

International Workshop on  
**Supporting User's Exploration on  
Digital Libraries**

collocated with

International Conference on Theory and  
Practice of Digital Libraries (TPDL)

Cyprus, September 27th, 2012

*Workshop proceedings*

This workshop has been supported by the Seventh Framework Programme of the European Commission through the PATHS project (ICT-2011-270082).



ISBN: 978-84-616-0682-5

Edited by:

Eneko Agirre (University of the Basque Country)

Kate Fernie (MDR partners)

Arantxa Otegi (University of the Basque Country)

Mark Stevenson (University of Sheffield)

# Preface

Welcome to the International Workshop on Supporting Users' Exploration of Digital Libraries (SUEDL), held as part of the International Conference on Theory and Practice of Digital Libraries (TPDL).

Large amounts of digital cultural heritage information have become available recently, partly due to the rise of large-scale aggregators such as Europeana<sup>1</sup>, the European aggregator for museums, archives, libraries, and galleries. These large collections, and similar digital libraries (DL), present an important challenge to new users: discovering what items are present in the collection. In current systems support for item discovery is mainly provided through the standard keyword-based search paradigm, which is well suited for DL professionals who are highly familiar with the collections, subject areas, and have specific search goals. However, for new users who may not have a good understanding of the contents of the collections, be unsure about appropriate keywords, and have vague search goals, this method of access is unsatisfactory. Alternative item discovery methodologies are required to introduce new users to DL collections.

The aim of the workshop is to stimulate collaboration from experts and stakeholders in Digital Libraries, Cultural Heritage, Natural Language Processing and Information Retrieval in order to explore methods and strategies to support exploration of Digital Libraries, beyond the white box paradigm of search and click.

This half-day workshop consists on oral presentations of the authors, and an additional presentation of the PATHS prototype<sup>2</sup> by Mark Hall. In addition, we will be honored to have a keynote speech by David Haskiya, and we will be hosting a panel with Milena Dobрева (University of Malta), Kate Fernie (MDR partners), Mark Hall (University of Sheffield) and David Haskiya (Europeana).

This workshop has been supported by the Seventh Framework Programme of the European Commission through the PATHS project (ICT-2011-270082).

We would like to thank all people who in one way or another helped in making this workshop a success. Our special thanks go to our invited speaker, the plenary speakers, to our sponsors, to the participants of the panel discussion, to the members of the program committee who did an excellent job in reviewing the submitted papers, and to the TPDL organizers. Last but not least we would like to thank our authors and the participants of the workshop.

The Organizers  
Cyprus, September 2012

---

<sup>1</sup><http://www.europeana.eu>

<sup>2</sup><http://paths-project.eu/eng/Prototype>



## **Organizing committee**

Eneko Agirre (University of the Basque Country)  
Kate Fernie (MDR partners)  
Arantxa Otegi (University of the Basque Country)  
Mark Stevenson (University of Sheffield)

## **Program committee**

Donatella Castelli (ISTI-CNR)  
Valentine Charles (Europeana)  
Paul Clough (University of Sheffield)  
Giorgio Maria Di Nunzio (University of Padua)  
Cathal Gurrin (Dublin City University)  
Mark Hall (University of Sheffield)  
Gareth Jones (Dublin City University)  
Jaap Kamps (University of Amsterdam)  
Johan Oomen (Netherlands Institute for Sound and Vision)  
Martin Theobald (Max Planck)  
Max Wilson (University of Nottingham)



# Contents

Contents	v
Browsing Europeana – Opportunities and Challenges <i>David Haskiya</i>	1
Query Rewriting Using Shallow Language Processing: Effects on Keyword Subject Searches <i>Anna Mastora and Sarantos Kapidakis</i>	3
Visualising Television Heritage: The EUscreen Virtual Exhibi- tions and the Linked Open Data Pilot <i>Johan Oomen, Vassilis Tzouvaras, Kati Hyypä and Daniel Ockeloen</i>	15
Providing Suitable Information Accesses for New Users of Dig- ital Libraries Case Study DMG-Lib <i>Rike Brecht, Torsten Brix, Ulf Dring, Sascha Falke and Michael Reeing</i>	23
Exploring Pelagios: A Visual Browser for Geo-Tagged Datasets <i>Rainer Simon, Elton Barker and Leif Isaksen</i>	29





# Browsing Europeana – Opportunities and Challenges

David Haskiya

Europeana

**Abstract.** Browsing allows an information seeker to find what they did not know they needed or what they could not articulate in the form of a keyword query. As such it is a mode of discovery which could work very well within the extremely heterogeneous Europeana repository. This is especially important as user studies indicate that many Europeana users consider the Europeana portal to not provide clear enough information on the scope and type of content thus making their initial searches difficult to articulate.

