# Language-Independent Access to History Textbooks - Utilizing GIS Functionality with Spatiotemporal and Thematic Metadata

Dieter Pfoser[1,3], Alexandros Efentakis[1], Thanasis Hadzilacos[1,4], Sophia Karagiorgou[1], Giorgos Vasiliou[2]

[1] RA Computer Technology Institute
University Campus Patras, 26500 Rion, Greece
{pfoser|efedakis|thh|karagior}@cti.gr

[2] Talent SA
Karytsi Square 4A, 10561 Athens, Greece
vasiliou@talent.gr

[3] RC "Athena", Institute for the Management of Information Systems
Bakou 17, Athens 11524, Greece
pfoser@imis.athena-innovation.gr

[4] Open University of Cyprus
13-15, Digeni Akrita Avenue
1055 Nicosia, Cyprus
thh@ouc.ac.cy

**Keywords:** content integration, spatiotemporal indexing, Multilanguage content, history content

**Corresponding author:**
*Dieter Pfoser*
Institute for the Management of Information Systems - IMIS/ATHENA
G. Mpakou 17, 11524 Athens, Greece
Work: +30 210 6990 522 (204)
email: pfoser@imis.athena-innovation.gr

## Abstract

Integrating and accessing structured textual content obtain from different sources is a challenging task and becomes even more so when dealing with multiple languages. The objective of this work is to showcase the technological efforts towards the creation of a digital European history textbook repository that integrates respective textbooks from various countries and publishers. The content integration is achieved by introducing language independent metadata based on space (locations), time (dates), and thematic categories (history gazetteer). Providing adequate interfaces such metadata can be used to provide language-independent access to Multilanguage history textbook content. The specific focus in this work will be on (i) presenting the metadata, (ii) the data management approach including indexing the history textbook content and (iii) the resulting textbook repository including its GIS-based interface allowing for a combination of map, timeline and keyword based search of the history content.

# 1 Introduction

The great number of languages in Europe (more than 20 official ones in the European Union alone) makes it from an individual's point of view often very difficult to identify relevant documents to one's search without adequate language skills. In the specific context of this work and the project it relates to, the focus is on providing means *to provide unified access to Multilanguage history content in a language-independent way*.

A *digital European history textbook base* will integrate and provide access to already existing digital material from various European publishers. The integration is achieved by introducing largely *language independent metadata* including (i) *space* (locations), (ii) *time* (dates), and (iii) *history concepts*. These three metadata aspects represent essential abstractions of the content and will be used to index and access the textbook content in a largely language-independent way. The aim is to use the history concepts to define a (near) language-independent (abstractions that can be easier translated than the entire texts) thematic categorization of content. In addition, to identify locations in arbitrary textbooks, a multilingual dataset that includes historic place names is needed. Finally, temporal identifiers are used to create a timeline for the respective text portions.

While in theory, the potential as an indexing means is substantial, a considerable challenge poses the *identification, collection and integration of said metadata*. This work focuses on creating a thematic ontology for history concepts and a multilingual location corpus. The former will contain thematic categories used in country-specific curricula to arrive at a history ontology covering all of Europe. The location corpus will include place name identifiers that can be used to geographically reference text portions. Integration of metadata sources entails also the translation of each source to several target languages. To make this process as efficient as possible, an metadata translation tool is created that utilizes the Wikipedia encyclopedia. Subsequent manual translation and verification provide for an overall semi-automatic translation tool.

The three metadata aspects will be used to *tag the historic textbooks*, i.e., relate metadata entries to the respective text portions. Essentially, all textbook content and metadata is stored by means of a relational database with the tagging software scanning the textbook content and relating it to the metadata. As a result a Multilanguage index based on temporal, spatial and thematic metadata is created that *provides language-independent access* to the textbooks, i.e., searching or content using Greek metadata will produce results also in other languages. Automatic translation tools will then be used to translate the content into the target language, e.g., Spanish text into Greek.

The work presented in this paper touches several research aspects that all belong to the larger complex of digital library research. Foremost, the spatial and temporal metadata-based access to content is been exploited in several projects and products. The Centennia Historical Atlas (Centennia (2010)) supports spatiotemporal exploration of history by means of interactive maps and displayed relevant content. In a similar fashion, HyperHistory Online (HyperHistory (2010)) allows one to explore selected history concepts by means of a simple temporal categorization and image maps of concept visualizations.

Work in the general area of Named Entity Recognition is manifold having resulted already in many research projects and products, many of them open-source. GATE (Cunningham et al. (2002)) is a general framework for information extraction tasks from content. In our work, we will use this tool to perform the essential content tagging task, i.e., automatically relating metadata to content. MinorThird (Cohen (2004)) is an open-source toolkit

(collection of Java classes) for storing text, annotating text, and learning to extract entities and to categorize text. As such its functionality is limited with respect to GATE but would fit the purposes of our project, i.e., relating metadata to content in terms of annotations. However, a severe shortcoming of MinorThird is that it does no support Unicode character encoding. Clearforest Gnosis (ClearForest(2010)) is a browser extension that performs categorization of terms for Web pages by annotating (color coding) them. Gnosis is a free service of Clearforest, which provides commercial solutions for text-driven business intelligence, i.e., text categorization and text analysis. A similar product suite is LexiQuest (SPSS (2010)) from SPSS providing text analysis functionality such as concept and relationship discovery combined with visualization interfaces for the results. The tool supports several languages including German, Italian and Spanish. LexiQuest has been evaluated and was found to provide limited flexibility with respect to storage of, both, metadata and content and provided no flexibility with respect to customizing the tagging solution. A limitation with most tools is that they focus on English as a working language and typically lack Unicode support. Further, available metadata for NER is limited and linked to supported languages.

In addition to general text mining tools, specialized approaches for geocoding content are available. Metacarta (Metacarta (2010) offers a tool for the geocoding of Web pages and texts in general. The tool recognizes geographic key words and relates them to coordinates. In connection with a spatial content browser, e.g., Google Earth (Google (2010)), this technology can be used as an additional powerful means to improve the quality of keyword-based search results. In the scope of this project, we will develop similar techniques that have a narrower content focus and are based on a general NER approach.

The remainder of this work is organized as follows. Section 2 describes the data scenario. Section 3 details how a spatiotemporal-thematic index is created from respective metadata for history textbooks. Section 4 describes the CITER tool for searching and accessing content. Finally, Section 5 provides conclusions and directions for future research.

## 2 The Data

To better understand the motivation behind our work of using three distinguished metadata aspects to index content, the following sections briefly describe the nature of history textbook content and subsequently introduce the used metadata in more detail.

### 2.1 History Textbook Content

Digital educational content and constructive educational software are two key “instruments” towards the essential introduction of new technologies in the learning process and the achievement of the anticipated goals. For the history discipline, the multilingual and multicultural digital content gains in particular importance as it allows comparative approaches to the discovery and interpretation of the historical facts and the development of a broader historic perspective by considering the existing multiple national views in a comparative teaching approach.

The specific content used in this work comprises

- *55 history textbooks*
- from a total of *eight publishers*

- from *seven countries*,
- representing *six languages* (English, Spanish, German, Italian, Greek and Slovenian).

The type of content will be mostly history textbooks available in digital form, but also include content from history CDs and picture libraries.

An important step towards the creation of the repository is the integration of content sources by means of an efficient storage mechanism. We achieved this by (i) defining an integrated data model and (ii) a methodology for transforming the textbooks according to this model and entering them in relational database (PostgreSQL DBMS in our case). While all "textbooks" are available in digital form, an elaborate process was designed to import books prepared using desktop-publishing software such as Adobe InDesign and Quark Xpress in the production process into the database. Here, a semi-automatic methodology was defined that includes manipulating the original content sources, i.e., improving the quality and correcting the original DTP files, translating them to (structured) XML documents and finally importing them to the database.

### 2.2 Metadata

Identifying and authoring proper metadata and indexing content is a next-to-impossible task when making no restrictions on the scope of the content. Focusing on the subset of history textbooks one can exploit specific metadata aspects of the data that present adequate abstractions of the content and provide a simple indexing means.

- *Space* (locations),
- *time* (dates) and
- *thematic metadata* (history concepts)

are used to structure history content and to provide a general means for (largely) *language independent access* and integration of content, i.e., comparatively little effort is needed to translate the metadata. In our endeavor involving content in six different languages, we use a combined approach involving manual as well as machine-based metadata collection. Sources of metadata are (i) already existing metadata for textbooks (indices, TOC), (ii) existing metadata collections (spatial feature name collections, datasets from publishers, ontologies), and (iii) text mining tools that extract relevant concepts from text analysis. The overall objective is to integrate all three sources to produce a comprehensive metadata base.

The two fundamental aspects for the study of historical facts are *space and time*. These two aspects have the advantage over classical semantic markups such as keywords in that they are non-ambiguous, i.e., there is no doubt as to what an $x$ and $y$ coordinate or a date mean! An important prerequisite to spatial metadata authoring is the identification of geographic feature names. Although simple at first sight, this task poses a challenge in that also Multilanguage issues and historical naming have to be considered. Thus, we have to include Multilanguage mappings of locations as well as historical to current mappings of locations. All these variations have to be captured in the geographic feature name corpus to afterwards identify place names in texts. Existing datasets for feature names that include geographic co-ordinates are the Geographic Names System (GNS) (GEONet (2010)) and the Getty Thesaurus of Geographic Names (TGN) (Getty Trust (2010)), with the latter comprising around 900k places and related information. Place names may include names in the vernacular language, English, other languages, historical names, names and in natural order and inverted order. Besides these content sources, we will evaluate and add any dataset that is available from the partners in this project.

Given the fact that large metadata sources do exist, identifying appropriate spatial metadata is simple when compared to the task of building a *history ontology* for thematic metadata. Accessing content based on location and time might not yield the desired results when trying to find information related to specific topics that cut across time and space. An example here would be "age of enlightenment," which affected all European countries in different ways. Hence, a thematic component is needed to better discriminate and index history content. For the definition of a history ontology, we followed a combined top-down/bottom-up approach. The CITER history ontology has been defined by a historian and history textbook author. The history ontology consists of 300 classes, including three main categories (event, organization, and person), a qualitative category (source), a time class, a location class and an attribute class (properties to additionally style instances). The maximum depth for each of the main classes is three (e.g., three levels below the "event" class). At the same time, concepts (instances) for this ontology are collected in a bottom-up fashion, i.e., collecting and integrating existing textbook metadata as well as using text analysis techniques. Finally, the instances are related to the respective classes of the ontology. Hence, while in a top-down fashion candidate entries of the history ontology are defined by domain experts such as historians, machine translation techniques and existing metadata sources provide concepts at the instance level. Figure 1 illustrates this approach.

### 2.3 Metadata Translation - or - the Wikipedia Trick

Space, time and thematic metadata were chosen since the represent (i) a powerful index for history information and (ii) need comparatively little effort to be translated. However, this translation task for metadata is still a considerable challenge and our goal was to provide automatic and high-quality means of translation as briefly outlined in the following.

Each metadata entry represents a specific "label" in its respective language. Any automatic translation would require and be based on dedicated Multilanguage resources, i.e., dictionaries. In case no dictionary entry is available, a manual translation is needed. One such dictionary that is publicly available is Wikipedia (Wikipedia Foundation (2010). In the present approach, the Wikipedia encyclopedia is viewed as a multi-language resource based on concepts. Each Wikipedia article page has a column that contains links to articles covering the same topic in another language. Figure 2 gives an example of an English article page on "World War II", where one can follow the "in other languages" links, to find respective articles in, among others, German, Italian, Spanish, Greek, and Slovenian. What is of interest is not the actual content of the article but its "label." The label is a human translation, i.e., by the article authors, of the respective concept into the various languages. Exploiting this fact, a Wikipedia translation tool was created that takes a label as input and tries to translate it into the various target languages. Two versions of the tool are publicly available (Efentakis (2010)), Java code for database-centered batch processing as well as an online interface.

