**Entry from vocabulary V1**	
3D objects *(240)*	amphora (↗), vase (↗), earring (↗), jewellery (↗)	http://scs.gr/vase	auto
		if filter in ["vase", "amphora"]	auto
		but **NOT** in ["statue"]	manual
		http://scs.gr/Jewellery	auto
		if filter in ["earring", "jewellery"]	auto
art items *(85)*	sculpture (↗), statue (↗), figurine (↗)	http://scs.gr/sculpture	auto
		if filter in ["sculpture", "statue"]	auto
		http://scs.gr/figurine	auto
		if filter in ["figurine"]	auto

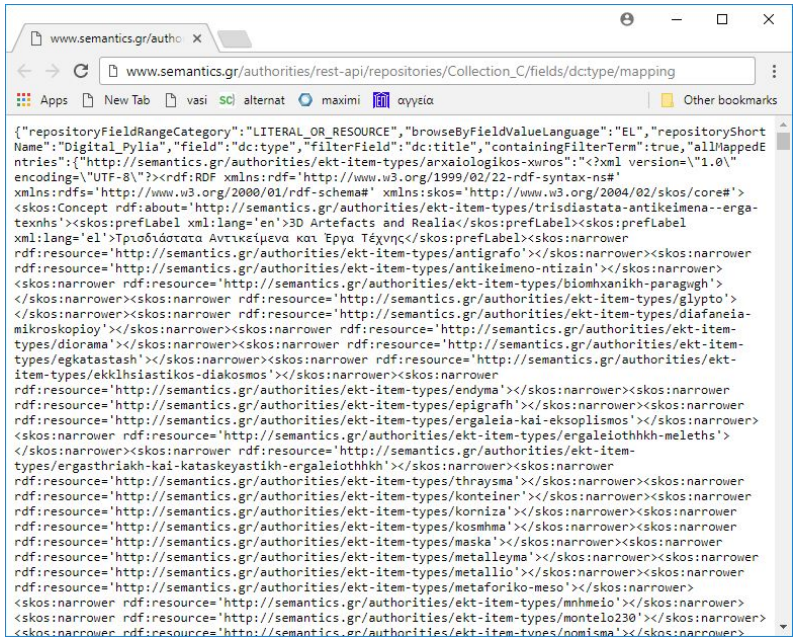
The GUI of the enrichment tool

Semantics.gr

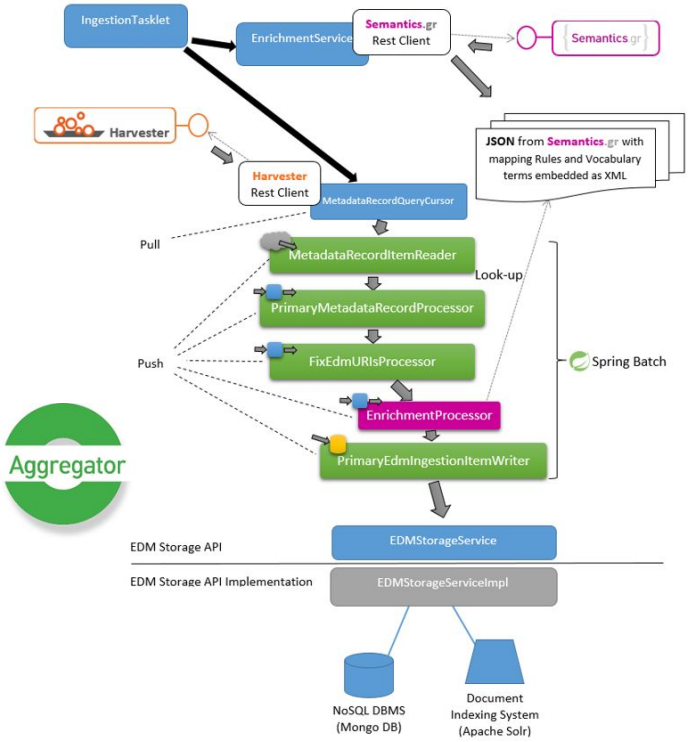


The actual enrichments are done by the Aggregator

ERMs are offered by **Semantics.gr** in JSON through a REST API



Aggregator uses **ERMs** as guidelines for the enrichment step of the ingestion data flow