In this presentation I will elaborate more on the opportunities and challenges of developing browse functions in the Europeana portal. I will touch upon a number of different forms of browsing, the opportunities they offer to users and the challenges of implementing them in the Europeana portal. The browsing types considered will be search with guided navigation, spatio-temporal browsing, subject browsing, tours/paths browsing and similarity browsing. Mockups and wireframes will be used throughout to illustrate the possible implementations.

While the presentation will focus on Europeana the principles should be applicable to any repository discovery service comprising cross-domain and multi-lingual content.



# Query Rewriting Using Shallow Language Processing: Effects on Keyword Subject Searches

Anna Mastora and Sarantos Kapidakis

Laboratory on Digital Libraries & Electronic Publishing, Archives & Library Science Department, Ionian University, 72 Ioanni Theotoki, Corfu, 49100 - Greece  
{mastora , sarantos}@ionio.gr

**Abstract.** The aim of this study is to investigate and report on potential implications of implementing shallow language processing towards rewriting keyword subject queries in Greek. The processing we report includes a speller and a lemmatiser along with stop word removal and query normalisation in terms of punctuation use. Among our findings is that users tend to submit morphologically variant words, which the Aspell tool, for spell checking and correcting, manages to process in a consistent way in 98.7% of the cases. We recorded a semantic drift of the initial query intent in approximately 8.2% of the overall submitted queries, after implementing the spell checker. The lemmatiser (ilsp\_nlp) performs extremely well for the words it identifies. Only five cases are recorded, among the initial 750 queries we submitted to the tool, which led to a semantic drift. However, the lemmatiser does not recognise either misspelled or truncated words which remained unaltered during this step of the process. Therefore, we conclude that, responding to the examined data, spelling prior to lemmatising is the appropriate sequence of implementing the specific shallow language processing.

**Keywords:** Digital libraries, Inflectional languages, Natural Language Processing, Spelling, Lemmatising, Query patterns

## 1 Introduction

Information seekers formulate their queries in a number of ways [1]. They tend, though, to query a system according to the paradigm and common practices of search engines, like Google, which involves throwing some keywords in a search box and expect to get the desired results. As Marchionini [2] puts it, the Web has legitimized browsing strategies that depend on selection, navigation, and trial-and-error tactics, which in turn facilitate increasing expectations to use the Web as a source for learning and exploratory discovery.

Searchers do not only experiment with different terms but also with various forms of the same word until they reach a satisfactory result set. But, they also make typing errors some of which they correct during the process. Other times they do not realise the mistake or do not know the correct form of the word. Such practices cause problems when string matching is implemented and lead to the users' frustration concern-

ing the retrieved results. This problem becomes more prominent when highly inflectional languages are involved which allow for the formulation of variant forms of the query words. Along with the phenomenon of cognitive polyrepresentation [3] the information searching task becomes a really demanding business for the non-expert user.

As stated by schraefel [4], successful paradigms can sometimes constrain our ability to imagine other ways to ask questions that might open up new and more powerful possibilities. Therefore, towards overcoming the aforementioned constraints for making exploratory search more efficient, especially since both browsing and searching are combined, language processing technologies can be implemented for query rewriting. Such techniques include spelling, stemming, lemmatising, part-of-speech tagging, named entity recognition, word sense disambiguation and others. Some of them are easier to implement and some others need either extreme pre-processing of the document corpus or are highly computationally intensive.

The special characteristics of each language dictate to a great extent which of the techniques can be implemented and whether they would secure a successful outcome. In this study we focus on recording possible implications from the implementation of shallow language processing, namely spelling, stop word removal and lemmatising, for keyword subject searches conducted in Greek focusing on the different input forms of the query words.

The remainder of the paper is organised as follows. Section 2 sets more specific aspects of our hypothesis and the research scenario, while section 3 discusses related works on the most prominent parts of our study. Section 4 presents our methodology, which is better clarified in section 5 where we present the results of our study. In section 6 we discuss particular issues which surfaced from the data processing. In section 7 we summarise the main conclusions of this work and, finally, in section 8 we refer to what has been planned as future steps of our research.

## 2 Aims and Objectives

The aim of this study is to investigate and report on potential implications of implementing shallow language processing towards rewriting keyword subject queries. Our study focuses on results concerning the use of the Greek language. The objectives of our study are:

- What are the identified query patterns in terms of word forms?
- How consistent is the speller's output when it treats different word forms?
- Does the implementation of the speller cause semantic drift of the initial query intent?
- How successful and reliable is the lemmatisation of words?
- Does the implementation of the lemmatiser cause semantic drift of the initial query intent?
- What is the outcome if, instead of doing either, perform both spelling and then lemmatising of the input query?

