Using type and temporal semantic enrichment to boost content discoverability and multilingualism in the Greek Cultural Aggregator SearchCulture.gr[1]

Haris Georgiadis, Agathi Papanoti, Maria Paschou, Alexandra Roubani, Despina Hardouveli, Evi Sachini

National Documentation Centre / National Hellenic Research Foundation
Athens, Greece
{hgeorgiadis, apapano, mpasxo, arouba, dxardo, esachin}@ekt.gr

Abstract. Most aggregators face challenges regarding searchability, discoverability and visual presentation of their content due to metadata heterogeneity across the collections. Particularly for cultural and historical material, keyword-based searching is far from sufficient. Structured item types and temporal information are key metadata for the discoverability of cultural heritage content. We developed an innovative metadata enrichment and homogenization scheme for types and temporal information that is both effective and user-friendly and we embedded it in the ingestion workflow of SearchCulture.gr, the greek cultural heritage aggregator developed by the National Documentation Centre (EKT). Two key components of the enrichment scheme are Semantics.gr, a platform for publishing vocabularies that contains a mapping tool for massive semantic enrichment, and a parametric tool for chronological normalization. We enriched and homogenized the aggregated content with respect to types and temporal information which subsequently allowed us to develop advanced multilingual search and browsing features, including hierarchical navigation on types and historical periods, searching and faceting on types, time spans and historical periods, a tag cloud of types and an interactive timeline/histogram.

[1] Preprint of: Georgiadis, H., Papanoti, A., Paschou, M., Roubani, A., Hardouveli, D. and Sachini, E (2018) '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.

Keywords: aggregator, semantic enrichment, linked data, automatic categorization, vocabularies, thesauri, cultural heritage, historical periods, time-driven search, temporal coverage, timeline, multilingualism

Biographical notes:

Haris Georgiadis works as the Head of the e-Services Unit of the E-infrastructure and Information Systems Department of the National Documentation Centre / National Hellenic Research Foundation (EKT/NHRF). He holds a Bachelor degree in Informatics from the University of Piraeus, a MSc at Information Systems and a PhD degree at Information Systems from the Athens University of Economics and Business.

Agathi Papanoti works as a Scientific Associate at National Documentation Centre / National Hellenic Research Foundation (EKT/NHRF). She holds a Bachelor of Arts (BA) in Archeology and Art History from the National and Kapodistrian University of Athens and a Master of Arts (MA) in Arts and Heritage Management from the University of Sheffield.

Maria Paschou works as a Sci Advisor to General Secretariat for Research and Technology of the Ministry of Education. She worked as the Head of Information services, Library Department of the National Documentation Centre / National Hellenic Research Foundation (EKT/NHRF). She holds a Bachelor in Biology from the National and Kapodistrian University of Athens.

Alexandra Roubani works as a Librarian (MLIS) at National Documentation Centre / National Hellenic Research Foundation (EKT/NHRF). She has Bachelor in Librarianship from the Technological Educational Institute of Athens, a Bachelor in Communication, Media and Culture from the Panteion University of Social and Political Sciences and a Master of Library & Information Science from the Ionian University.

Despina Hardouveli works as the Head of Documentation Unit at the National Documentation Centre / National Hellenic Research Foundation (EKT/NHRF). She holds a Bachelor in Biology and a master degree from the National and Kapodistrian University of Athens.

Evi Sachini is the Director of National Documentation Centre / National Hellenic Research Foundation (EKT/NHRF). She holds a Bachelor degree in Chemical

Engineering from the National Technical University of Athens and a PhD degree at Chemical Engineering from the same University.

1 Introduction

SearchCulture.gr (https://www.searchculture.gr) is the Greek Aggregator for Cultural Heritage Content and National Provider for Europeana. It is being developed by EKT as part of an aggregation and preservation framework established and implemented by EKT in collaboration with the "Digital Convergence" OP (National Strategic Reference Framework). The objective was to ensure the sustainability and reusability of content produced by publicly funded digitization projects, to provide central access to digital cultural resources and to integrate them to Europeana. EKT was assigned with the important role of National Aggregator and Preservation Infrastructure for projects funded by Cultural Heritage Digitisation related Calls of the OP. The role involves setting the quality and interoperability specifications for publicly funded projects (metadata, digital files and systems), validating and certifying systems and content, publishing the produced content in SearchCulture.gr, depositing it in a preservation platform, and, eventually, delivering it to Europeana.

The digital resources accessible via SearchCulture.gr include digital representations of archaeological items, historical documents and manuscripts, folklore items, works of art, cartographic material, books and oral history. The digital files are mainly photographs and other images, pdfs, 3D digital representations and audiovisual material. So far, SearchCulture.gr hosts more than 430,000 items from 67 collections contributed by 53 institutions that include museums, archives, ephorates of antiquities, municipalities and cultural foundations. SearchCulture.gr has contributed to Europeana 31 collections (more than 114,000 items).

The portal went into production in 2015. Soon our aim became to increase searchability and discoverability of the content that is aggregated in SearchCulture.gr by providing new means of search, filtering, browsing and visual presentation of the content.

When it comes to cultural and historical material, keyword-based searching is far from sufficient. Structured item types and temporal information are key metadata for cultural heritage content as presented in Fig. 1. Users expect to be able to: search by type and temporal criteria (year ranges or historical periods), explore by browsing and filtering through types, historical periods and year ranges and submitting combined queries such as "icons from the late byzantine period", "manuscripts dated from 1850 to 1900" and "sculptures dated strictly within the middle classical period of Greece".

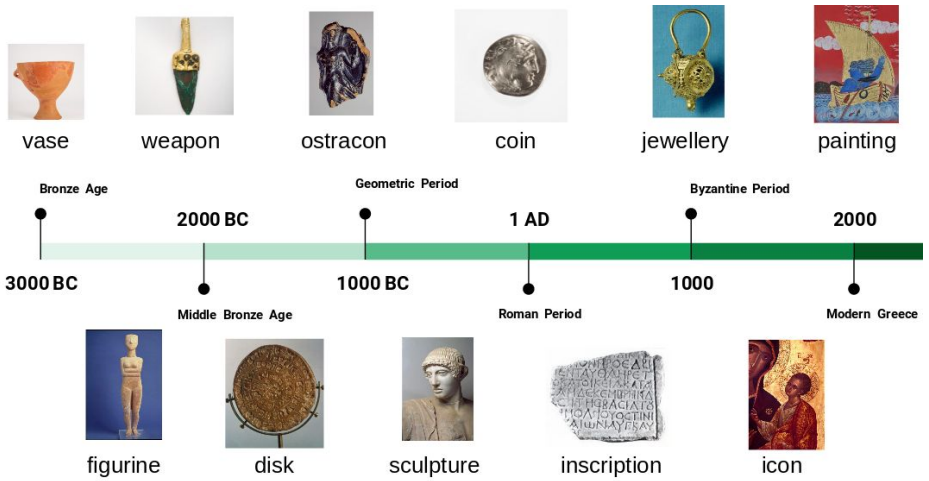


Fig. 1. Item types and temporal information: key metadata for cultural content discoverability

To support these kind of search and browsing functionalities, we focused on two types of metadata, item type (dc:type) and chronological information (dc:date, dcterms:created or dcterms:temporal, depending on the collection and the material genre). However as is the case with aggregators, the original documentation is extremely heterogeneous across the collections making impossible to support these features without intensive semantic enrichment and homogenization.

The heterogeneity of dc:type derives from term variations (different languages, synonyms, mixing plural and singular etc) and from different documentation methodologies (ranging from extremely general terms, such as “exhibit”, to very specialized terms, such as “oenochoe”, a specific type of vase) (Fig. 2 (a)).

When it comes to temporal fields, some providers use *period labels* which are as problematic as the aforementioned type labels (e.g. “Late Helladic Period” vs. “Mycenaean Period”, “Classical Period” vs. “Classic times”). Most providers use *chronological values*, such as dates, centuries, years and their *interval* counterparts (year and centuries ranges) which are also highly heterogeneous due to the use of different time encodings (e.g. “198?”, “-1500”, “~600 BC”, “11-12-1932”), languages (e.g. “BC” “π.X.”), time granularity (century, decade, year, date, datetime) or literal descriptors (e.g. “Early”, “First half of”)(Fig. 2 (b)).

(a) (b)

Fig. 2. The heterogeneity of types (a) and temporal values (b)

As a solution we developed a semantic enrichment and homogenization scheme. It is based on *www.semantics.gr*, a platform for publishing vocabularies and thesauri that includes *a mapping tool* for content enrichment and contextualization. The tool sets mapping rules from multiple metadata field values to vocabulary terms. It uses a self-improving automatic suggestion mechanism and additionally supports the curator when intervening in the enrichment process. We developed and published two vocabularies in Semantics.gr, one for item types and another for Greek historical periods. We used the tool to map types and period labels included in the original metadata to terms of the two vocabularies. We also developed a *parametric tool for the normalization of chronological values* which is based on regular expression processing. As a result the content of SearchCulture.gr was enriched and homogenized in respect to types, chronologies and historical periods, allowing us to publish it as Linked Data and to enhance the portal with new multilingual search and navigation features.

This work is an extended version of a paper published and presented in the 11th International Conference on Metadata and Semantic Research (MTSR 2017), which took place in Tallinn, Estonia, from November 28 to December 1, 2017 [15].