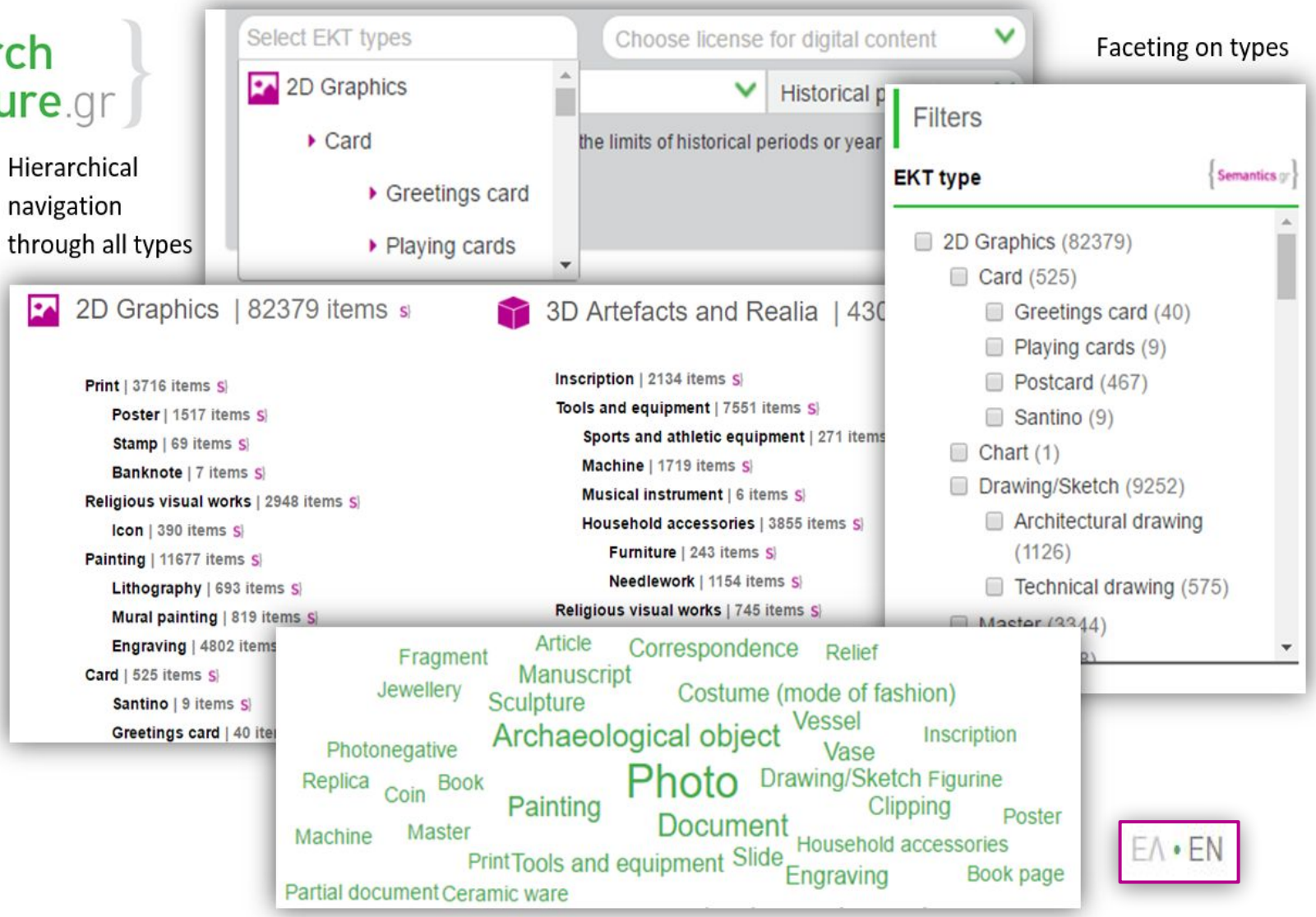


New searching, filtering & browsing features | based on type

Search Culture.gr

Search by type

Hierarchical navigation through all types



Faceting on types

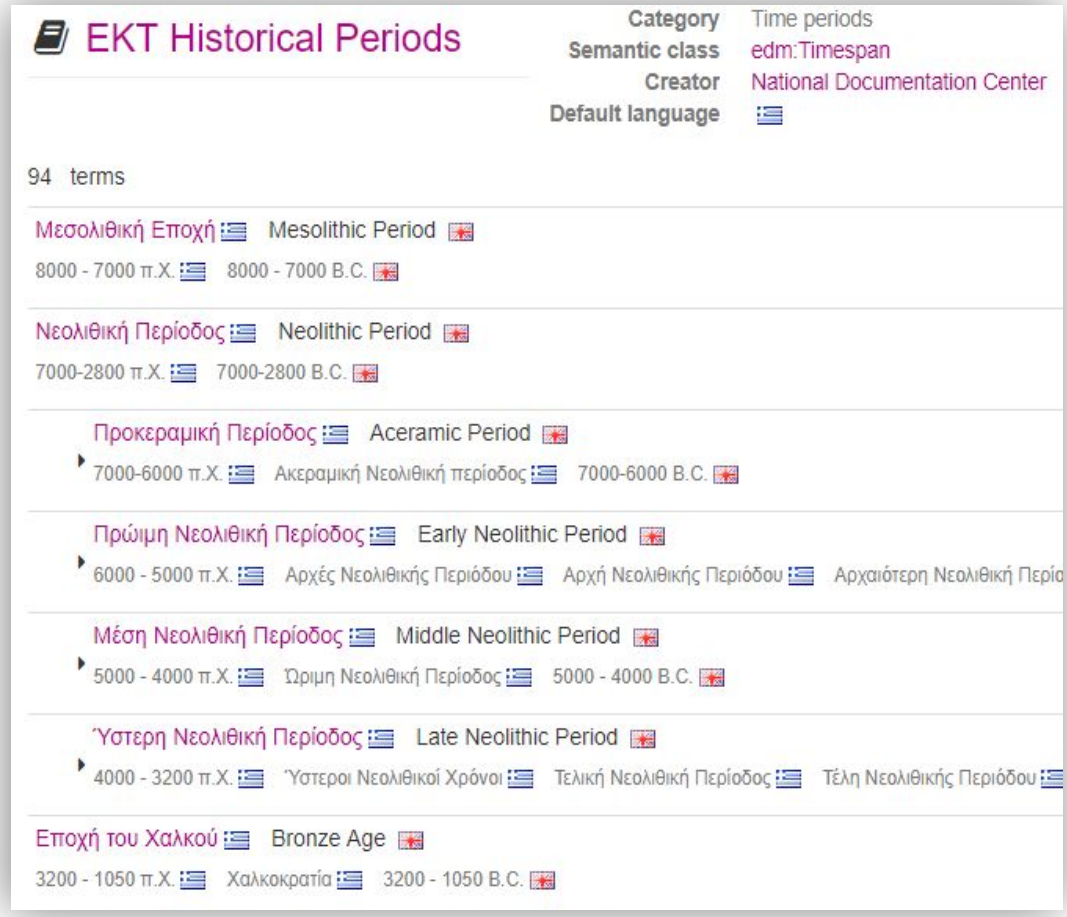
ΕΛ • EN

Tag cloud of popular types

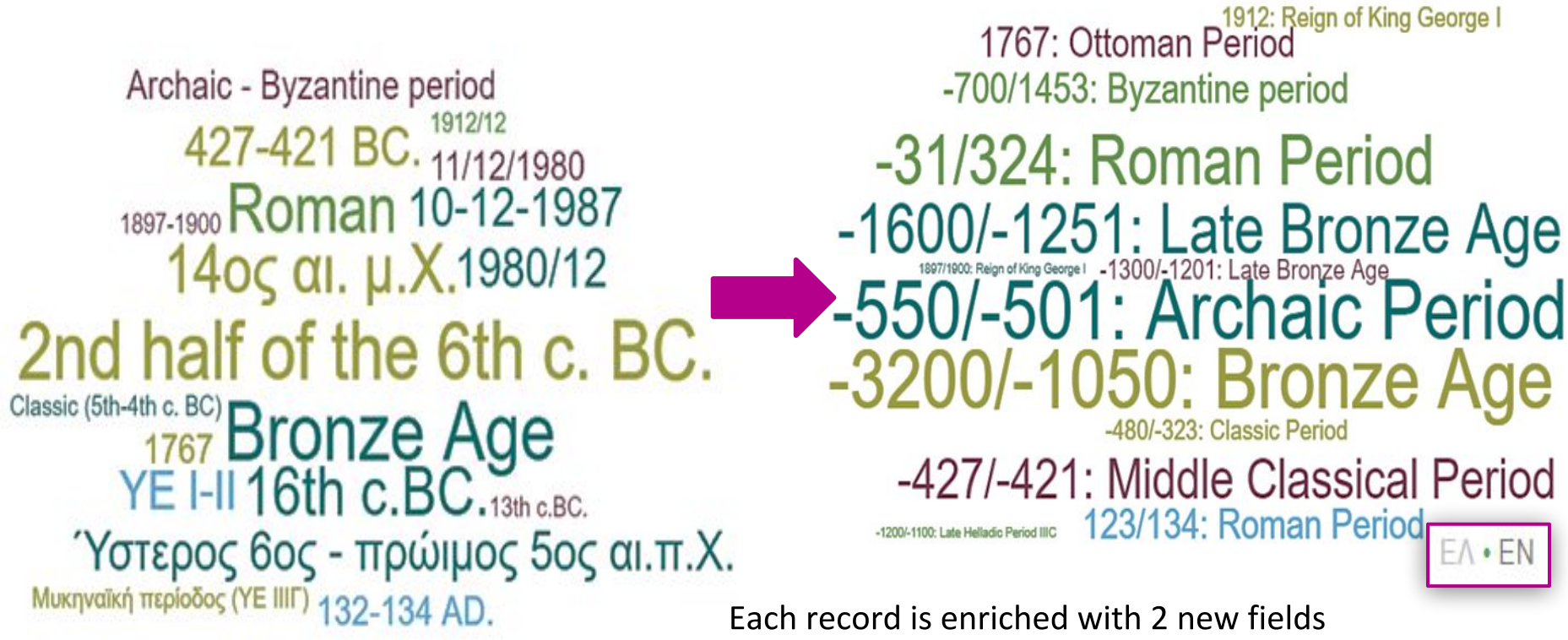
Temporal enrichment |

The vocabulary of EKT Historical Periods

- 94 terms
- 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)



Temporal enrichment and homogenization | Our goal



Each record is enriched with 2 new fields

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

Exhaustive temporal homogenization | Two approaches

Historical period-driven enrichment

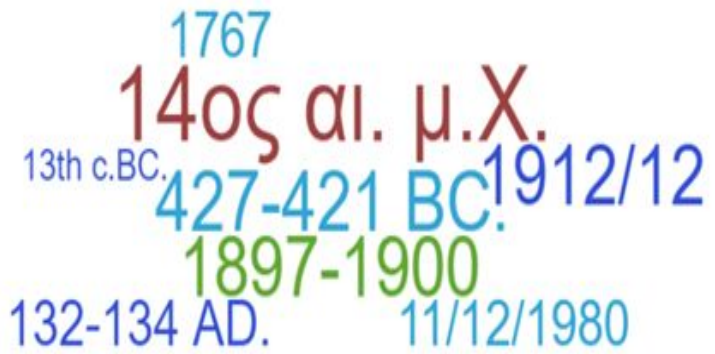
Period Label →

EKT Historical Period → EKT Chronology

year ranges in voc terms

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

Chronology-driven enrichment



Chronology →

EKT Chronology → EKT Historical Period

year ranges in voc terms

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

Historical period-driven enrichment |

Steps per collection

Semantics.gr

1. Insert the collection in **acceptance portal** (if not already)
2. Inspect the documentation quality. Is there a metadata field containing period labels?
   - **dc:date? dcterms:temporal?** (primary field)
   - should we use keywords in **dc:title** as filters? (secondary field)
3. Create **EMRs** in the **Enrichment tool** of Semantics.gr (just like types)
4. Re-index the collection in **acceptance portal and check**
5. Insert (or re-index) and publish the collection in SearchCulture.gr