### 3 Related Work

As stated by Hearst [5], with the rise of the Web came the dominance of keywords as the primary query input type. Keyword queries consist of a list of one or more words or phrases –rather than full natural language statements. Some keyword queries consist of lists of different words and phrases, many others are noun compounds or proper nouns and, less frequently, contain syntactic fragments including prepositions or full syntactic phrases.

Concerning the language processing of queries, Schafer [6] states that deep natural language processing (DNLP) systems try to apply as much linguistic knowledge as possible to analyse natural language utterance. Schafer continues that, unlike DNLP, shallow natural language processing (SNLP) systems do not attempt to achieve an exhaustive linguistic analysis. They are designed for specific tasks ignoring many details in input and linguistic (grammar) framework. And, finally, some examples are provided for SNLP like tokenisation, part-of-speech tagging, chunking, named entity recognition, and shallow sentence parsing.

Lemmatisation is used for handling morphological variations in highly inflectional languages. As stated in Kettunen et al. [7], it is most often used to describe the process when inflected word form and its (dictionary) base form are related to each other with an algorithm. Lazarinis [8] argues that the Greek language is grammatically more complex than the English language. It has conjugations and morphologically complex words. Articles, verbs, nouns, first names and surnames may be in various cases (nominative, genitive, etc), in singular or plural form and they are differentiated according to their gender (masculine, feminine, neuter).

### 4 Methodology

For the purpose of our study we used 750 queries which were collected during a hands-on experiment. More specifically, we used as the test set the queries submitted to a bibliographic database during a hands-on experiment with the participation of 47 undergraduate and postgraduate students at the Department of Archives and Library Science at the Ionian University in Corfu (Greece). They were given thirteen (13) information needs related to environmental issues and asked to submit appropriate queries in order to retrieve relevant documents. Some of the information needs were about *migratory birds*, the *greenhouse effect*, *laboratory animals* and *national parks* to name a few. The database contained material mainly from the environmental domain and participants could only submit queries in the Subject field using keywords in Greek. For the purpose of this experiment we selected and customized approximately 14,400 bibliographic records of the Evonymos Ecological Library<sup>1</sup>.

The queries we present here were the ones which resulted to zero hits when submitted to the database. The next step of the examination process involved the spell checking and correcting of these queries while making transformations in terms of the

---

<sup>1</sup> Full database available at <http://www.evonymos.org/index.html> (Accessed July 2012)

use of accents and the case of the letters. The following step involved the lemmatisation of the initial queries with no prior editing and, finally, the lemmatisation of the queries after them being spelled checked and corrected. Although the reported results for the tools' performance are derived from statistics from the mediator programs we developed for processing and analysing our data, the qualitative evaluation of the tools' performance was done manually by checking the results one by one in order to make the appropriate judgments.

#### 4.1 The tools: Speller and Lemmatiser

For our study we used Aspell, v.0.60.6<sup>2</sup> for spell checking and correcting the queries. Aspell is a utility program that connects to the Aspell library so that it can function as an `ispell -a` replacement, as an independent spell checker, as a test utility to test out Aspell library features, and as a utility for managing dictionaries used by the library. The Aspell library contains an interface allowing other programs direct access to its functions and therefore reducing the complex task of spell checking to simple library calls.

We also used the `ilsp_nlp`<sup>3</sup> tool, which lemmatises Greek texts. The input is either plain text or an XCES document with text segmented in paragraphs. The service detects paragraph sentence and token boundaries and generates Parts Of Speech and lemma annotations for each token. The output by default is an XCES document. The tool is developed by the Greek Institute for Language and Speech Processing<sup>4</sup> and is freely available through a web interface.

## 5 Results

This section presents the results of our study. We first report on the identified query patterns, and then on the outcome of each of the tests we performed with the language processing tools we used, namely the speller and lemmatiser.

### 5.1 Users' Queries: Observations on Word Formulation Patterns

Our observations concerning the users' word formulation patterns during the search process focus on specific characteristics of the Greek language. Following we present the most common characteristics providing examples from the users' queries during the laboratory experiment.

- i. Singular or plural of the same concept (*park* - *parks*).
  - o e.g. *πάρκο* - *πάρκα*
- ii. Use of accents or not. Concerning the Greek language accents define the way each word is pronounced and, sometimes, the meaning of the word. In the

---

<sup>2</sup> Aspell, v.0.60.6. ©2000-2004 by Kevin Atkinson, //aspell.net/ (last accessed July 2012)

<sup>3</sup> The lemmatiser `ilsp_nlp` is available here //nlp.ilsp.gr/soaplab2-axis/ (last accessed July 2012)

<sup>4</sup> <http://www.ilsp.gr/en> (last accessed July 2012)

example given below the first word stands for *oranges*, the second for the *orange tree* and we cannot be certain of the user's intention concerning the third word, since the accent is missing.