Given the availability of the Wikipedia tool, the following procedure was adopted to translate all metadata from and to the six target languages (cf. Figure 3). In Step 1 all potential metadata is collected from respective sources. In Step 2, all concepts are translated to English, which is adopted as a language of reference. Step 2a involves an automatic translation approach and Step 2b a complementary translation step and verification of the automatic translation results. Step 3 involves then the translation into the respective target languages, with again, Step 3a involving automatic translation and Step 3b manual translation and verification of the automatic translation results.

Using English as a reference language has the advantage of improving the results in Step 3a using the Wikipedia tool.

Using the Wikipedia translation tool and translating the metadata into all six languages, the complete amount of available metadata is shown in Table 1. The columns represent (i) the number of original metadata entries in their respective language as collected from the various sources, (ii) the number of automatically translated terms (Wikipedia tool) and (iii) the final number of metadata entries available in all languages. Essentially, using the Wikipedia tool, roughly half of all entries have been translated automatically.

# 3 Spatiotemporal-Thematic Index

While creating spatial and thematic metadata involves a lot of work, tagging the actual content with this metadata is even more so challenging. Such a process is generally labeled Named Entity Recognition (NER), i.e., assigning a (group of) words to a set of predefined categories. In our specific context it would mean to discover temporal, spatial and thematic identifiers that can be linked to a timeline, locations recorded as spatial metadata, and statements referring to concepts in the history ontology.

## 3.1 Software Framework

When faced with a large amount of content, the tagging content with metadata has to be automated to the largest extent possible. Information extraction (IE) systems automate tagging process and can be integrated as software infrastructure. In the specific context, the software that is used is GATE (A General Architecture for Text Engineering) (Cunningham (2002)), a software framework for natural language processing and engineering. In the context of this work, we assume that terms with a common *stem* will usually have similar meanings. The performance of an information retrieval system will be improved if term groups are conflated into a single term, e.g., “connect”, “connected”, “connecting”, “connection”, and “connections” to the stem “connect” (Porter (1980)). In the specific context, we used Snowball (Porter (2010), a framework for writing stemming algorithms. Out of our six target languages, stemmers for English, German, Spanish and Italian were readily available. We developed a Greek stemmer based on Ntais (2006). A Slovenian stemmer has been developed based on Popovic (1992). The Snowball framework is available as a plugin for GATE.

**Spatial and Thematic Tagging**

For tagging content with metadata and considering the natural language processing aspect, a workbench was created integrating GATE and specific Snowball stemming functionality. Our approach of text processing is (i) word level oriented, uses (ii) stop-words (articles, conjunctions, etc.) and (iii) stemming. In our framework, the textbook content and the metadata are stored by means of a relational DBMS, in our case PostgreSQL. This is in contrast to the original GATE implementation that uses files. The Content Tagger is used to relate terms found in the content to terms contained in the metadata.

The principal approach for tagging content (history textbooks) is illustrated in Figure 4. The Content Tagger is used to relate terms found in the content to terms contained in the metadata. The overall content tagging approach can be grouped in the following phases. Phase 1, tokenization, represents a pre-processing step to extract candidate

words from text by removing stop words and punctuation and applying stemming. Stemming is applied to metadata as well. In Phase 2, the tagging process, content is loaded from the relational database in main memory. Each document is represented by a sorted vector of tokens. To identify tokens in the content (among complex words with dashes and punctuation) a Unicode tokenizer is used to account for the variety of texts and languages. With metadata also loaded from the database into a main memory data structure, in a first pass, for each token in the vector of tokens, a candidate list of metadata entries is created. In essence, the tagging process itself is language independent and word level oriented, i.e., initially we do not search for complete phrases but independent words (tokens). Consider the following example. For a token "republic" that is found in the text, all metadata entries that contain the word "republic" are identified. In a second pass, tokens prior and past "republic" are used to form candidate phrases. One such example would be "Weimar republic." Is any such phrase contained in the candidate list of metadata entries for "republic", the respective sequence is tagged. In our case "Weimar republic" would be such a phrase.

The tagging result is stored in tables as relationships between metadata and content. Figure 5 shows another tagging result. The concept "Second World War" was found in various texts. The visualization of the result directly reflects the result table. Text portions that were tagged are highlighted. Terms with the same number of words are highlighted using the same color. The highlighting is primarily used as a debugging tool to verify tagging results and improve the algorithm.

**Temporal Tagging.**

Temporal tagging, i.e., the discovery of temporal identifiers requires a different approach since a priori temporal metadata is limited! A temporal identifier in the respective languages is composed of (i) numbers that may represent everything dates, month, years and centuries, (ii) metadata comprising labels for weekdays, month, seasons, decades, centuries, etc. and (iii) the respective rules to from form valid temporal identifiers, e.g., dates. A parser was developed to detect date, month and year numbers from the textual representation of a temporal phrase. The implementation is based on Glowacki (2010). Figure 6 gives some temporal identifiers that were detected in English, German and Greek texts.

### 3.2 Resulting Spatiotemporal-Thematic Index

To assess the quality of the resulting spatiotemporal-thematic index, Table 2 shows tagging results for some exemplary books covering topics of the 20th century. The results given are in that order Words – number of words in book with stop words excluded, Found all – number of concepts found measured in terms of number of words, Found % - number of concepts found in terms of percentage of total words, and Found dist. – distinct number of concepts found.

Given the prominent examples of German and Slovenian textbooks it is shown that a significant part of the content is essentially related to metadata. In the case of only considering thematic metadata, 15% of the content can be linked to metadata, while when also considering spatial and temporal metadata *the actual words in the content that are linked to metadata increases to close to 30%*!

**Metadata Analysis, Content Overlap.**

A more detailed metadata analysis was conducted in order to realize the *semantic overlap between books*, i.e., given a specific metadata entry how often and in how many books does it appear. The analysis focuses on the same textbooks also given in Table 2, i.e., including German, English, Italian, Slovenian and Greek. All books focus on the 20th century, i.e., an a-priori overlap is given. What needs to be established is whether this overlap can be also found by examining the overlap in metadata as well.