Historical period-driven enrichment | Collection A

Semantics.gr

Period Label → EKT Historical Period → EKT Chronology

Metadata records

dcterms:temporal	EKT historical period Step 1: EMR - 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

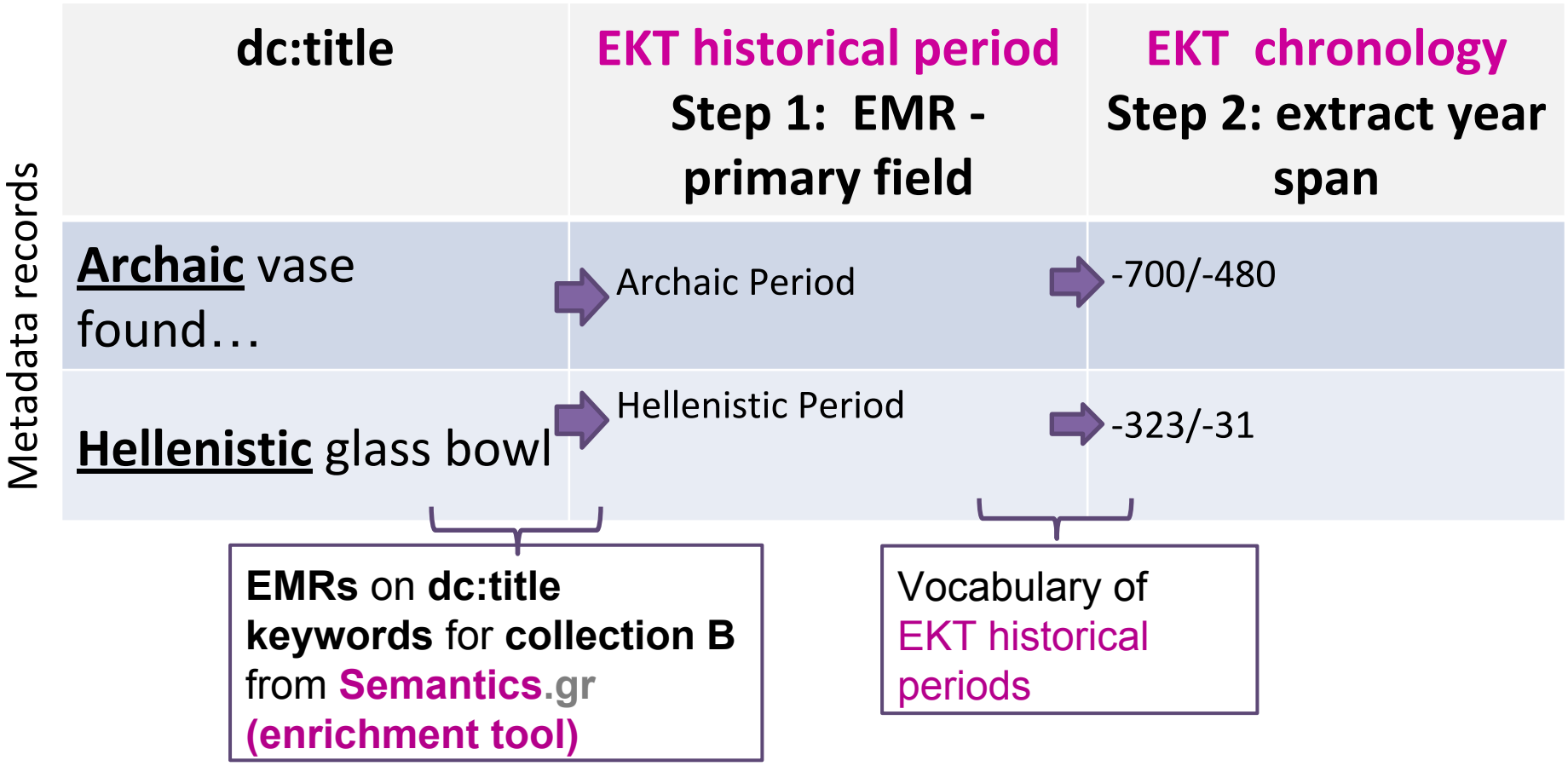
EMRs on **dcterms:temporal** for **collection A** from **Semantics.gr (enrichment tool)**

Vocabulary of EKT historical periods

Historical period-driven enrichment | Collection B

Semantics.gr

Period Label → EKT Historical Period → EKT Chronology



A tool for time normalization

Chronology → EKT Chronology → EKT Historical Period

- 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 EMRs 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

A tool for time normalization | a pattern example

Chronology → EKT Chronology → EKT Historical Period

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 | our pool of 30 patterns

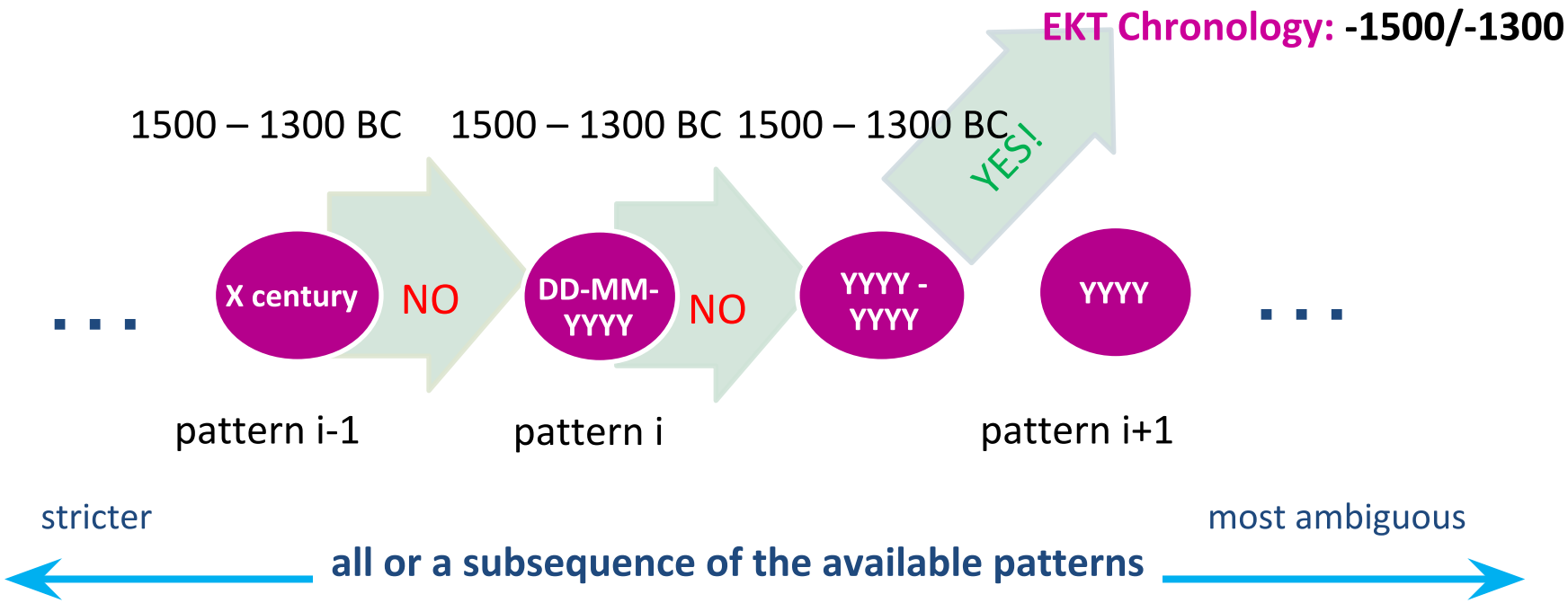
Chronology → EKT Chronology → EKT Historical Period

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**

A tool for time normalization | how it works?

Chronology → EKT Chronology → EKT Historical Period

- patterns have a natural order: from the stricter to the most ambitious
- 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 | Steps per collection

Chronology → EKT Chronology → EKT Historical Period

1. Insert the collection in **acceptance portal** (if not already)
2. Inspect the documentation quality. Which metadata field better describes chronologies?
   - **dc:date? dcterms:temporal? dcterms:issued?**
   - or a descriptive field such as **dc:title**?
3. Add all *common* chronological patterns in collection's configuration
   - or add only specific ones (subsequence)
4. Re-index the collection in **acceptance portal and check**
5. Insert (or re-index) and publish the collection in **SearchCulture.gr**

Chronology-driven enrichment | Collection C

Chronology → EKT Chronology → EKT Historical Period

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

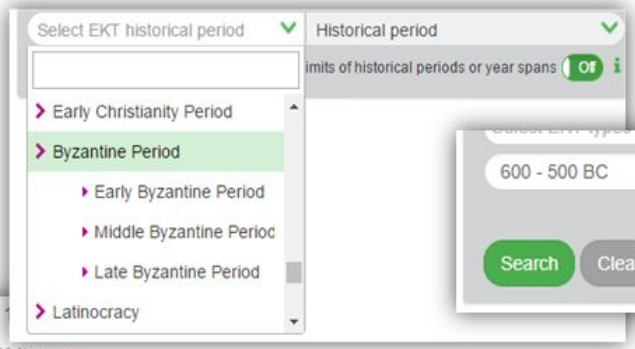
Subsequence of chronological patterns chosen for **collection C**
(time normalization tool)

Vocabulary of
EKT historical periods
Index from year range to terms

New searching, filtering & browsing features | based on temporal enrichment

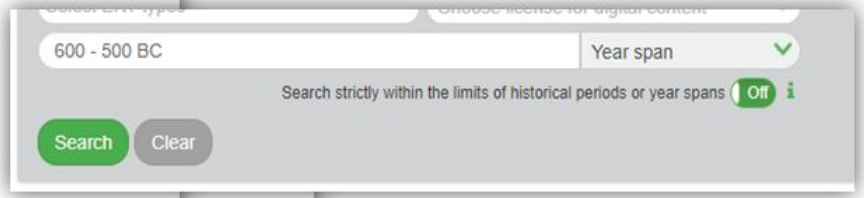
Search Culture.gr

Search by historical period

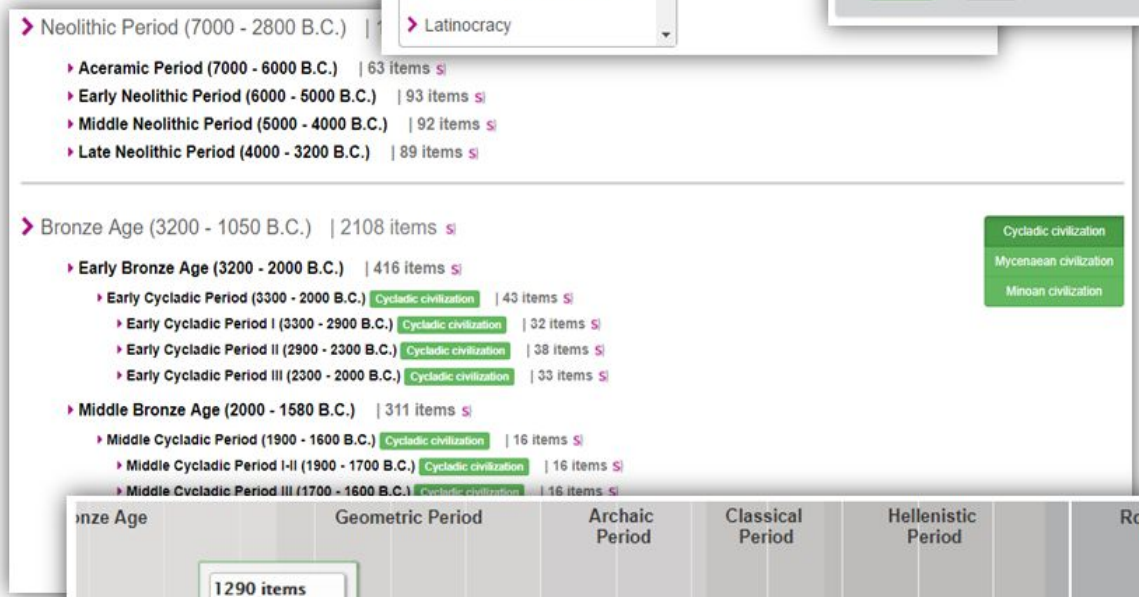


EA • EN

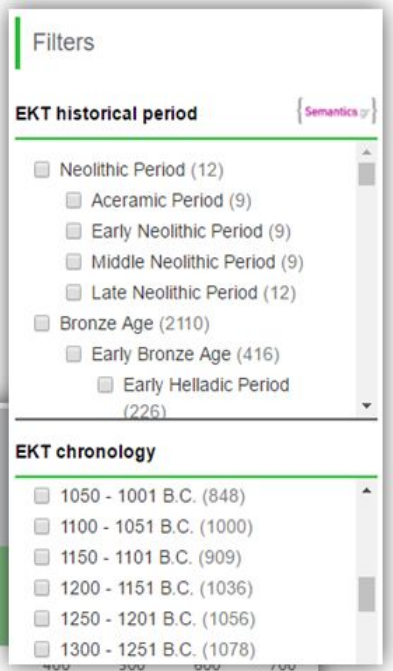
Search by date / chronology

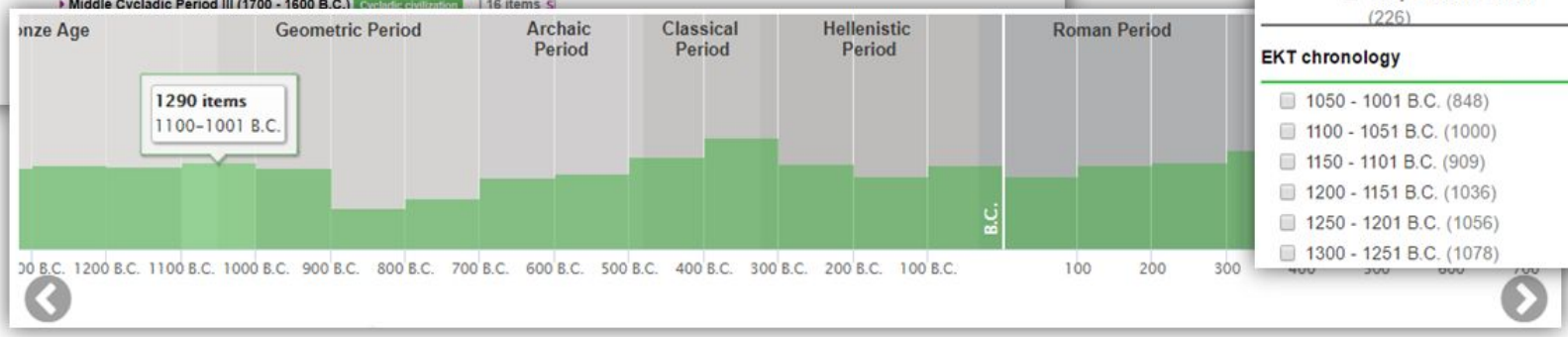


Hierarchical navigation through all historical periods



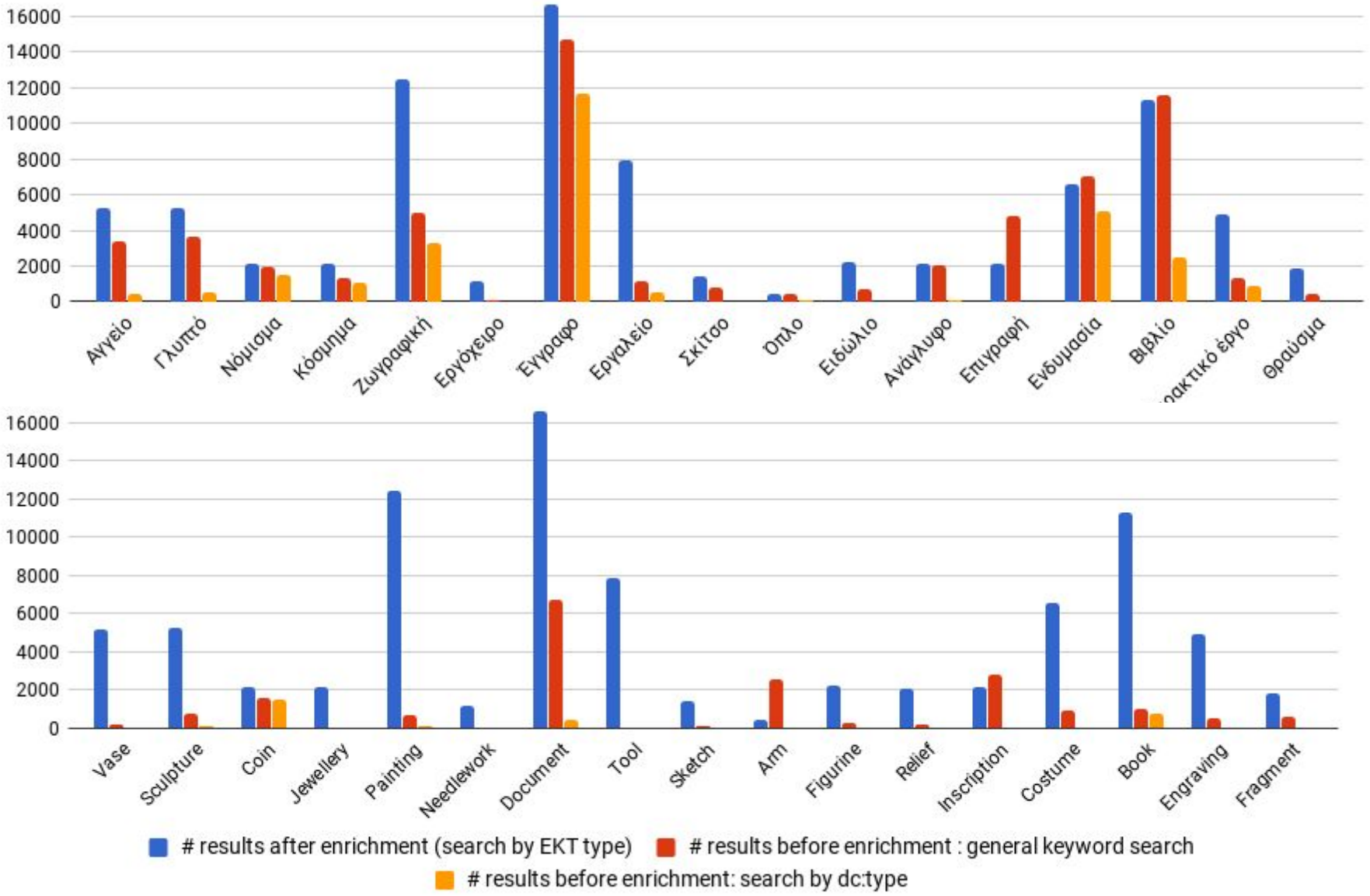
Faceting on periods and chronology



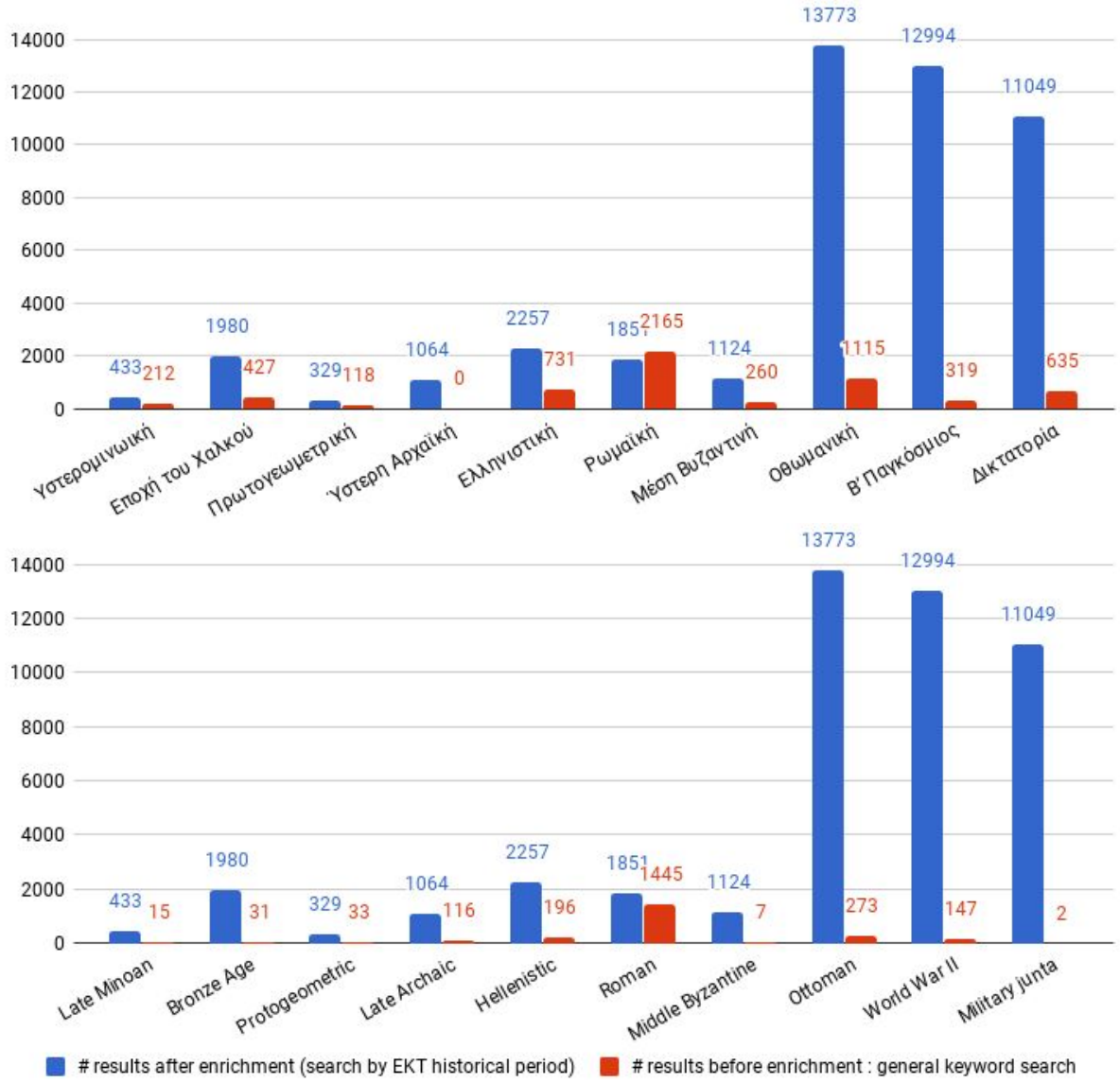


Navigation through an interactive timeline

Type enrichment: improve in searchability



Temporal enrichment: improve in searchability



Learn more

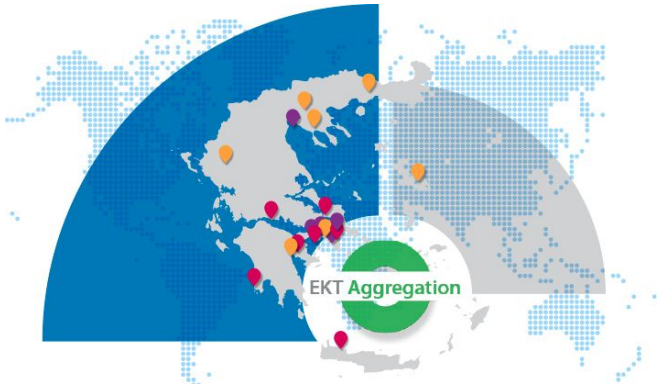
1. Haris Georgiadis, Agathi Papanoti, et al: **The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.gr**, In Proc. **MTSR 2017**
2. Haris Georgiadis, Agathi Papanoti, et al: **Semantics.gr: A self-improving service to repositories and aggregators for massively enriching their content.** In Proc. of the **DHC Workshop of MTSR 2016**
3. Ioanna Ourania Stathopoulou, Haris Georgiadis, Vangelis Banos, et al: **An Open Cultural Digital Content Infrastructure**, In Proc. **DL'2014**
4. Haris Georgiadis, Vangelis Banos, Ioanna Ourania Stathopoulou et al: **Ensuring the quality and interoperability of open cultural digital content: System architecture and scalability**, In Proc. of **IISA'2014**

EKT
ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΤΕΚΜΗΡΙΩΣΗΣ
NATIONAL DOCUMENTATION CENTRE

National Documentation Center

Thank you!

searchculture@ekt.gr



Search Culture.gr | www.searchculture.gr

Semantics.gr | www.semantics.gr

Haris Georgiadis PhD | Computer scientist | hgeorgiadis@ekt.gr
Agathi Papanoti MSc | Archeologist, Information specialist | apapano@ekt.gr
Maria Paschou MSc | Information specialist | mpasxo@ekt.gr
Alexandra Roubani MSc | Librarian, Information scientist | arouba@ekt.gr
Despoina Chardoyveli MSc | Information specialist | dxardo@ekt.gr
Evi Sachini PhD | Director | esachin@ekt.gr

Special thanks to Dimitra Pelekanou | Graphic designer | pelekanou@ekt.gr