- e.g. πορτοκάλια - πορτοκαλιά – πορτοκαλία

iii. Use of capitalised or small case characters. The example below stands for *national park*.

- e.g. ΔΡΥΜΟΣ - δρυμός - Δρυμός

iv. Use of words in various grammatical cases. Greek is a highly inflectional language; ergo grammatical cases are common in use. The example below shows the use of, in order of appearance, nominative, genitive and accusative cases of the plural number. The word presented stands for *national*.

- e.g. εθνικοί - εθνικών - εθνικούς

v. Use of very common words which are usually ignored by the information systems, i.e. stop words, like articles or prepositions. The example given here is an alternative search for the *greenhouse effect*, except that there is the definite article between the two nouns.

- e.g. φαινόμενο του θερμοκηπίου

vi. Use of different valid forms of a word. This case refers to words that may have different spelling but represent the same concept and can be used interchangeably. The words appearing below stand for the word *tomato*.

- e.g. τομάτα - ντομάτα

vii. Misspelled words.

- e.g. ANITPPHΣEΙΣ, instead of ANTIPPHΣEΙΣ, which is obviously a simple typing error because it changed the correct sequence of the letters. The word ANTIPPHΣEΙΣ means *objections*.
- e.g. ΥΓEΙHNH, instead of ΥΓEΙINH, which is more likely an error due to lack of knowledge of the correct spelling. “H” and “I” represent the same sound but are not used interchangeably. The word here means *hygiene*.

In this category we could also include the cases where two words were typed as one, losing the space between them, like ΟΠΩΡΟΦΟΡΑΔΕΝΤΡΑ, instead of two separate words, i.e. ΟΠΩΡΟΦΟΡΑ ΔΕΝΤΡΑ, which stands for *fruit trees*.

viii. Use of Greeklish. Greeklish is the use of the Latin alphabet (along with some numbers replacing look-alike letters) to write Greek words based, mainly, on the way they sound. It is an informal way of writing, which lacks strict rules and it basically started to be of use for short communications where, in electronic media, the Greek alphabet was not supported. Nowadays it is broadly used, especially in social media or for text messaging. However, no official communication is established in Greeklish and they are certainly not used for indexing purposes in the libraries' catalogues.

- e.g. PERIBALLON, instead of ΠΕΡΙΒΑΛΛΟΝ, which stands for *environment*.

ix. Use of Named Entities and abbreviations. This category includes geographical names and persons' or organisations' names. The examples below stand

for *Ionian (sea)*, *Princess Sissy* and *ARCTUROS*. The latter is the name of a Greek environmental organisation.

- e.g. *Ιόνιο, ΠΡΙΓΚΙΠΙΣΣΑ ΣΙΣΣΥ, αρκτούρος*.

Some abbreviations were also used by the participants for their searches like in the case of *Γενετικά Τροποποιημένους Οργανισμούς* which stands for *Genetically Modified Organisms* (GTO).

- e.g. *ΓΤΟ*
- x. Use of truncated words. The participants in the experiment submitted a great deal of truncated words, aiming, most probably, to increase recall in their searches. The words below stand for *genet-ically modif-ied organis-ms*.
- e.g. *γενετ-ικά τροποποιημεν-οι οργανισ-μοί*
- xi. Additional characteristics of the queries the participants formulated are:
- The use of punctuation signs, like the hyphen and the quotes.
    - e.g. *ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΑΩΠΛΙΑ (national park-flora)*.
  - The use of nouns. It outnumbers the use of any other part of speech. Somewhat common is the use of adjectives and participles, while the use of verbs is extremely limited. To give an example of the proportion, among the 750 queries we processed, we came across with only one verb.
  - The formulation of keyword queries with very few words.

## 5.2 Spelling: Results and Considerations

As described in section 4.1, we used Aspell for spell checking and correcting the queries of our study. The outcome of the tests is presented in this section.

In order to examine the outcome of the speller in terms of its consistency, we performed a number of runs of the initial query corpus to the speller. The logic behind the steps' sequence is based on the morphology of the Greek language as well as on the users' identified query patterns along with the speller's functions. The steps of the process were as follows.

- i. Ran the initial –not edited– queries' corpus to the speller.
- ii. Removed punctuation and phrase quotes. Ran queries to the speller.
- iii. Removed punctuation, phrase quotes and stop words. Ran queries to the speller.
- iv. Removed punctuation, phrase quotes and stop words. Then transformed all queries to upper case letters without accents. