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

{SearchCulture.gr} {Semantics.gr}
eContent EKT

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
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- historical documents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

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**

- **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**

MTSR 2017 01.12.2017 | EKT | The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.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
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.
The portal [www.semantics.gr](http://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

- **159** terms
- Hierarchical
- Bilingual (Greek and English)
- Links to Getty AAT
- Schema: skos:Concept
Type enrichment | Steps per collection

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**
Each record is enriched with a new field:

**EKT Type**: from the vocabulary of types
## EKT Types Vocabulary

<table>
<thead>
<tr>
<th>dc:type value</th>
<th>Entry from vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>sculpture art (120 items)</td>
<td><a href="http://scs.gr/sculpture">http://scs.gr/sculpture</a></td>
</tr>
<tr>
<td>greek vases (230 items)</td>
<td><a href="http://scs.gr/vase">http://scs.gr/vase</a></td>
</tr>
<tr>
<td>statuette (15 items)</td>
<td><a href="http://scs.gr/figurine">http://scs.gr/figurine</a></td>
</tr>
</tbody>
</table>
## Type enrichment | EMRs on dc:type & dc:subject values

### Vocabulary

- **EKT Types**
  - [sculpture](http://scs.gr/sculpture)
    - skos:prefLabel "Sculpture"@en | "Γλυπτό"@el **keywords:** "statue"
  - [figurine](http://scs.gr/figurine)
    - skos:prefLabel "Figurine"@en | "Ειδώλιο"@el **keywords:** "statuette"
  - [Jewellery](http://scs.gr/Jewellery)
    - skos:prefLabel "Jewellery"@en | "Κόσμημα"@el **keywords:** "earring"
  - [vessel](http://scs.gr/vessel)
    - skos:prefLabel "Vessel"@en | "Σκύθος"@el
  - [vase](http://scs.gr/vase)
    - skos:prefLabel "Vase"@en | "Αγγείο"@el **keywords:** "amphora", "oenochoe"

### dc:type values | Filters: dc:subject | Entry from vocabulary V1

<table>
<thead>
<tr>
<th>dc:type values</th>
<th>Filters: dc:subject</th>
<th>Entry from vocabulary V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ceramic objects (101 items)</td>
<td>amphora (↗), vase (↗), statuette (↗), figurine (↗),</td>
<td><strong><a href="http://scs.gr/vase">http://scs.gr/vase</a></strong> if filter in [&quot;vase&quot;, &quot;amphora&quot;] auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td>auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong><a href="http://scs.gr/figurine">http://scs.gr/figurine</a></strong> if filter in [&quot;statuette&quot;, &quot;figurine&quot;] auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td>auto</td>
</tr>
<tr>
<td>exhibits (55 items)</td>
<td>earring (↗), amphora (↗),</td>
<td><strong><a href="http://scs.gr/Jewellery">http://scs.gr/Jewellery</a></strong> if filter in [&quot;earring&quot;] auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td>auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong><a href="http://scs.gr/vase">http://scs.gr/vase</a></strong> if filter in [&quot;amphora&quot;] auto</td>
</tr>
<tr>
<td></td>
<td>but NOT in [&quot;earring&quot;] manual</td>
<td></td>
</tr>
</tbody>
</table>
**Type enrichment | EMRs on dc:type values & dc:title**

<table>
<thead>
<tr>
<th>dc:type values</th>
<th>Filters (terms found in dc:title values)</th>
<th>Entry from vocabulary V1</th>
<th>auto</th>
<th>manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D objects (240)</td>
<td>amphora (↗), vase (↗), earring (↗), jewellery (↗)</td>
<td><a href="http://scs.gr/vase">http://scs.gr/vase</a> if filter in [“vase”, “amphora”] but NOT in [“statue”]</td>
<td>auto</td>
<td>manual</td>
</tr>
<tr>
<td>art items (85)</td>
<td>sculpture (↗), statue (↗), figurine (↗)</td>
<td><a href="http://scs.gr/sculpture">http://scs.gr/sculpture</a> if filter in [“sculpture”, “statue”]</td>
<td>auto</td>
<td>auto</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://scs.gr/figurine">http://scs.gr/figurine</a> if filter in [“figurine”]</td>
<td>auto</td>
<td>auto</td>
</tr>
</tbody>
</table>

**Example of dc:title value:** “An amphora from the Mycenaean period”

**Vocabulary EKT Types**

- http://scs.gr/sculpture
  - skos:prefLabel "Sculpture"@en | “Γλυπτό”@el keywords: “statue”
- http://scs.gr/figurine
- http://scs.gr/jewellery
  - skos:prefLabel “Jewellery”@en | “Κόσμημα”@el keywords: “earring”
- http://scs.gr/vessel
  - skos:prefLabel “Vessel”@en | “Σκήπτρο”@el
- http://scs.gr/vase
The GUI of the enrichment tool
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

Hierarchical navigation through all types

Faceting on types

Tag cloud of popular types
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)

<table>
<thead>
<tr>
<th>EKT Historical Periods</th>
<th>Category</th>
<th>Time periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesolithic Period</td>
<td>edm:Timespan</td>
<td>8000 - 7000 BC</td>
</tr>
<tr>
<td>Neolithic Period</td>
<td>edm:Timespan</td>
<td>7000-2800 BC</td>
</tr>
<tr>
<td>Aceramic Period</td>
<td>edm:Timespan</td>
<td>7000-6000 BC</td>
</tr>
<tr>
<td>Early Neolithic Period</td>
<td>edm:Timespan</td>
<td>6000 - 5000 BC</td>
</tr>
<tr>
<td>Middle Neolithic Period</td>
<td>edm:Timespan</td>
<td>5000 - 4000 BC</td>
</tr>
<tr>
<td>Late Neolithic Period</td>
<td>edm:Timespan</td>
<td>4000 - 3200 BC</td>
</tr>
<tr>
<td>Bronze Age</td>
<td>edm:Timespan</td>
<td>3200 - 1050 BC</td>
</tr>
</tbody>
</table>

MTSR 2017 01.12.2017 | EKT | The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.gr
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**

According to EMR set in the enrichment tool of Semantics.gr

Period Label → EKT Historical Period → EKT Chronology

Year ranges in voc terms

---

**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

---

MTSR 2017 01.12.2017 | EKT | The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.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

**Period Label → EKT Historical Period → EKT Chronology**

<table>
<thead>
<tr>
<th>dcterms:temporal</th>
<th>EKT historical period</th>
<th>EKT chronology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Step 1: EMR - primary field</td>
<td>Step 2: extract year span</td>
</tr>
<tr>
<td>Post-Byzantine Period</td>
<td>Ottoman Period</td>
<td>1453/1821</td>
</tr>
<tr>
<td>Second War</td>
<td>World War II</td>
<td>1940/1944</td>
</tr>
<tr>
<td>Middle - Late Hellenistic Years</td>
<td>Middle Hellenistic Period - Late Hellenistic Period</td>
<td>-220/-31</td>
</tr>
</tbody>
</table>

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

**Vocabulary of EKT historical periods**
### Historical period-driven enrichment

**Period Label → EKT Historical Period → EKT Chronology**

<table>
<thead>
<tr>
<th>Metadata records</th>
<th>dc:title</th>
<th>EKT historical period</th>
<th>EKT chronology</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Archaic</em> vase found...</td>
<td></td>
<td>Archaic Period</td>
<td>-700/-480</td>
</tr>
<tr>
<td><em>Hellenistic</em> glass bowl</td>
<td></td>
<td>Hellenistic Period</td>
<td>-323/-31</td>
</tr>
</tbody>
</table>

**EMRs on dc:title keywords for collection B from Semantics.gr (enrichment tool)**

**Vocabulary of EKT historical periods**
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”, “πΧ.”
  - placeholder for “first half of”: “first half of”, “1st half of”, “first A”, “πρώτο μισό του”
- 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 SearchCulture.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: \\[(?#\text{century\_identifier})\(.*\)?(\d{1,4})\]?(\#s0)?(\(\#bc\_ad\)\(.*\))?\(\(\#s1\).?)

Extraction pointers: 4

Predefined placeholders:
- #century_identifier
  - early: “early”, “first quarter of”, “beginning of”, “αιρχές”...
  - late: “late”, “end of”, “end of”, “τελος”, ...
- #bc_ad

Custom placeholders:
- #s0: ος|ος|ου|st|nd|rd|th
- #s1: αιωνας|ai|a|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
### Chronology → EKT Chronology → EKT Historical Period

<table>
<thead>
<tr>
<th>Chronological Pattern Class</th>
<th># of patterns</th>
<th>Examples</th>
</tr>
</thead>
</table>
| 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               
tέλος 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.

EKT Chronology: \(-1500/-1300\)

1500 – 1300 BC 1500 – 1300 BC 1500 – 1300 BC

\[ \text{pattern } i-1 \quad \text{pattern } i \quad \text{pattern } i+1 \]

stricterversion

all or a subsequence of the available patterns

most ambiguous
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**

<table>
<thead>
<tr>
<th>Metadata records</th>
<th>dc:date</th>
<th>EKT chronology</th>
<th>EKT historical period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Late 5th century</td>
<td>471/500</td>
<td>Early Byzantine Period</td>
</tr>
<tr>
<td></td>
<td>7th c. B.C - mid 6th c. BC</td>
<td>-700/-551</td>
<td>Early Archaic to Middle Archaic Period</td>
</tr>
<tr>
<td></td>
<td>03/11/1980</td>
<td>1980</td>
<td>Regime change</td>
</tr>
</tbody>
</table>

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

**Vocabulary of EKT historical periods**
Index from year range to terms

---

MTSR 2017 01.12.2017 | EKT | The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.gr
New searching, filtering & browsing features based on temporal enrichment

Search Culture.gr

Hierarchical navigation through all historical periods

Search by historical period

- Select EKT historical period
- Historical period
  - Early Christianity Period
  - Byzantine Period
- Neolithic Period (7000 - 2800 B.C.)
  - Aceramic Period (7000 - 6000 B.C.)
  - Early Neolithic Period (6000 - 5000 B.C.)
  - Middle Neolithic Period (5000 - 4000 B.C.)
  - Late Neolithic Period (4000 - 3200 B.C.)
- Latincocy

Bronze Age (3200 - 1050 B.C.)
- Early Bronze Age (3200 - 2000 B.C.)
- Middle Bronze Age (2000 - 1500 B.C.)

Faceting on periods and chronology

Filters
- EKT historical period
  - Neolithic Period
  - Aceramic Period
  - Early Neolithic Period
  - Middle Neolithic Period
  - Late Neolithic Period
  - Bronze Age
- EKT chronology
  - 1050 - 1001 B.C.
  - 1100 - 1051 B.C.

Navigation through an interactive timeline
Type enrichment: improve in searchability
Temporal enrichment: improve in searchability

MTSR 2017 01.12.2017 | EKT | The semantic enrichment strategy for types, chronologies and historical periods in SearchCulture.gr
Learn more

Thank you!

searchculture@ekt.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 Chardoeyveli MSc | Information specialist | dxardo@ekt.gr
Evi Sachini PhD | Director | esachin@ekt.gr

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