Figure 7 shows the overlap among the discovered metadata concepts for sets of textbooks. Consider the example of thematic metadata found for the German textbook in Figure 7(a), which is 921. This textbook shares 460 terms with the Italian book and 116 with the Greek book. The overlap of all three books is 91 terms. The overall of all five textbooks results in 35 terms (center of Figure 7(a)). Similar figures are shown for geographic metadata. Temporal metadata differ in that here temporal overlaps are considered. For example if one books mentions March 1945 and the other March 25, 1945, this is considered an overlap. The temporal overlap is largely similar to the numbers reported for spatial metadata.

**Overall Content and Index Size.**

Did the previous section show some specific facts on the respective coverage of the index, so will we in the following give the overall statistics concerning the index and its coverage.

*History Textbook Content***.** The 55 history textbooks comprise a database consisting of

- a total of 82800 paragraphs, which
- constitute 8.5 million words, or
- an average paragraph size of 103 words.

*Metadata*. The collected metadata comprises a total of

- 3690 thematic metadata entries
- 3215 geographic metadata entries

The temporal metadata is individually discovered, i.e., the number of hits correspond to the actual size of the metadata. Temporal metadata is stored by means of time, i.e., temporal queries such as range are supported.

*Spatiotemporal-thematic index*. An important measure for our spatiotemporal-thematic index is its size. The following numbers give the total number of hits per thematic category, i.e., number of occurrences in content.

- Total thematic hits: 757,536
- Total spatial hits: 195,546
- Total temporal hits: 159,161

It can be observed that while spatial and temporal metadata produce roughly the same number of hits, the thematic metadata produces 4 times as many hits. The reason is simply that thematic metadata was derived from the content itself using indices, TOCs and text mining tools. The spatial metadata was derived from external sources and was then actually "discovered" in the content. As such the number of hits is astounding.

Consequently, the total number of paragraphs that have been indexed is 75949 out of total 82791, corresponding to a coverage of the index of 92%. This means that at least 92% of all paragraphs have been indexed at least once. Elaborating further on the index characteristics, it can be observed that for

- the *thematic metadata*, *on average* each entry has
    - 223 hits, i.e., references to content

- spread over 6.7 history textbooks
- and 2.63 languages (out of 6).

- the *spatial metadata*, on average each entry has
  - 142 hits, i.e., references to content
  - spread over 5.7 history textbooks
  - and 2.6 languages (out of 6).

Again, considering the fact that spatial metadata was simple compiled using external sources, the actually hit rate when compared to thematic metadata is very good.

**Summary**

Given the right metadata, a combined spatiotemporal-thematic index for history content becomes feasible. The metadata gazetteer includes roughly 10000 concepts. Using such a gazetteer to tag history content, results in up to 30% of the content (i.e., the actual words not considering stop words) being related to metadata! The textbooks further exhibit a considerable overlap with respect to the metadata they share. This overlap is important when we consider the objective of the metadata, namely to find content sources in Multilanguage textbooks related to the same thematic, geographic and temporal concept/metadata.

## 4 Accessing History Content – the Modified GIS Case

Tagging content with *spatial, temporal and thematic metadata creates a formidable index to access history textbook content*. This section describes the CITER platform and interface specifically supporting content search based on the three distinguished metadata aspects.

The CITER platform is a client-server application resembling a somewhat peculiar GIS desktop application. The application is based on and uses the Cruiser platform (Talent (2010)). The application has to be installed on a Windows XP/Vista computer with continuous and preferably broadband Internet connection. As such, the application is similar to, e.g., Google Earth. All the CITER platform data (books, photographs, indices etc.) are stored on a server. On the client side (desktop app), apart from the application, map data is installed to achieve fast response time and to avoid overloading network traffic. In this guide we will have a look at the most useful and exciting platform features.

Figure 8 showcases the basic user interfaces to query the history textbooks. Properly indexed content based on space, time and thematic metadata is accessed using the three respective querying mechanisms. The *temporal parameter* is fixed using a temporal slider that allows for the selection of time points as well as ranges. The *spatial scope* is determined by a spatial range and thus selected landmarks and alternatively also based on keywords, e.g., Greece. This functionality is supported using a base map of Europe to show important geographical landmarks the user can query. A next step in the application will be to integrate historic maps (vector data with changing boundaries in time). The *thematic range* is determined by browsing/querying keywords in the respective language, thus effectively identifying concepts in the history ontology.

*Query results* are displayed on the right side using an embedded Web-browser (see below for examples and explanation). The retrieved content is showed in various tabs, each representing the content in its original form or a

respective machine translation. Each type of content is marked as such. The software used for the translation of the history content to the various target languages is the ESTEAM Translator (ESTEAM (2010)). The interface allows temporal, spatial and thematic parameters to be used simultaneously to query the textbook content. For example, given the spatial range of Greece, a temporal range of the year 1923 and the thematic concept of "war", the respective metadata entries in the database are queried to follow their pointers to the respective text portions they index. In this case, texts relating to the Greek-Turkish war will be retrieved.

In the following, the various query parameters are described in more detail.

**Spatial parameter**

The user can add spatial parameters to the query (cf. Figure 9), thus indicating that he/she wants to have results that are only related to the places that have been selected. Spatial data is selected by dragging a place from the map and dropping it to the reserved area. Range queries for selection are supported.

**Temporal parameter**

The timeline (cf. Figure 10) is an active component in the application that can zoom or pan to select a respective timespan (point). It has a highlighted area, which can be moved, shrunk or expanded fitting the query needs. This highlighted area restricts the query results only to those that refer to the time-period represented by the area to retrieve content referenced by the temporal index.

**Adding thematic parameters**

Thematic terms (history concepts) (cf. Figure 7) can be added by opening a dialog and allowing the user to select one or many entities either by browsing the history ontology or by using (keyword) search.

**Results**

The results are shown in a tree-like structure, in the panel at the bottom (cf. Figure 12). The results are divided into main texts (and photos) and sources of the texts. The results can be browsed freely and the user can see a paragraph (cf. Figure 13), or even a whole section (chapter) containing the paragraphs that match the query (cf. Figure 14).

## 5 Conclusions and Future Work

Content integration involving multilingual sources poses an interesting challenge, especially when coupled with the ambition of language-independent access. This work demonstrates that given the right content, specific metadata can be identified that (i) provides a sufficiently accurate abstraction of the content and can be used as an index to provide for sufficiently accurate search and (ii) at the same time is either language independent or requires a comparatively small effort to be translated.

In our case of history textbooks, such *metadata includes time, space and thematic concepts*. This work showcased (i) the data – history textbook content, (ii) spatial, temporal, and thematic metadata resulting in a multilingual history ontology and spatial metadata gazetteer and (iii) the resulting largely language-independent spatiotemporal-thematic index. This index covers the content adequately and also shows sufficient overlap between books and languages.

Directions for future work are manifold. As was expected for the geocoding of content, i.e., performing NER for geographic features, disambiguation of geographic terms is needed to avoid false hits. This task is challenging in that we have to deal with a multitude of languages, i.e., taking into account context and grammar at the individual language level. The objective however should be to provide a language independent approach for the disambiguation of such terms. An interesting approach when deriving metadata is automatic ontology creation (Ogata (2004)). Here, texts related to the same topic are analyzed to automatically extract classes and relationships to construct a domain ontology. While the history ontology has been created with the specific purpose of indexing content in mind and a large number of instances were available a priori, such an approach can be used to verify the current ontology structure as well as to produce additional instances. Using GATE as the basic software framework has led to the using Snowball-based stemming algorithms. However, promising work in the area of language independent stemming (Bacchin (2005)) needs to be evaluated in the present context as it can significantly reduce the effort needed for adding new languages to the history textbook repository. Besides technological improvements, an important effort will be to collect additional metadata to improve the coverage of the repository index. In the context of content tagging, we want to experiment with fuzzy matching of concepts to texts. Besides stemming, which relaxes the matching words, we want to experiment with relaxing the word sequences of phrases themselves. For example, instead of trying to identify the metadata concept "Treaty of Versailles" in the text, we would search for occurrences of such a word sequence within some proximity, thus also linking the text "the peace treaty, which was signed in Versailles, France" to the above concept. Finally, we want to develop this project into a Wiki-based annotation system that utilizes the history textbooks as core content and provides a means for students, teachers, and other interested third parties to re-structure content so as to develop individualized teaching materials as well as additions to the content itself.

## Acknowledgments

This work is partially supported by the CITER project (Creation of a European History Textbook Repository – http://citer.cti.gr) funded by European Commission, eContentplus programme, grant agreement number ECP-2005-EDU-038193. The authors would like to thank the project partners for their contributions to this work.

## References


Bacchin M, Ferro N, and Melucci M 2005 A probabilistic model for stemmer generation. *Information Processing and Management,* 41(1): 121-137

Centennia Software 2010 The Centennia Atlas. Web document, http://www.clockwk.com

ClearForest 2010 ClearForest Gnosis. Web document http://www.opencalais.com/gnosis/

Cohen W 2004 Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data, Web document http://minorthird.sourceforge.net

Cunningham H, Maynard D, Bontcheva K and Tablan V 2002 GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02)

Efentakis A and Pfoser D 2008 Wikipedia Translation Tool. Web document http://citer.cti.gr/the-project/wikipedia-phrase-translation-tool/

ESTeam AB 2010 Automatic Translation Solutions. Web document http://www.esteam.gr

Getty Trust, The J. Paul 2010 Getty Thesaurus of Geographic Names. Web document http://www.getty.edu/research/conducting_research/vocabularies/tgn/index.html

Glowacki D 2010 CalendarParser, a Java object class. Web document http://icecube.wisc.edu/~dglo/software/calparse/index.html

Google 2010 Google Earth. Web document http://earth.google.com.

NGA 2010 GEONet Names Server. Web document http://earth-info.nga.mil/gns/html/index.html

Metacarta 2010 MetaCarta Platform. Web document http://www.metacarta.com.

Ntais G 2006 Development of a Stemmer for the Greek Language. *Master Thesis, Department of Computer and System Sciences, Royal Institute of Technology, Sweden*. Online available at http://www.dsv.su.se/~hercules/papers/Ntais_greek_stemmer_thesis_final.pdf

Ogata N and Collier N 2004 Ontology Express: Statistical and Non-Monotonic Learning of Domain Ontologies from Text. In *Proceedings ECAI-2004 Workshop on Ontology Learning and Population*

Popovic M and Willett P 1992 The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5): 384-390

Porter M F 1980 An algorithm for suffix stripping. *Program*, 14(3): 130-137.

Porter M F 2001 Snowball: A Language for Stemming Algorithms. Web document http://snowball.tartarus.org/texts/introduction.html

SPSS 2010 LexiQuest Mine. Web document http://www.spss.com/lexiquest/lexiquest_mine.htm

Talent 2010 Cruiser platform. Web document http://www.cruiser.gr

Wikipedia Foundation 2010 Wikipedia – The Free Encyclopedia. Web document http://www.wikipedia.org

World History Online 2010 HyperHistory Online. Web document http://www.hyperhistory.com/

**Table 1: Concepts after translation**

| | **Concepts incl. Organizations** | | | **Places** | | |
|---|---|---|---|---|---|---|
| | Orig. | After WP Transl. | After Manual Transl. | Orig. | After WP Transl. | After Manual Transl. |
| German | 3123 | 3317 | 3690 | 473 | 1313 | 3215 |
| Slovenian | 611 | 733 | 3690 | 352 | 678 | 3215 |
| Italian | 191 | 1592 | 3690 | 327 | 1094 | 3215 |
| English | 327 | 3690 | 3690 | 1789 | 3215 | 3215 |
| Spanish | | 1451 | 3690 | 148 | 488 | 3215 |
| Greek | | 464 | 3690 | 206 | 654 | 3215 |
| Total | 4253 | 11247 | 3690 | 3294 | 7443 | 3215 |