2 Related Work

The issue of appropriate digitization, documentation and preservation of cultural and scientific heritage has been widely recognized for its significance. A great number of publicly funded digitization projects have produced numerous repositories, digital libraries, archives and publishing platforms. Key factors in achieving the appropriate return from the corresponding investments are i) ensuring, to the greatest extent possible, the quality of the output at the level of both the metadata records and the digital files, ii) establishing actions towards the digital long-term preservation of the

content [23], iii) ensuring interoperability for systems and content (including provision via open APIs according to established standards, use of standard metadata and file formats etc.), iv) setting clear licences for the digital resources, preferably open access licenses when possible v) establishing large-scale national and international aggregation schemes that offer central access to the fragmented content - thus increasing its discoverability and re-use potentials - and often set specifications and guides with good practices that boost all previous factors [1][21][22]. However, discoverability and re-use can be influenced by the level of metadata heterogeneity and semantic interoperability [19]. Especially for cultural heritage, all aggregative data infrastructures face the problem of defining a conversion model that limits the loss of information and implies that a deep knowledge of vocabularies and ontologies is required in order to propose a functional and exhaustive model [18]. Overall for a successful aggregation and management of extremely heterogeneous digital data (such as this of SearchCulture.gr) requires a multi-disciplinary collaboration with extended knowledge of the cultural content as noted in [19][20].

Many aggregators and libraries use semantic enrichment techniques to deal with heterogeneity. Europeana [3], the European Library [16], aggregation platforms such as MoRe [2] , MINT [11] and LoCloud [12], research projects and creative industries such as PATHS [8] and Ontotext[2] have developed and use automatic semantic enrichment tools that cover mainly concepts, agents and places. Complete automated enrichment on structured fields (such as dc:type) adopts an “enrich-if-you-can” strategy, horizontally, resulting in non negligible percentages of mistakes [6] and in relatively low enrichment coverage - despite using extremely large target thesauri, such as DBpedia and Geonames [17]. Automated annotation methods on more descriptive fields (such as dc:title) yield similar results [8]. All these techniques do increase searchability and multilingualism. However, due to the relatively low enrichment coverage, the large target thesauri and the non negligible percentage of enrichment mistakes, they cannot not achieve sufficient homogenization that would allow aggregators to offer advanced ways of content exploration (browsing, faceting on enriched fields). Our semantic enrichment scheme achieves homogenization because i) it can be adjusted to the documentation particularities of the individual collections which increases the enrichment coverage ii) it combines self-improving automatic and fuzzy-based suggestions with a suit of tools that allows easy and effective curation and disambiguation, which increases further enrichment coverage while eliminating enrichment mistakes and iii) uses smaller and more compact target vocabularies. TMP tool [13] of the AthenaPlus project is a platform for creating vocabularies that offers a mapping functionality which allows users to define equivalent relations between concepts from different vocabularies. However, unlike

[2] https://ontotext.com

Semantics.gr, it supports SKOS vocabularies (which is not suitable for time periods), while the mapping tool cannot perform more complex mappings (such as mappings from combined multiple fields or from keywords contained in descriptive fields) and does not have a self-improving auto-suggestion mechanism.

Particularly for temporal enrichment, some aggregators enrich items described with period labels using timespan vocabularies (e.g. [3]), suffering, though, the abovementioned problems. Some attempt to homogenize chronological values to some extend as far as they conform to specific date formats ([2][11]). The work in [14] presents an automatic method for the extraction of time periods related to ontological concepts from the web. The method involves an information extraction phase which uses simple regular expressions to extract years from documents. However, the method uses simple regular expressions that cover only year descriptions missing language-depended patterns and time descriptions of different granularity (e.g. "the NNth Century"). Our time normalization method is fully extentensible and parametric, takes into consideration language descriptors and covers four types of temporal expressions, centuries, range of centuries, years/dates and year ranges.

Temporal enrichment and normalization schemes for aggregated metadata don't handle items with temporal information uniformly, i.e. they don't assign period labels to items described with chronologies (a complex and error-prone task as highlighted in [9][10]) and vice versa (chronologies to items described with period labels). Our enrichment scheme supports chronological search and browsing both by year ranges and historical periods across all items with original chronological metadata, either explicit (temporal fields) or implicit (e.g. keywords in titles), regardless whether they were originally described with chronologies or with period labels.

3 SearchCulture.gr and the aggregation infrastructure that lies underneath

SearchCulture.gr is the public portal of the Aggregator platform, a digital content aggregation system. The Aggregator platform is a component of a broader infrastructure (Fig. 3) comprised of four other systems: Harvester, a digital content harvesting system, Validator, a system for validating content against interoperability and quality specifications, Semantics.gr (Section 3) and Preservator system used for the secured and long-term preservation for digital resources. The infrastructure along with the individual systems were designed and implemented by EKT.

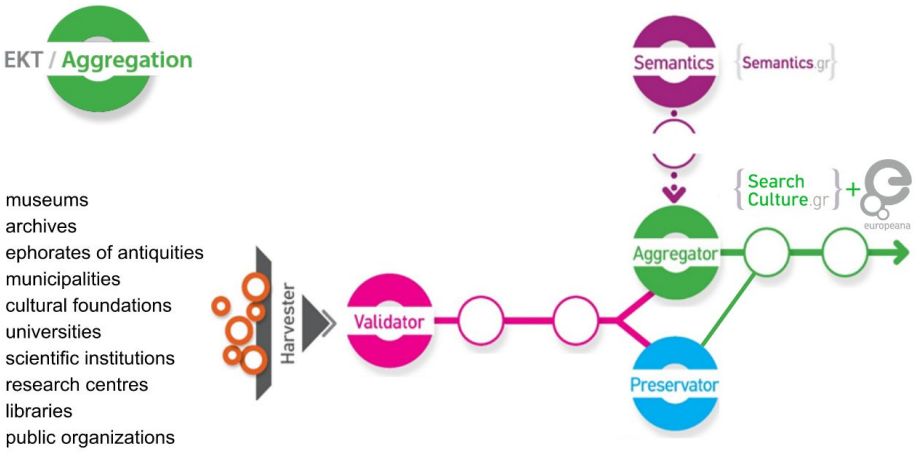


Fig. 3. The EKT aggregation infrastructure

3.1 The Harvester system

The Harvester system collects metadata and digital files from remote repositories and digital libraries using various channels, such as the OAI-PMH and OAI-ORE protocols. It can be configured to append fields in the metadata records that are being harvested by OAI-PMH by visiting and parsing the respective item web pages on-the-fly and extracting additional information.

The Harvester system stores the content and provides it on request by a user or an external system via the graphical web user interface or its RESTful API, respectively. The collected digital files can be converted in preview images (thumbnails) through a Harvester's embedded service. Harvester feeds with metadata and digital files the Validator, Aggregator, Semantics.gr and Preservator systems.

3.2 The Validator system

Validator is the information system that confirms the extent to which individual repositories comply with content quality and interoperability standards [1]. To impose interoperability, the aggregation and preservation framework established by the "Digital Convergence" OP and EKT included a set of specifications to be followed by the project contractors and the necessary checkpoints that are linked to funding. EKT has published the specifications, developed the Validator tool and validated and certified all projects funded by the OP (42 so far) before publishing their content to SearchCulture.gr and Europeana [1]. The specifications include interoperability guidelines at the systems level (support of OAI-PMH, OAI-ORE and Open Search protocols and APIs, use of persistent identifiers), at the level of metadata (common

metadata schemata such as OAI-DC, ESE and EDM, documentation guidelines, encoding standards for specific types of values, use of controlled vocabularies, use of licenses for digital resources and metadata etc.) and at the digitization level (image quality specifications per material category, OCR for PDFs etc.).

The Validator tool works closely with the Harvester to retrieve content which is then validated against a set of extensible and configurable validation rules that encode the content quality and interoperability specifications. Highly granular results are recorded and are made available through a web GUI as analytical and aggregated reports.

3.3 **The Aggregator platform**

The Aggregator retrieves metadata and thumbnails from the Harvester system. It transforms and stores the accumulated metadata in EDM metadata format. EDM is an RDF model recommended by Europeana for the representation of cultural content metadata [4]. It incorporates mechanisms that allow metadata to retain semantic references to vocabularies, thesauri and other resources, thus making them available as Linked Data. As a result SearchCulture.gr can host naitively semantically linked content. The Aggregator supports metadata input in seven alternative formats: EDM, ESE, OAI-DC, DC-DS-XML, Qualified DC (QDC), HEAL and MODS. Institutions that do not contain semantically linked content can provide their metadata in OAI-DC, ESE, QDC, HEAL and MODS. Institutions that have semantically linked content can provide their metadata in EDM (public version) or DC-DS-XML. The DC-DS-XML type is a Dublin Core XML representation (Metadata Terms) that allows semantic references in vocabularies and thesauri terms[3]. The aggregator transforms the original metadata from the supported input formats to the EDM-internal type in order to store them in its database.

The Aggregator platform has a complete graphic management environment that allows the authorized user to perform as automated as possible all the necessary aggregation procedures: managing information for providers and collections/repositories, setting normalization, transformation and cleansing rules per collection, initiating new metadata or thumbnail ingestion processes, semantically enriching collections and dynamically parameterising the web portal. It is integrated with the enrichment tool of Semantics,gr (Section 4) and with the a time normalization tool that extracts chronologies from temporal metadata (Section 5).

[3] http://dublincore.org/documents/dc-ds-xml/

3.4 The Preservator system

Digital preservation was a key component of the framework established by EKT and the "Digital Convergence" OP. The Preservator system keeps back-up copies of the aggregated content (metadata and digital objects) in the trustworthy Cloud-based infrastructure of EKT, to ensure its long-term secure storage and accessibility, protecting the born-digital and digitised resources from physical and technical calamities. EKT has recently loosely coupled the Preservator system with SearchCulture.gr in order to guarantee that the digital resources will still be accessible via SearchCulture.gr and Europeana even when individual repositories face technical issues. Thus, for repositories with dysfunctions, EKT temporally integrates their metadata that are published in SearchCulture.gr (and Europeana) with the digital resources deposited in the Preservator system. So far, 7 out of 42 repositories of beneficiaries of the OP Calls are out of order and their digital resources remain accessible, temporally served by EKT's infrastructure.

4 Semantics.gr: the semantic enrichment tool

Semantics.gr is a platform developed by EKT for creating and publishing RDF-based vocabularies and thesauri and a tool for semantic enrichment allowing repositories and aggregators to enrich their metadata records with references to vocabulary terms.

4.1 Creating and publishing vocabularies and thesauri

Semantics.gr was initially created as a platform where EKT and other institutions will create and publish RDF-based vocabularies and thesauri [5][7]. The published vocabularies are disseminated as Linked Data through an open portal that contains a search engine and presents the vocabularies for hierarchical navigation. Institutions can be registered in the platform and obtain user accounts to create, process and publish their own vocabularies. The vocabularies can be linked to other vocabulary or thesaurus entries, both internal or external ones.

Semantics.gr has a parametric mechanism for defining vocabulary schemata which are modeled as owl classes (for example skos:Concept) that group parametric owl properties. The creation and configuration of owl classes and their properties are built via a user friendly management web UI. Until now we have successfully modelled skos:Concept and the contextual classes introduced by Europeana, edm:Timespan, edm:Agent and edm:Place [4].

When institutions create a new vocabulary they first have to choose one of the registered owl classes. After that, they can start creating vocabulary entries using a

dynamic form that embeds all the properties of the respective owl class as form components whose functional and validation behavior reflects the respective property parameters. Semantics.gr embeds functions for collaborative curation and editorial process, such as posting comments and approving changes. At the completion of a vocabulary, the owning institution can choose to publish it, thus making it publicly accessible through Semantics.gr open portal.

4.2 The mapping-based enrichment tool

The semantic enrichment tool of Semantics.gr has a GUI environment with advanced automated functionalities that help the curator easily define *Enrichment Mapping Rules (EMRs)* per collection from distinct metadata values to vocabulary entries. The tool accesses collection metadata via OAI-PMH in order to run count aggregations on specific metadata fields. Note that the tool only stores distinct metadata field values and not the entire metadata. The tool can be used by both repositories and aggregators to enrich their content. Particularly for an aggregator, it is recommended that the EMRs are set per collection in order to handle separately the documentation particularities of each institution. After EMRs for a collection are set, they can be served on request via a REST API in json format which can be used by the aggregator or repository to enrich the collection in a bulk and straightforward one-pass fashion.

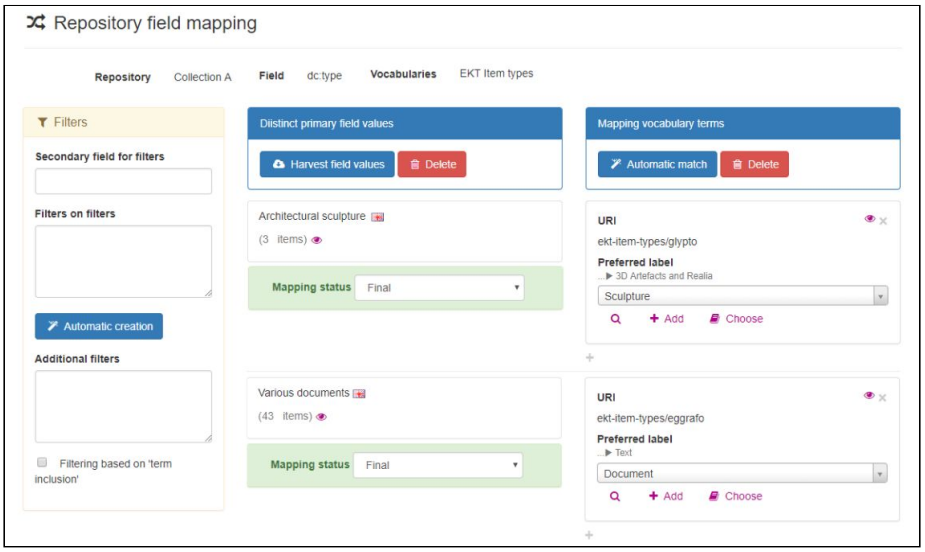


Fig. 4. EMRs on primary field values (dc:type)

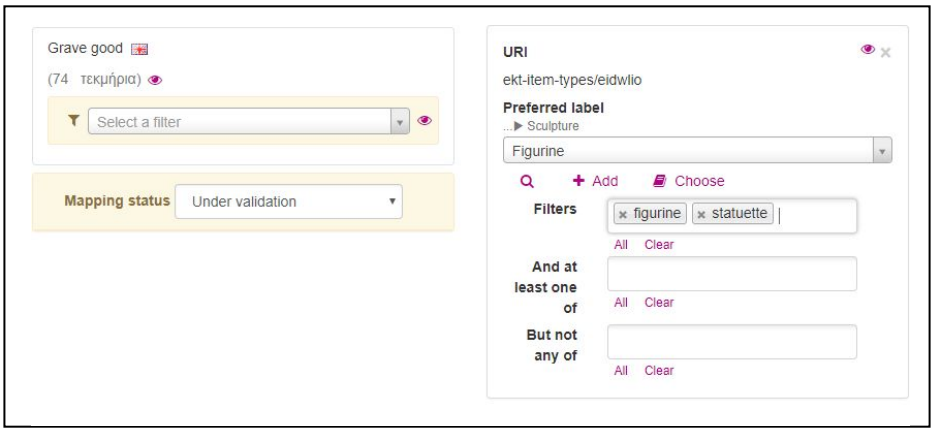


Fig. 5. EMRs on primary field values (dc:type) and secondary field filters (dc:subject)

The EMRs in their simplest form are defined per distinct value of a predefined metadata field (for example dc:type or dcterms:temporal), which is called *primary field*. Fig. 4 is a screenshot of the tool with primary field-based EMRs.

In special cases the curator can choose a second metadata field (for example dc:subject) to create more precise EMRs in case the documentation of the primary field is poor. We call this metadata field *secondary field* and its values *filters*. For example, a metadata record may have a dc:type value "folklore object" but a dc:subject value "Jewel" that reveals a much more accurate type. Fig. 5 is a screenshot of the tool with an example of an EMR using secondary field. In this example items with dc:type "Grave good" are enriched with the term "figurine" when their dc:subject value (filter) is either "figurine" or "statuette".

The enrichment tool supports automatic suggestion of EMRs which is based on string similarity matching between metadata field values and indexed labels of vocabulary entries (e.g. skos:prefLabel and skos:altLabel). The automatic mapping suggestion is very effective and efficient leveraging the indexing system of Semantics.gr search engine, namely Apache Solr. The tool can be easily configured to be loosely coupled to the aggregator (or repository) search portal (using deep linking) allowing the curator to easily search the collection for items having the specific values on primary and secondary fields. This is very crucial especially when the curators resort to using a secondary field with different semantics, since it allows them to easily inspect the validity of EMRs. The curator can create complex logical expressions using the logical operators AND, OR and NOT on the filters of an EMR in order to create finer and more precise mappings and avoid false positives. For instance an EMR may assign items with dc:type "image" to the vocabulary term "vase" if they have a dc:subject value "vase" or "oenochoe" but NOT a dc:subject

value “drawing representation”. Another EMR could map items with dc:type “image” and dc:subject “drawing representation” to the vocabulary term “drawing”.

When the automatic suggestion function fails to produce correct rules, the curator can set EMRs manually. The set of manual mappings from metadata values to - usually similar or broader - vocabulary terms, constitute valuable knowledge that we leverage to improve effectiveness of auto-suggestion in future enrichments hence reducing manual assignments. The curator decides whether a manual EMR should be remembered by *bookmarking* it. When an EMR is bookmarked, its original metadata value is stored in a hidden, special kind of label field called *keyword* inside the mapped vocabulary entry. The keyword field is indexed by Apache Solr just like the preferable and alternative labels. This way, the vocabulary entry can also be automatically suggested in a future mapping if one of its keywords matches the metadata value.

In certain cases, the curator can choose a highly selective descriptive field (the number of its distinct values approaches the number of all items) as a secondary field, such as dc:title or dc:description, if the values contain words or phrases that can reveal the appropriate vocabulary entry. For example a dc:title “An amphora from Attica” implies that the item is a vase. The tool searches inside such values for specific words or phrases. We call these words and phrases *search term space*. The search term space can be defined manually by the curator or can be set automatically by the tool. To do so, the tool scans all distinct values of the descriptive field (e.g. all titles) and searches for inclusion matches against all the indexed labels of the vocabulary (preferable labels, alternative labels and keywords). Only the matching vocabulary labels are exposed as available filters.

5 A tool for extracting years or year ranges from various time formats

We developed an autonomous parametric tool that extracts years or year ranges from temporal metadata fields according to an extensible and configurable set of rules. It is integrated with the ingestion data-flow of the Aggregator platform.

The tool is based on regular expression processing and can handle 4 classes of chronological patterns, namely, “century range”, “century”, “year range” and “year/date”. Users can create many regular expression patterns for each class in order to capture as many chronological formats as possible. Each class has specific parameters that must be set. Common parameters for all classes are the regular expression template and the matching position for the numeric value in the regular expression (e.g the number 500 of the temporal value “500 AD”). Range classes (“century range” and “year range”) accept two numeric matching positions. The

regular expression template of a pattern can include custom and predefined placeholders that are associated with lists of keywords in many languages. Placeholders eliminate the number of different patterns needed for each class.

Custom placeholders have no practical impact in the time extraction algorithm and are used to allow users to easily accumulate alternative terms. These terms however may affect whether a time value is matched by the particular pattern.

Predefined placeholders are fixed for each pattern class and affect the actual time extraction algorithm. They have multiple descriptors that are associated with specific keywords in chronological patterns. For example the "century identifier" predefined placeholder is used by the extraction algorithm to capture a more specific year range within the century and applies to patterns of the "century range" and "century" classes. It has two descriptors, the "early" and the "late". For a specific chronological pattern of one of these classes, the "early" descriptor may have keywords like "early", "beginning of" and the Greek counterparts.

For each pattern class there is a dedicated time extraction algorithm. An algorithm for a pattern class takes as input a pattern (along with all associated parameters) and a temporal value. The algorithm first checks whether the temporal value is matched against the "unfolded" regular expression of the pattern, i.e. the regular expression that occurs after all placeholders are substituted with the associated keywords. If so, it outputs a year (for class "year/date") or a year range (for classes "century range", "century", "year range").

Let's suppose a pattern named "early Xth century" of the class "century" that can handle century values such as "early 6th c. BCE", "first quarter of the 2nd c. AD" and "αρχές 5ου αι. π.Χ." All parameters of the pattern are shown in Table 1.

Table 1. Parameters for pattern "early Xth century" of class "century"

pattern name	early Xth century		
pattern class	Century		
regular expression template	\[?(**#century_identifier**)(.*\s)?(**\d{1,4}**)\s?**#s0**?(\s)?(**#bc_ad**(\s*)?)?(\s**#s1**\.		
numerical matching pos	4		
predefined placeholders	**#century_identifier**	**early**	early, first quarter of, beginning of, α ρ χ ές, late
		late	late, end of, τ ε λ ος
	#bc_ad	**BC**	bce, bc, b.c.e., b.c., b.c, π.χ., π.χ , π χ
		AD	μ.χ., μ.χ , μχ , ad, a.d., a.d, ce, c.e., c.e
custom placeholders	**#s0**		st, nd, rd, th, ος, ο υ
	#s1		century, cent., c., α ι ώ ν ας, α ι .

The numerical matching position parameter is 4 matching the part "\d{1,4}". The "#century_identifier" and "#bc_ad" placeholders are the predefined placeholders for class "century". For the former we set keywords "early", "first quarter of", "beginning of" and "αρχές" for descriptor "early". For the descriptor "late" we ser keywords "late", "end of", and "τελος". The "#bc_ad" predefined placeholder is used by the extraction algorithm to define whether the temporal value refers to a BC or an AD century (resulting in a negative or positive year number). It has two descriptors, the "BC" and "AD". We set keywords "bce", "bc", "b.c.e.", "b.c.", "b.c", "π.χ.", "π.χ" and "πχ" in the "BC" descriptor and keywords "ad", "a.d.", "a.d", "ce", "c.e.", "c.e", "μ.χ.", "μ.χ" and "μχ" in the "AD" descriptor . The "#s1" is a custom placeholder that accumulates alternative keywords for the term century: "century", "cent.", "c,", "αιώνας", "αι." etc.

Some examples of temporal values that are successfully matched by this pattern and can be normalized by the corresponding extraction algorithm are depicted below:

early 6th c. BCE → **-600/-571**

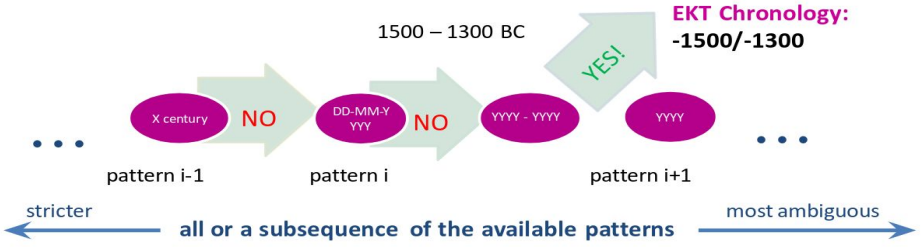
first quarter of the 2nd c. AD → **100/130**

end of the 12th cent. → **1171/1200**

αρχές 5ου αι. π.Χ. → **-500/-471**

The chronological patterns must be arranged in a specific order, from the stricter to the most ambiguous. When a chronological value is to be normalized, it passes through an ordered list (sequence) of chronological patterns, until the first match is found. Based on that pattern, the normalized year or year range is calculated. Fig.6 illustrates an example where a temporal value "1500-1300 BC" is eventually matched by pattern i+1 named "YYYY-YYYY" of the "year range" pattern class.

We created 30 different chronological patterns[4] in order to capture the different time formats we come across with in the aggregated content. Table 2 shows some typical normalization examples per pattern class.



[4] Available at https://www.searchculture.gr/aggregator/resource/docs/Chronological_Patterns.pdf

Fig. 6. Date value “1500 - 1300 BC” is normalized to “-1500/-1300” by pattern “i+1” with label “YYYY-YYYY” that handles year ranges.

Table 2. Normailization of chronologies using regular expression matching

Chronological Pattern Class	Examples
century range	2nd half of 5th c. BC until 4th c. BC → **-450/-301**
century	18th century → **1701/1800** early 18th century → **1700/1730** first half of 5th c. BC → **-500/-451**
year range	1342/48 → **1342/1348** 1342 - 1654 → **1342/1654** 500 BC - 400 BC → **-500/-400**
date/year	526 BC → **-526** 198; → **1980/1989** 11/03/2000 → **2000**

6 The semantic enrichment scheme used in SearchCulture.gr

We enriched the aggregated content with terms from a vocabulary of cultural item types, with homogenized chronological values (years or year intervals) and with terms from a vocabulary of greek historical periods. Both vocabularies were created with specific assumptions to facilitate the enrichment process. Metadata records are enhanced with three separate *EKT fields*. Note that the original documentation is not modified and is normally indexed and searchable.

6.1 The enrichment strategy for item types

We enriched and homogenized the aggregated content of SearchCulture.gr using a vocabulary of item types that we created and published in Semantics.gr[5]. The vocabulary is hierarchical (Is-a hierarchy) and bilingual (Greek and English) consisting of 159 distinctive terms. The schema of the vocabulary conforms to SKOS (skos:Concept owl class). Each term, apart from the different labels (skos:prefLabel, skos:altLabel), has references to broader and narrower terms (skos:broader,

[5]http://www.semantics.gr/authorities/vocabularies/ekt-item-types/vocabulary-entries/tree

skos:narrower) from the vocabulary and also links to Getty AAT[6] (via skos:exactMatch) and DBPedia[7].

Metadata records were enriched with a separate field *EKT type* that holds references to the vocabulary. The type enrichment of a collection involves the following actions: i) examination of the documentation quality of the collection to decide whether a secondary field is needed or not ii) registration of the repository in Semantics.gr iii) creation of EMRs for the collection iv) ingestion (or re-indexing) of the collection in the aggregator in order for the actual enrichment to take place. Depending on the collection, the enrichment is based on original values of "dc:type" and, for special cases, of "dc:subject" or "dc:title". Table 3 summarizes three different documentation qualities, namely Type-A, Type-B and Type-C, and the respective mapping methodology.

Table 3. Documentations classes & type enrichment methodologies.

Class	Documentation quality class description	Methodology
Type-A	Good documentation of dc:type.	**EMR**: primary field
Type-B	Insufficient documentation on dc:type for the entire collection or part of it, useful dc:subject	**EMR**: primary and **secondary** fields
Type-C	Insufficient documentation on dc:type for the entire collection or part of it, useful dc:title/dc:description	**EMR**: primary field and **descriptive secondary** field

We will demonstrate the mapping process with the following example. Suppose that an aggregator-institution wishes to enrich its collections with references to a SKOS vocabulary (*V*) previously published in Semantics.gr. Vocabulary *V* contains the following 5 entries:

➔ http://scs.gr/sculpture **skos:prefLabel** "Sculpture"@en | "Γλυπτό"@el
  ➔ http://scs.gr/figurine **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

For a *Type-A* collection, the curator initializes a new EMR form in the enrichment tool where he/she sets the metadata field dc:type as the primary field and chooses *V* as the target vocabulary. Then, the enrichment tool harvests metadata records from the repository and creates a list of distinct dc:type values with their cardinalities (1st column of Table 4).

[6]The Getty Art & Architecture Thesaurus, http://www.getty.edu/research/tools/vocabularies/aat/

[7] https://wiki.dbpedia.org/

Table 4. EMR for a Type-A collection

dc:type value	**Entry from vocabulary *V1***	
sculpture art *(120 items)*	**http://scs.gr/sculpture**	**auto**
greek vases *(230 items)*	**http://scs.gr/vase**	**auto**
jewelleries *(135 items*)	**http://scs.gr/jewellery**	**auto**
amphora *(100 items)*	**http://scs.gr/vase**	**manual**
oenochoe *(12 items)*	**http://scs.gr/vase**	**manual**
earring *(13 items)*	**http://scs.gr/jewellery**	**manual**

Next, the curator triggers the auto-suggestion functionality which successfully maps 3 distinct dc:type values to the correct vocabulary entries. The curator assigns the correct vocabulary term for the three remaining values manually. He/she bookmarks these three EMRs so as to be taken into account in future mapping suggestions. This creates three keyword values in vocabulary V, "amphora", "oenochoe" and "earring":

➔ http://scs.gr/sculpture **skos:prefLabel** "Sculpture"@en | "Γλυπτό"@el
  ➔ http://scs.gr/figurine **skos:prefLabel** "Figurine"@en | "Ειδώλιο"@el
➔ 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
    keyword: ["amphora", "oenochoe"]

Finally, the curator confirms the EMRs and the mapping phase is completed. In Table 4, label "auto" indicates that the EMRs was automatically created.

Type-B collection has insufficient documentation of the primary field (either for all or for some of the items) but has another metadata field (secondary) that can contribute in the enrichment process. An example is shown in Table 5. Focus on the first mapping rule for dc:type value "ceramic objects": a metadata record with this dc:type value will be enriched with the reference http://scs.gr/vase only if it has one of the following dc:subject filters: "vase" or "amphora" or with the reference http://scs.gr/figurine if it has a dc:subject value "figurine". The auto-suggestion mechanism can easily set this EMR as long as there are vocabulary matches for these filters. Items with dc:type value "exhibits" will be enriched with http://scs.gr/vase if they have a dc:subject "amphora" but they do NOT have a dc:subject "earing" (suppose that an image shows an earring shaped as an amphora).

Table 5. EMR for a Type-B collection

dc:type	Filters (dc:subject)	Entry from vocabulary *V1*	
ceramic objects *(101 items)*	amphora , vase, statuette …	**http://scs.gr/vase**	**auto**
		if filter in [“vase”, “amphora”]	**auto**
		http://scs.gr/figurine	**auto**
		if filter in [“statuette”]	**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”] & **NOT** in [“earing”]	**manual**

In a *Type-C* collection, the documentation of dc:type is very poor for some items, but its dc:title values may contain useful words or phrases. The enrichment tool will search all titles against a set of words that form the *search term space*. It is derived from all skos:prefLabel and skos:altLabel values of V (Sculpture, Γλυπτό, Figurine, Ειδώλιο, Jewellery, Κόσμημα, Vessel, Σκεύος, Vase, Αγγείο) as well as the keywords from previous bookmarked EMRs (earring, amphora, oenochoe). The tool will set only the matching words as available filters for each dc:type value. The rest of the mapping process is identical with the one described for *Type-B* collections.

6.2 The enrichment strategy for chronologies and historical periods

We enriched the aggregated content with homogenized (normalized) chronologies and with historical periods using a hierarchical bilingual vocabulary of Greek historical periods. Metadata records are enriched with two separated fields, *EKT chronology* and *EKT historical period*.

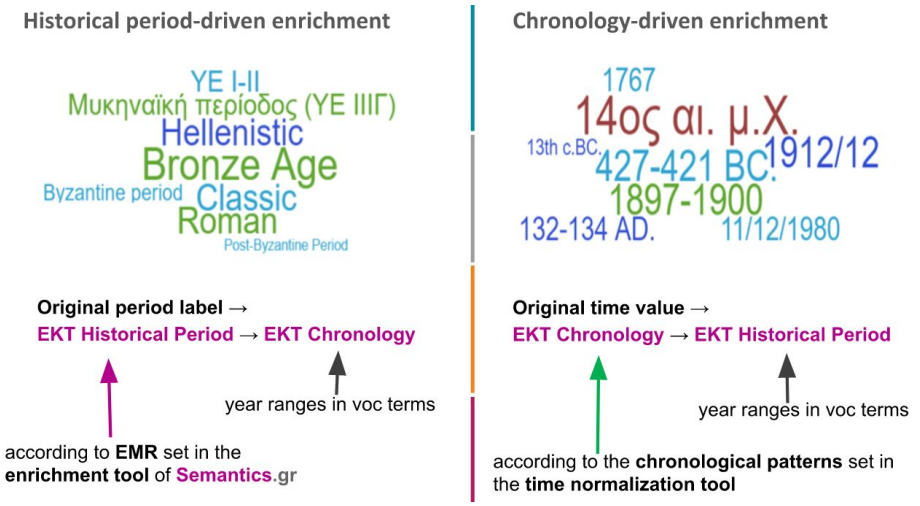


Fig. 7. The two strategies used for temporal enrichment that both lead to the same result.

Depending on whether the original temporal documentation is based on period labels or chronologies, we adopted two fundamentally different enrichment strategies, *historical period-driven enrichment* and *chronology-driven enrichment*, respectively, as illustrated in Figure 7. The former involves setting EMRs in the enrichment tool of Semantics.gr, similarly to the enrichment procedure for types; items originally described with period labels, are mapped to vocabulary terms but now they are also enriched with the respective year ranges. In the chronology-driven enrichment, chronological values are being homogenized into years or year ranges and then, based on the results, the items are enhanced with the corresponding terms from the historical periods vocabulary. The enrichment is based on the original values of a temporal field ("dc:date", "dc:created" or "dcterms:temporal", depending on the collection) and in special cases taking into account keywords in descriptive field values, such as of "dc:description" and "dc:title".

Table 6 summarizes 4 typical collection types, namely *Temp-A*, *Temp-B*, *Temp-C* and *Temp-D*, their qualitative characteristics with respect to temporal documentation and the enrichment methodologies used. The methodologies used for *Temp-A* and *Temp-B* fall into the historical period-driven enrichment strategy, while the methodology used for *Temp-C* falls into the chronology-driven one. *Temp-D* collections are handled using both strategies: items described with period labels are handled as *Temp-A* or *Temp-B* and items described with chronologies are handled as *Temp-C*. We use regular expressions to distinguish chronologies from period labels.

Table 6. Documentations classes & temporal coverage enrichment methodologies.

Class	Documentation quality class description	Methodology
Temp-A	Temporal field (dcterms:temporal) with period labels (e.g. "archaic era")	1) **EMR**, primary field (EKT historical period) 2) extract year span from voc term (EKT chronology)
Temp-B	Insufficient documentation on temporal field for part or all the collection, useful titles (e.g. "archaic vase", "sculpture from the hellenistic period")	1) **EMR**, primary field, descriptive secondary field (EKT historical period) 2) extract year span (EKT chronology)
Temp-C	Temporal field (dc:date, dcterms:created or dcterms:temporal) with chronologies (e.g. "1981", "late 12th c. AD", "1100-1200 AD")	1) **Normalization of chronology values** (EKT chronology) 2) Enrich with EKT historical period
Temp-D	Temporal field (dcterms:temporal) with some values containing historical periods and others containing chronologies	Items with chronological values are handled as Temp-C and the remaining as Temp-A or Temp-B

The vocabulary of Greek Historical Periods

We created a Greek historical periods' vocabulary that ranges from 8,000 BC (Mesolithic Period) to present and we published it in Semantics.gr[8]. It is hierarchical and bilingual (Greek and English) consisting of 94 distinctive terms. The schema of the vocabulary conforms to the edm:Timespan contextual class introduced by Europeana [4]. For each term, apart from the different labels (skos:prefLabel, skos:altLabel) the year range is also defined in properties edm:begin and edm:end. The vocabulary is linked to DBpedia (via skos:relatedMatch or skos:exactMatch properties).

We created the thesaurus taking into consideration reputable sources about Greek history as well as established vocabularies such as Getty AAT. Some periods have a strict local scope (e.g. minoan, cycladic and helladic periods) and as a result their year ranges tend to overlap. We call those periods *relative* and marked them accordingly in a special administrative field. The rest of the periods cover the entirety of Hellenic territory and are less debatable with respect to their timespans. We call those *absolute*. In our vocabulary, absolute periods have neither overlaps nor gaps when they have the same parent and relative periods have at least one absolute ancestor. A simplified part of the historical period vocabulary with both absolute and relative periods is presented below:

[8] http://www.Semantics.gr/authorities/vocabularies/historical-periods/vocabulary-entries/tree

➔ http://scs.gr/bronze | ABSOLUTE
skos:prefLabel "Bronze Age"@en | "Εποχή του Χαλκού"@el
edm:begin -3200 **edm:end** -1050
➔ http://scs.gr/cycladic | RELATIVE
skos:prefLabel "Cycladic Period"@en | "Κυκλαδική Περίοδος"@el
edm:begin -3300 **edm:end** -1100
➔ http://scs.gr/helladic | RELATIVE
skos:prefLabel "Helladic Period"@en | "Ελλαδική Περίοδος"@el
edm:begin -3300 **edm:end** -1000
➔ http://scs.gr/minoan | RELATIVE
skos:prefLabel "Minoan Period@en | "Μινωικη Περίοδος"@el
edm:begin -3200 **edm:end** -970
...

➔ http://scs.gr/archaic | ABSOLUTE
skos:prefLabel "Archaic Period"@en | "Αρχαϊκή περίοδος"@el
edm:begin -700 **edm:end** -480
➔ http://scs.gr/early_archaic | ABSOLUTE
skos:prefLabel "Early Archaic"@en | "Πρώιμη Αρχαϊκή"@el
edm:begin -700 **edm:end** -575
➔ http://scs.gr/early_archaic | ABSOLUTE
skos:prefLabel "Middle Archaic"@en | "Μέση Αρχαϊκή"@el
edm:begin -575 **edm:end** -535
➔ http://scs.gr/late_archaic | ABSOLUTE
skos:prefLabel "Late Archaic"@en | "Υστερη Αρχαϊκή"@el
edm:begin -535 **edm:end** -480

Table 7. Enrichment steps for a Temp-A collection

dcterms:temporal	Step 1: EMR - primary field (EKT historical period)	Step 2: extract year span (EKT chronology)
Post-Byzantine Period	→ Ottoman Period	→1453/1821
Middle - Late Hellenistic Years	→ Middle Hellenistic Period - Late Hellenistic Period	→-220/-31

Historical period-driven enrichment

The aggregator enriches items originally described with period labels with the mapped historical periods from the vocabulary (step 1) and computes their chronologies according to the assigned periods (step 2). The temporal enrichment of a collection involves the following actions: i) registration of the repository in Semantics.gr, ii) creation of the EMRs for the time field of the collection in the enrichment tool, as described in Section 3.2 (only for the period label values; if there are chronological values as well, these are automatically ignored by the mapping tool using regular expression filtering) iii) enabling historical period-driven enrichment for the particular collection in the aggregator which includes setting the metadata field

used as primary field in the EMRs iv) ingestion (or re-indexing) of the collection in the aggregator in order for the actual enrichment to take place. Tables 7 and 8 illustrate examples of Temp-A and Temp-B collections. The result of each enrichment step is presented for each item.

Table 8. Enrichment steps for a Temp-B collection

Dc:title (secondary field)	Step 1: EMR -primary & descriptive secondary field (EKT period)	Step 2: extract year span (EKT chronology)
Archaic vase	→ Archaic Period	→ -700/-480
Hellenistic sculpture	→ Hellenistic Period	→ -323/-31

Chronology-driven enrichment

The effectiveness of chronology-driven enrichment is heavily based on the normalization of chronological values. The aggregator normalises the original chronologies (step 1) and based on the year or year ranges, chooses the corresponding absolute historical periods from the vocabulary (step 2). The chronology-driven temporal enrichment of a collection involves the following actions: i) enabling chronology-driven temporal enrichment for the particular collection in the aggregator which includes setting the metadata field that contain chronological values and associating with a subsequence of the available chronological patterns iii) ingestion (or re-indexing) of the collection in the aggregator in order for the actual enrichment to take place.

Table 9. Enrichment steps for a Temp-C collection

dc:date, dctemrs:created or dcterms:temporal	Step 1: Normalize chronologies (EKT chronology)	Step 2:Enrich with corresponding period (EKT historical period)
Late 5th century	→ 471/500	→ Early Byzantine Period
7th c. B.C-mid 6th c. BC	→ -700/-551	→ Early Archaic - Middle Archaic Period
03/11/1980	→ 1980	→ Regime change

Note that for step 2, we ignore the relative historical periods, since we did not always have spatial information [9] to assign a correct relative period given a normalized chronology. For example, Middle Bronze Age is an absolute period, covering the timespan 2000-1580 BC. It includes Middle Minoan, Middle Cycladic and Middle Helladic periods which are marked as relative since they refer to different civilisations that flourished in different territories. Therefore, an item dated in 1700 BC (-1700 in normalized form) will be assigned with the Middle Bronze Age term.

The chronological-driven enrichment is completely automated since there is no need for creating EMRs. Table 9 illustrates an example of a Temp-C collection.

7 Leveraging semantic enrichment to improve portal's search and browsing functionality

We stored EKT type, EKT historical period and EKT chronology fields as dc:type, dcterms:temporal and dc:date properties respectively in a separated ore:Proxy object of the internal-EDM model that represents EKT's perspective on the items and wee indexed them using several indexed fields in Apache Solr of various types.

7.1 Searching, browsing and faceting by type

The visual representation of EKT type field (as shown in item pages of SearchCulture.gr) consists of one or more types (e.g. "Figurine, Souvenir"). However, in order to support advanced hierarchical searching and faceting on types that capture the semantics of *is-a* relationships (browder-narrower) between types, we index for each item its browder types as well, using a separate auxiliary Solr field. This way, for example, when a user searches vessels the results will also include vases.

Leveraging the Apache Solr search platform and our indexing scheme, we enhanced SearchCulture.gr with new multilingual search and browsing functionalities that improve discoverability including searching by type using a controlled hierarchical list of values, hierarchical navigation on all types through a separated page, hierarchical faceting on types and an interactive tag cloud (Fig. 8).

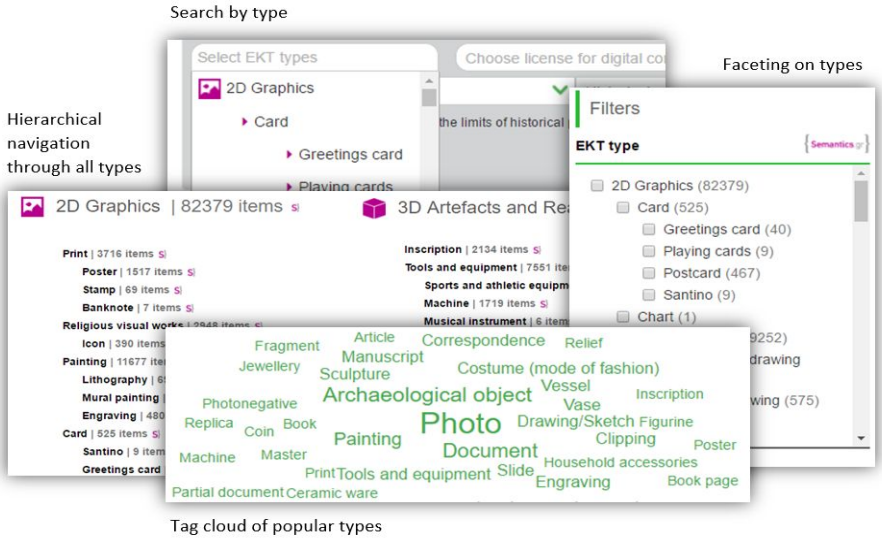


Fig. 8. Bilingual search, filtering (facets) and navigation (browsing) on types

7.2 Searching, browsing and faceting by chronologies and historical periods

For the EKT chronology field, we used the Date Range Field of Apache Solr which supports time interval indexing, time range queries and interval facets[9]. The Date Range Field supports effectively and efficiently year range queries and facets (e.g 1700 - 1950) on indexed year intervals.

Regarding the EKT historical period field, its visual representation consists of either one (e.g. "Hellenistic Period") or two – in case of period intervals – historical periods (e.g. "Middle Archaic Period – Late Hellenistic Period"). In order for the search engine to support advanced hierarchical searching and faceting on periods that capture the semantics of *part-of* relationships between periods, we index intervening periods (e.g. those between the upper and lower bounds of a period interval) as well as both super (ancestor) and sub (descendant) periods in separated auxiliary Solr fields.

Users can choose between two search modes both for year range-based and for historical period-based search, the "loose" one and the "strict" one.

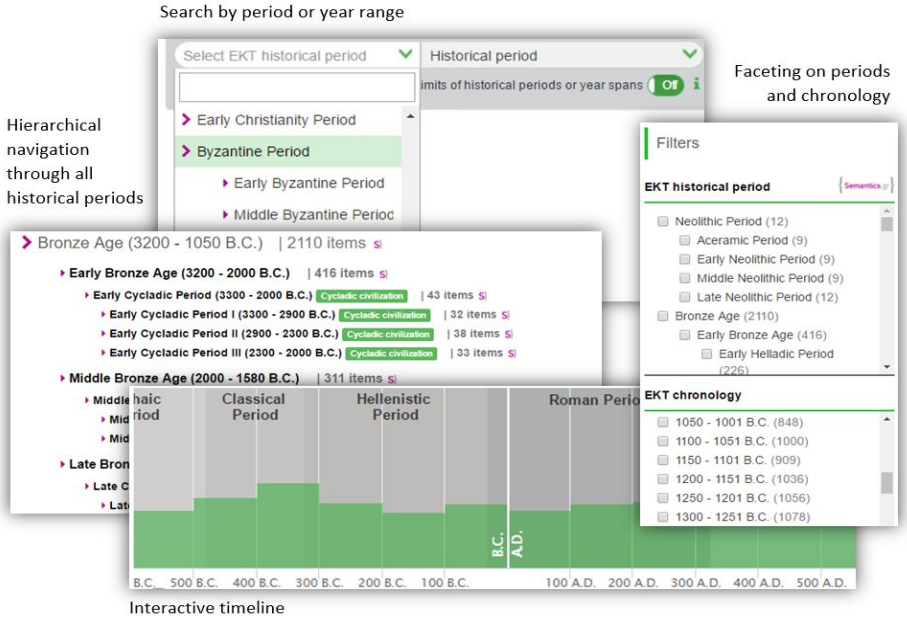


Fig. 9. Bilingual search, filtering (facets) and navigation (browsing) on periods and chronology

[9] https://cwiki.apache.org/confluence/display/solr/Working+with+Dates

In the “loose” mode, which is the default, temporal search returns items with a year or period interval that *intersects* that of the search criterion. For example, for a search criterion: “1500-1600 AD”, an item dated “1550-1750 AD” will appear in the results. Similarly, for a search criterion “Classical Period”, an item dated “From Classical to Hellenistic period” will also appear in the results.

In the “strict” mode, temporal search is more precise bringing only items with a year or period interval strictly *within* (or coinciding) the one defined by the search criterion. For example, for a search year range: “1500-1600 AD”, an item dated “1550-1750 AD” will not be included in the results, while an item with “1550-1570 AD” date will. Similarly, for a search criterion “Classical Period”, an item dated “From Classical to Hellenistic period” will not be included in the results, while an item dated “Early Classical Period” will. The “strict” mode is very useful when the user wants to find items dated exclusively within as specific year or period interval.

We achieved the supporting of the “loose” and “strict” modes for EKT chronologies by using the native capability of the Date Range Field of Apache Solr (relational predicates “Intersects” and “Within”, respectively). For EKT historical periods we support the two modes by using different auxiliary Solr indexed fields. The “loose” mode uses an indexed field that stores intervening, super and sub periods while the “strict” mode uses an indexed field that stores only intervening and their super periods.

Leveraging the Apache Solr search platform and our indexing scheme, we enhanced SearchCulture.gr with advanced time-driven multilingual search and browsing functionalities including searching by historical period via a hierarchical list of values, searching by year or year range, hierarchical navigation through all historical periods, faceting on year-ranges (intervals of 50 years) and historical periods and an interactive timeline/histogram (Fig. 9).

7.3 Combined queries

The user can easily submit complex queries such as "middle-Byzantine coins" (Fig. 10-a), "postwar paintings" (Fig. 10-b) "Sculptures dating from 600 to 500 BC" (Fig. 10-c) using the controlled criterium “EKT Type” combined with the controlled criterium “EKT Historical Period” or the free text criterium “Year Span”. Of course these queries can be combined with keyword-based searching in various fields, for example “Sculptures dating from 600 to 500 BC having the word ‘Apollo’ in the title metadata field”. The user can also choose a type, for example "vase," and then, using the facets, navigate through historical or chronological periods, exploring this way the evolution of pottery art in Greek culture.

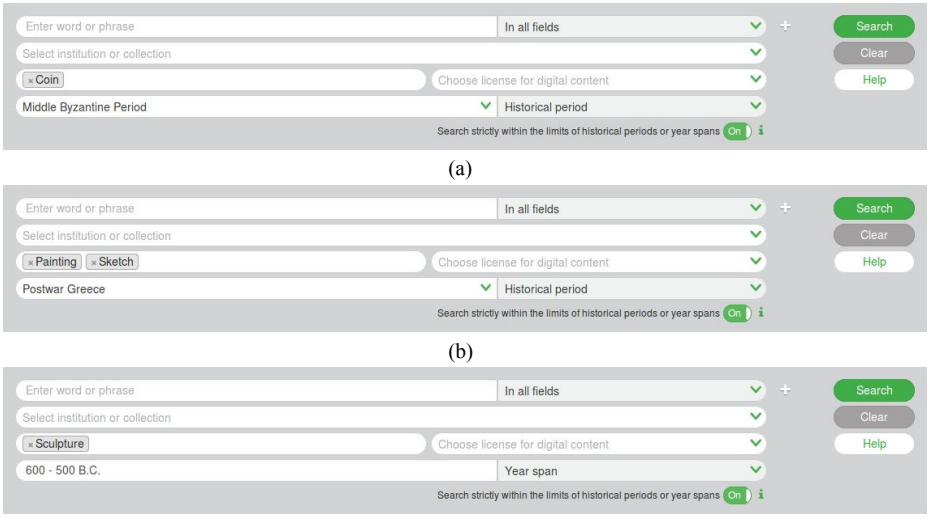


Fig. 10. Combined queries by types, historical periods and year spans

7.4 Publishing content as Linked Data

SearchCulture.gr publishes the aggregated content as Linked Data. Resources on SearchCulture.gr have unique, permanent HTTP URIs so that they can be referred to by users and applications. The HTTP URIs return the description of the resources as HTML or as RDF (structured as EDM) either on XML (rdf/xml) or json (json-ld) serialization. Thanks to the semantic enrichment, cultural heritage items retain links to the Vocabulary of Item Types of EKT - which in turn links to the Getty Art & Architecture Thesaurus (AAT) and the semantic thesaurus DBpedia of Wikipedia - and to the Vocabulary of Greek Historical Periods, linked to DBpedia. There are also links that were included in the original metadata by the content providers (ie GeoNames).

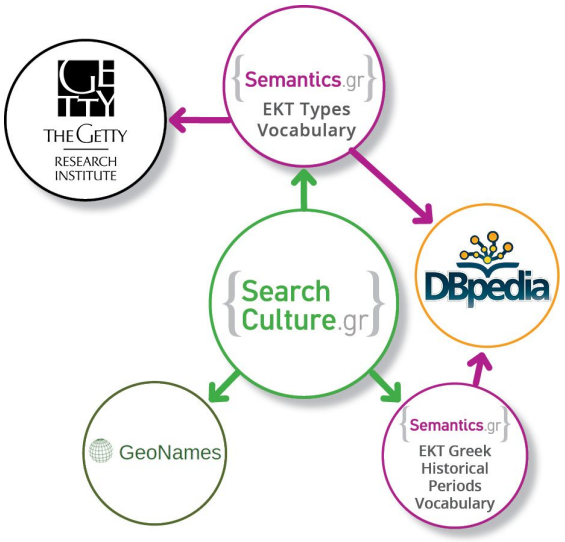


Fig. 11. Semantic links to other data - vocabularies and thesauri- also offered as Linked Data

8 Enriching the content of SearchCulture.gr – The results

More than 395K items of SearchCulture.gr - 91% of the content - were classified into a compact and balanced set of 146 types. Table 10 illustrates the number of collections and enriched items per documentation class (Sec. 3.2).

Table 10. Collections and number of items per type documentation class

Documentation Class		# of collections	# of items
Type-A:	sufficient existing dc:type values	30	287128
Type-B:	insufficient dc:type values – useful dc:subject	29	74224
Type-C:	insufficient dc:type – resorting to dc:title values	7	71519
Total		**66**	**432871**

The enrichment improved remarkably the searchability of the content as illustrated in the experiment shown in Fig. 12 where we compared the number of search results returned by SearchCulture.gr for 17 popular types in Greek before and after the enrichment. Before the enrichment, for each type we run two searches, a general keyword search (in all indexed metadata) and a keyword search restricted on the original dc:type metadata field. After the enrichment, we used the EKT type search criterium that allows users to pick a type from a hierarchical list of values. Comparing the number of results of the EKT type-based search and the general keyword search

before the enrichment, we can see that the former returns significantly more results in all but three cases. For these three cases we should take into account that the general keyword search is far from accurate for searching for specific types because it may return a large number of false positives. For example, when searching using the keyword "ενδυμασία" ("costume" in english) in all indexed metadata fields, the results may also return "book" items that happen to have dc:description values that include the word "ενδυμασία". Comparing the number of results of the EKT type-based search and the keyword search on the original dc:type field before the enrichment, we can see that the former returns significantly more results in all queries. We repeated the experiment, this time using the same search keys in English, as shown in Fig. 13. Since the majority of the items were documented only in Greek, the improvement was even more impressive.

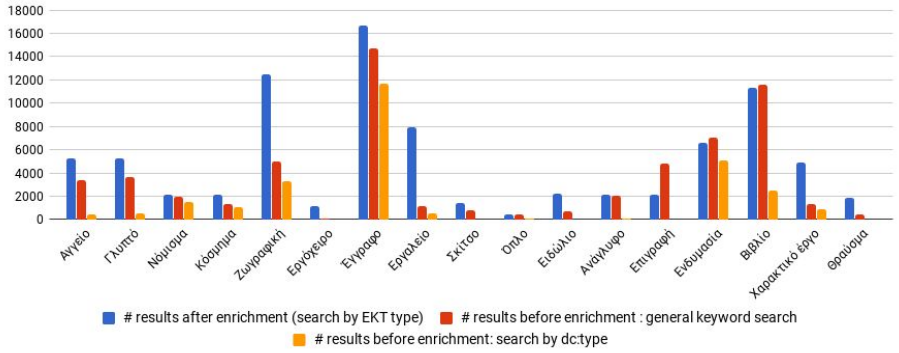


Fig. 12. Type enrichment: improve in searchability of 17 types in Greek

A total of 312,000 items of SearchCulture.gr - the 72% of the aggregated content - were enriched with normalized chronologies and assigned with historical periods. Note that 129,258 items did not have any explicit temporal information, however, we managed to enrich 8,387 of them by identifying keywords in their titles or other descriptive fields. Table 11 illustrates the number of collections per documentation class as introduced in Sec. 5.1 and the total of enriched items per class.

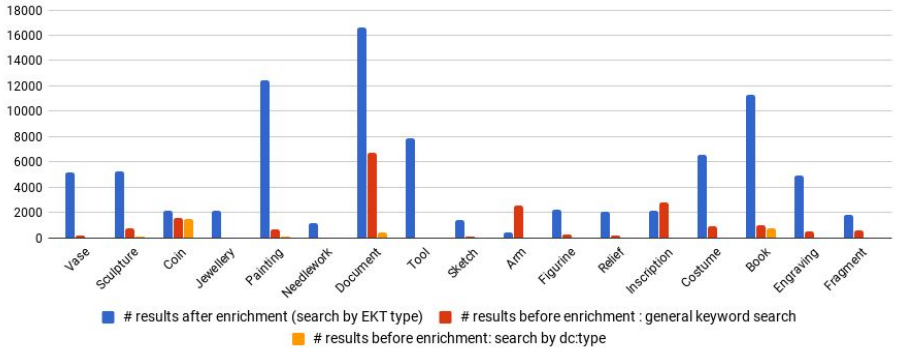


Fig. 13. Type enrichment and multilingualism: improve in searchability of 17 types in English

Table 11. Collection and number of items per temporal documentation class

Documentation Class		# of collections	# of items
Temp-A	Temporal field (usually dcterms:temporal) with period labels	4	6870
Temp-B	Partly insufficient temporal field, useful titles or descriptions	3	6646
Temp-C	Temporal field with chronologies (usually dc:date, dcterms:created)	56	286810
Temp-D	Mixed values: periods (based on dcterms:temporal, dc:title or dc:description) and chronologies (Temp-A, B and C)	2	11674
Total		**65**	**312000**

The temporal enrichment improved further the discoverability of the content as illustrated in the experiment shown in Fig. 14 where we compared the number of search results returned by SearchCulture.gr for 10 greek historical periods in Greek before and after the enrichment. Before the enrichment, for each period we run a general keyword search (in all indexed metadata). After the enrichment, we used the EKT historical period search criterium that allows users to pick periods from a hierarchical list of values. Comparing the number of results of the EKT period-based search and the keyword search before the enrichment, we can see that the former returns significantly more results in all but one queries. For this case we should once again take into consideration that the general keyword search is very fuzzy when searching by periods due to the large number of false positives. For example, when searching using the keyword “Ρωμαϊκή Περίοδος” (“Roman Period” in english) in all

indexed metadata fields, the results may also return “book” items dated in recent years that happen to have dc:description values that include the word “Ρωμαϊκή”.

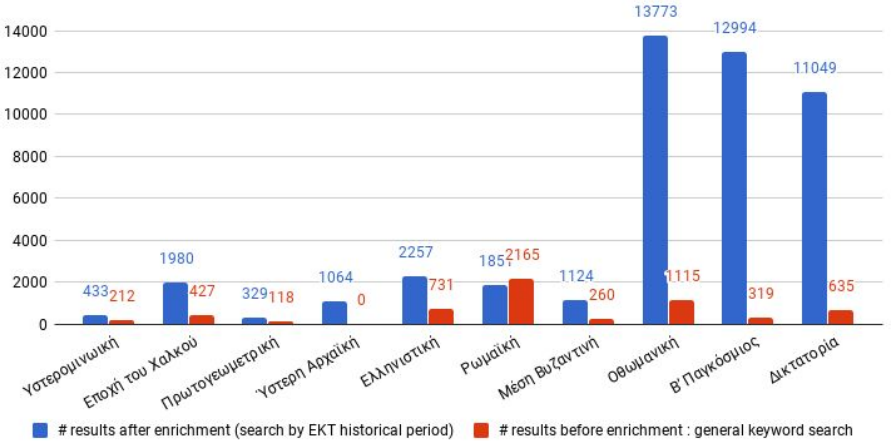


Fig. 14. Temporal enrichment: improve in searchability for 10 periods in Greek

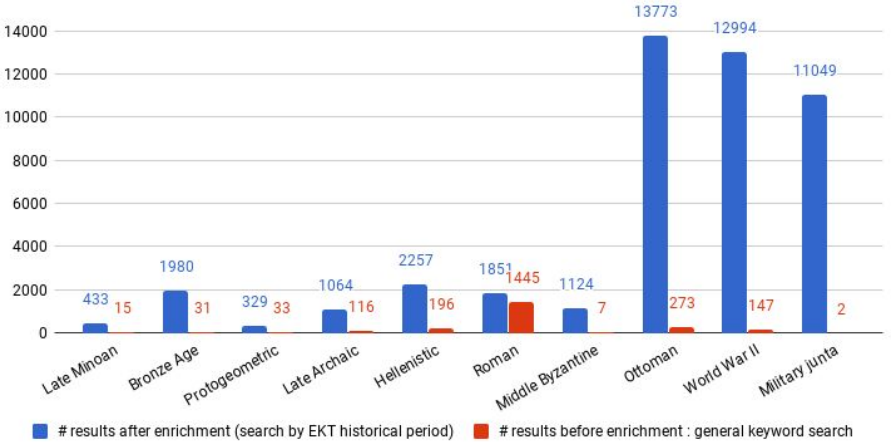


Fig. 15. Temporal enrichment and multilingualism: improve in searchability of 10 periods in English

We repeated the experiment, this time using the same historical periods in English, as shown in Fig. 15. As expected, since the majority of the items were documented only in Greek, the improvement was even more drastic.

Our future plans focus on extending the enrichment scheme in order to deal with spatial information and subjects. This will allow the enhancement of SearchCulture.gr with new features such as map-based navigation as well as searching, browsing and

faceting on thematic categories. Moreover, the multi-dimensional semantic enrichment will facilitate the creation of thematic exhibitions and “similar object” functionality.

Acknowledgments. The work presented in this article has been partly supported by the project "Platform for provision of services for deposit, management and dissemination of Open Public Data and Digital Content" of the Operational Programme "Digital Convergence" (NSFR), co-funded by Greece and the European Union. We would like to thank Dimitra Pelekanou, the graphic designer of SearchCulture.gr, Elena Lagoudi, digital engagement specialist, as well as Manolis Megalooikonomou and George Zachariadis from EKT Systems and Networks Unit.

References

[1] Georgiadis, H., Banos, V., Stathopoulou, I.O., Stathopoulos, P., Houssos, N., Sachini, E.: Ensuring the quality and interoperability of open cultural digital content: IISA: 178-183 (2014)

[2] Gavrilis, D. and Ioannides, M. and Theofanous, E.: Cultural Heritage Content Re-Use: An Aggregators's Point of View, ISPRS. II-5/W3: 83-87 (2015)

[3] Stiller, J., Petras, V., Gäde, M., Isaac, A.: Automatic Enrichments with Controlled Vocabularies in Europeana: Challenges and Consequences. EuroMed: 238-247 (2014)

[4] Europeana Data Model Primer, available at http://pro.europeana.eu/page/edm-documentation

[5] Georgiadis, H, Papanoti, A.:Semantics.gr: Information system for vocabularies & semantic enrichment. Technical report, EKT(2016)

[6] Manguinhas H. et al. Exploring Comparative Evaluation of Semantic Enrichment Tools for Cultural Heritage Metadata. TPDL 2016. Springer (2016)

[7] Georgiadis, H., Papanoti, A., et al: Semantics.gr: A self-improving service to repositories and aggregators for massively enriching their content. In Proc. of DHC Workshop of MTSR 2016

[8] Agirre, E., Barrena, A., Lopez de Lacalle, O., Soroa A., Fernando S., Stevenson, M.: Matching Cultural Heritage tems to Wikipedia. In: Proc. LREC 2012. Istanbul, Turkey. (2012)

[9] Rabinowitz, A.:It’s about time: historical periodization and Linked Ancient World Data. ISAW Papers 7.22 (2014)

[10] Doerr, M., A. Kritsotaki , S. Stead: Which period is it? A methodology to create thesauri of historical periods. Proceedings of CAA2004 (2004)

[11] MINT. (2013). MINT Metadata Interoperability Services. http://mint.image.ece.ntua.gr/

[12] LoCloud enrichment, http://www.locloud.eu/Resources/LoCloud-enrichment-services

[13] Coudyzer, E. et al: The Terminology Management Platform: a Tool for Creating Linked Open Data. TOTh Workshop 2014

[14] V. de Boer, M. W. van Someren and B.J. Wielinga. Extracting Historical Time Periods from the Web. Journal of the American Society for Information Science and Technology (JASIST). 2010 DOI 10.1002/asi.21378

[15] Georgiadis, H., Papanoti, A., et al: The Semantic Enrichment Strategy for Types, Chronologies and Historical Periods in SearchCulture.gr. In Proceedings of MTSR 2017, Tallinn, Estonia (2017)

[16] Charles, V., Freire, N., Antoine, I.: Links, languages and semantics: linked data approaches in The European Library and Europeana. In Linked Data in Libraries: Let's make it happen!, IFLA 2014, Satellite Meeting on Linked Data in Libraries (2014).

[17] Manguinhas H., Freire N., Isaac A. et al. Exploring Comparative Evaluation of Semantic Enrichment Tools for Cultural Heritage Metadata. TPDL 2016, Lecture Notes in Computer Science, vol 9819. Springer, Cham

[18] Peroni, S., Tomasi F., Vitali F.: The aggregation of heterogeneous metadata in Web-based cultural heritage collections. A case study. International journal of Web Engineering and Technology, Vol. 8 (4), pp 412-432 (2013)

[19] Garoufallou, E., and Papatheodorou, C.: A critical introduction to Metadata for e-Science and e-Research, International Journal of Metadata, Semantics and Ontologies ((IJMSO), 9(1), pp.1–4. http://dx.doi.org/10.1504/IJMSO.2014.059143

[20] Greenberg, J. and Garoufallou, E.: Change and a Future for Metadata. In Proceedings of the 7th Metadata and Semantic Research (MTSR), 2013

[21] Henning S. and Federica F. : Europeana Publishin Guide: A guide to the metadata and content requirements for data partners publishing their collections in Europeana. Europeana, 2018, available at https://pro.europeana.eu/post/publication-policy

[22] Matienzo, M. A., Rudersdorf, A.:: The Digital Public Library of America Ingestion Ecosystem: Lessons Learned After One Year of Large-Scale Collaborative Metadata Aggregation. Dublin Core Conference 2014: 12-23

[23] Evens, T. and Hauttekeete, L.: Challenges of digital preservation for cultural heritage institutions. Journal of Librarianship and Information Science, Vol 43, Issue 3, pp. 157 - 165 (2011)