**Table 2: Thematic index coverage**

| Content | Lang. | Words | Found all | Found % | Found dist. |
|---|---|---|---|---|---|
| **Concepts and Organizations** | | | | | |
| Geschichte und Geschehen 4 | DE | 38651 | 5986 | **15.5%** | 921 |
| Britain - World War II | EN | 14794 | 1739 | **11.8%** | 185 |
| OMNIA | IT | 2714250 | 80332 | **3.0%** | 1188 |
| 20. stoletje | SLO | 24784 | 5102 | **20.6%** | 597 |
| Ιστορία ΣΤ Δημοτικού | GR | 12833 | 1373 | **10.7%** | 223 |

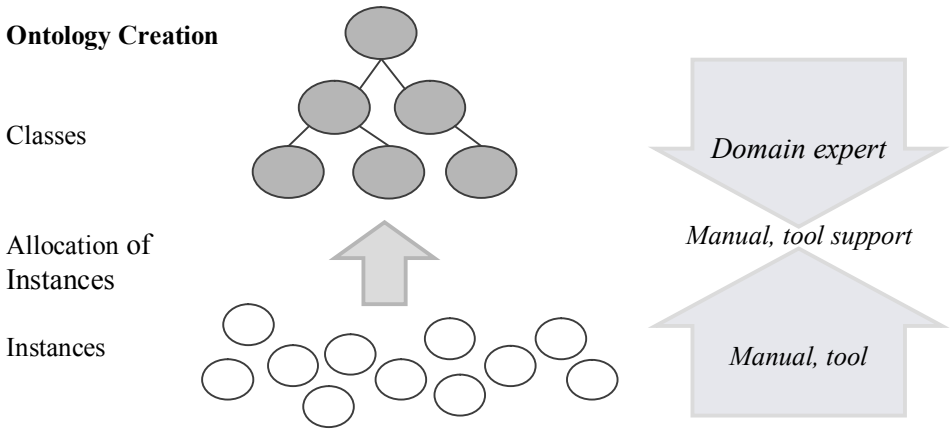



**Figure 1: History ontology creation overview**

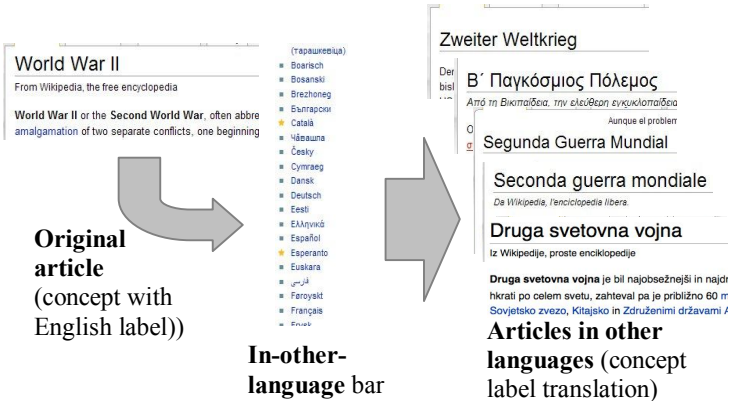



**Figure 2: Wikipedia "term" translation**

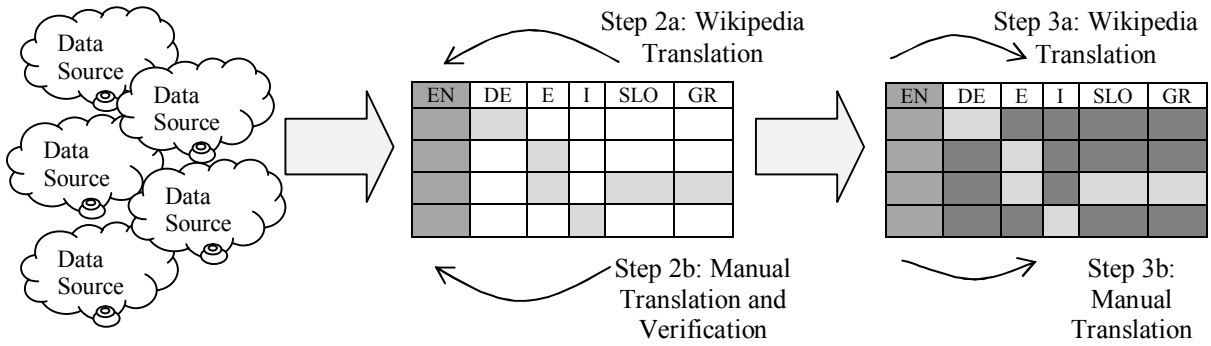
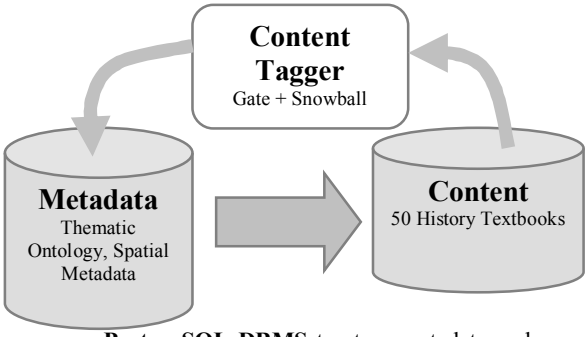



**Figure 3: Metadata translation process**




**PostgreSQL DBMS** to store metadata and

**Figure 4: Content Tagging: linking metadata and content**

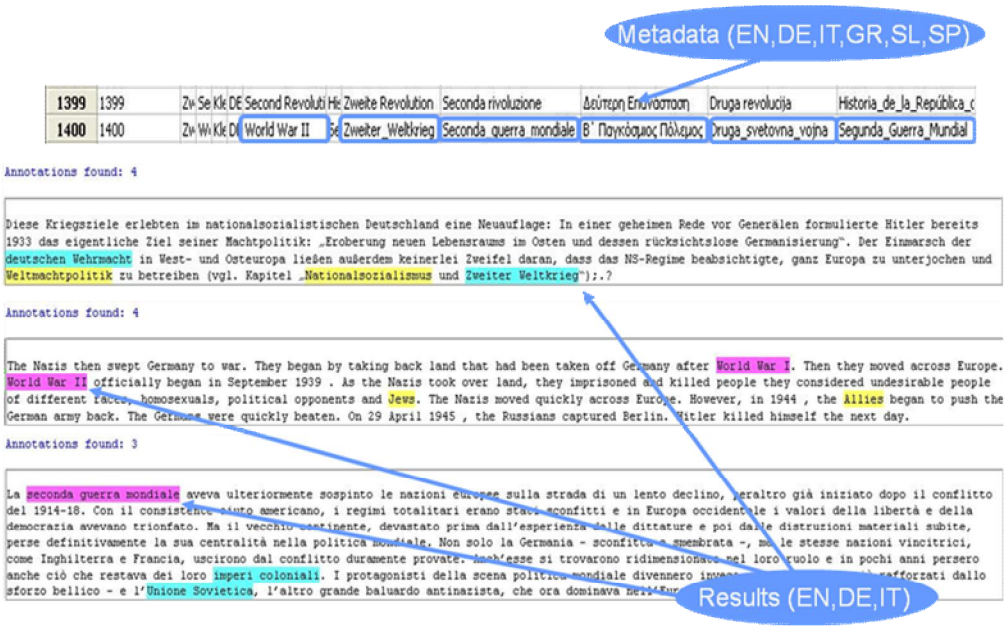



**Figure 5: Tagging example: visualization**

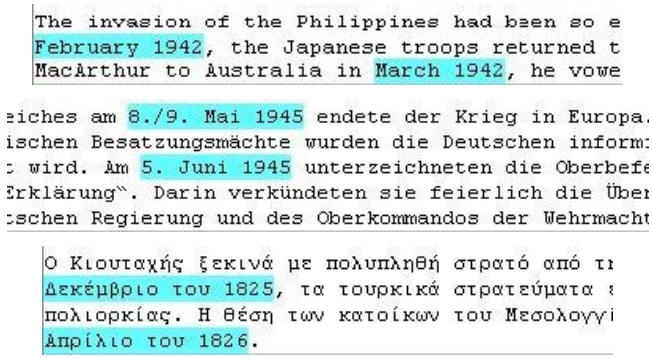
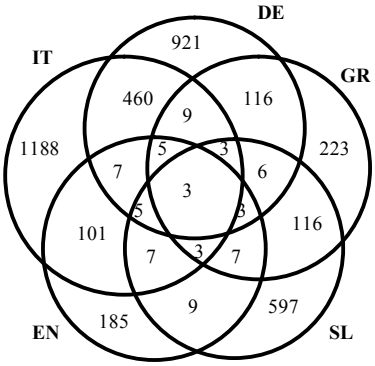
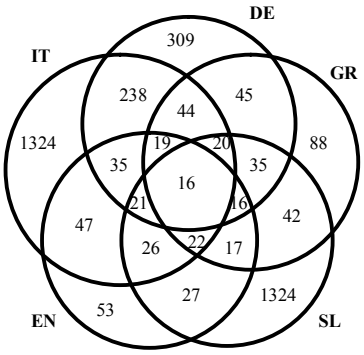



**Figure 6: Temporal tagging examples**




(a) Thematic metadata overlap




(b) Geographic metadata overlap

**Figure 7: Metadata overlap**

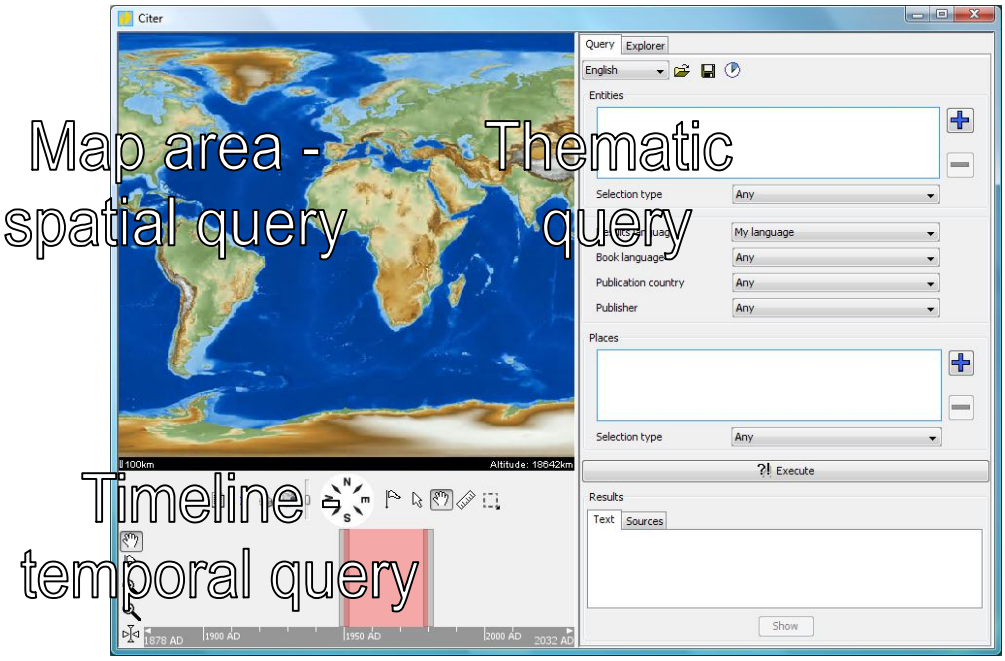



**Figure 8: Integration of new functionality and existing e-content portal solutions**

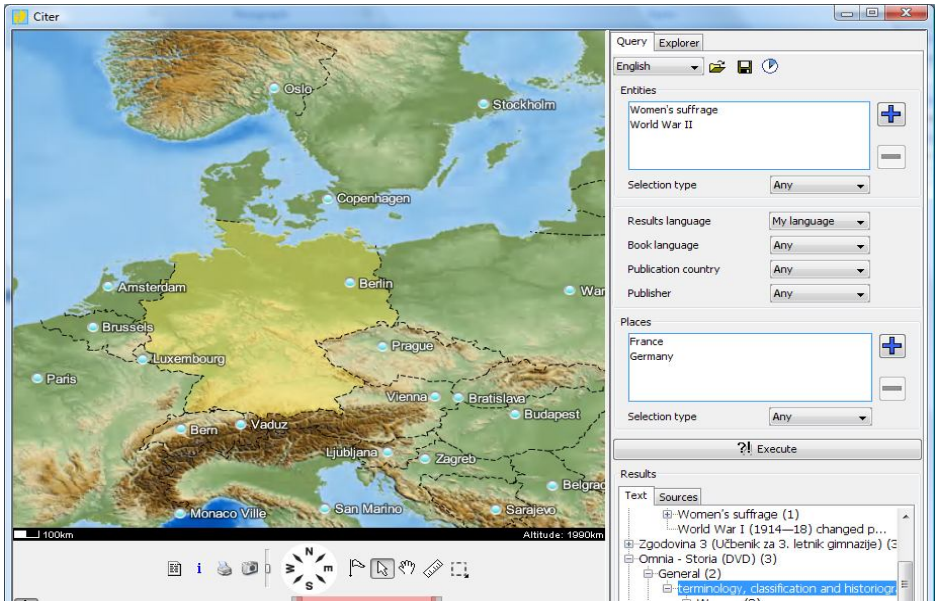



**Figure 9: Spatial search parameter**

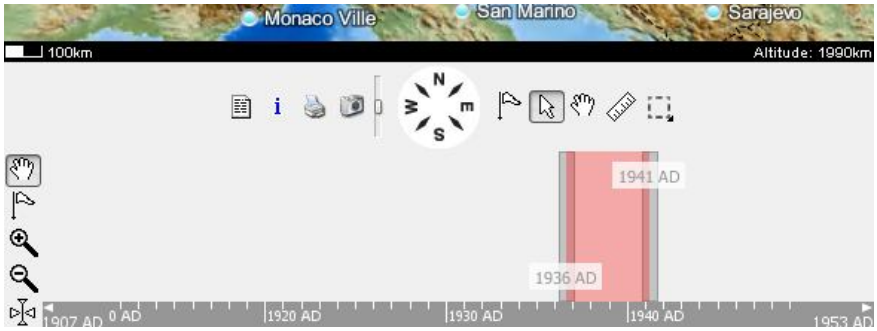



**Figure 10: Temporal search parameter**

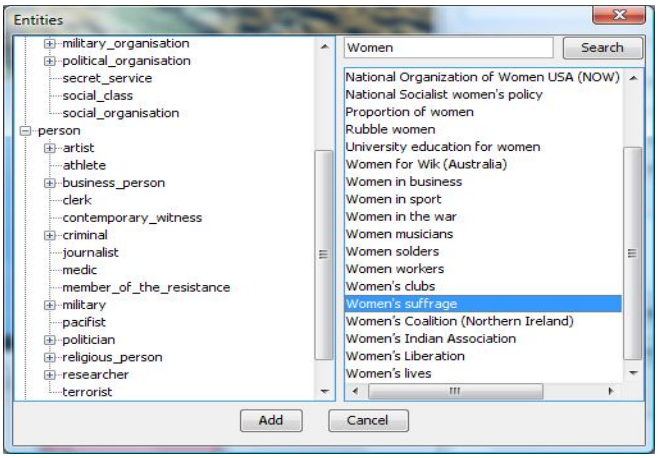



**Figure 11: Thematic search parameter**

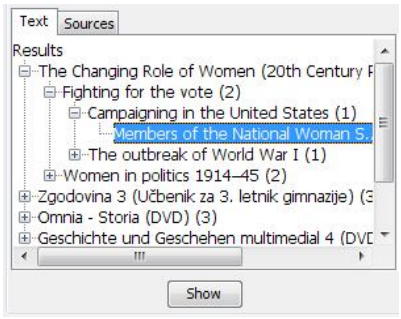



**Figure 12: Search result: tree structure**

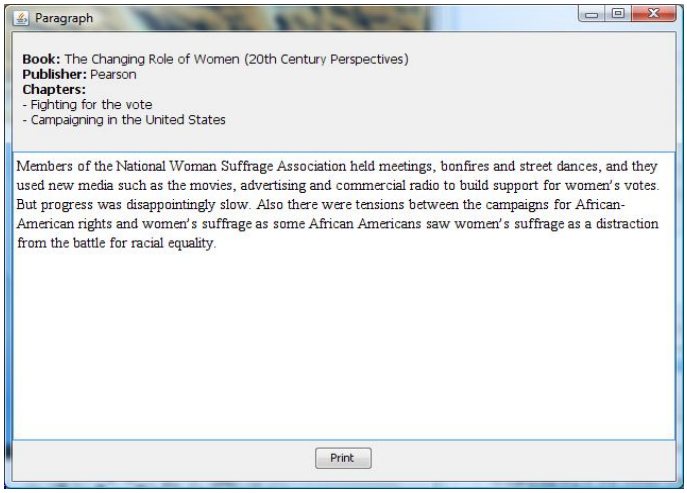



**Figure 13: Individual paragraph view**

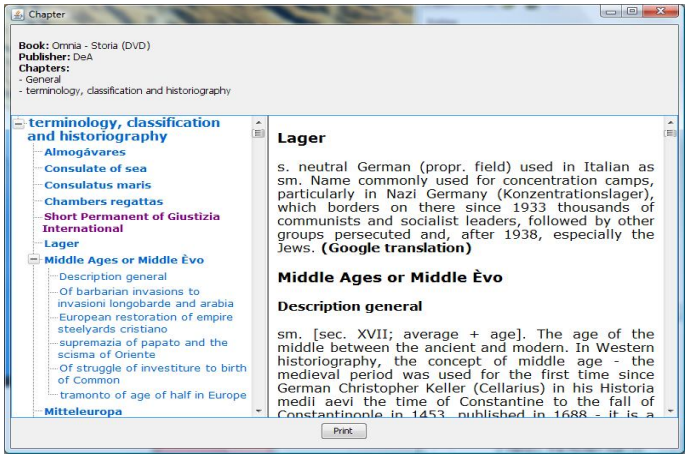



**Figure 14: Book chapter view**

# List of Figures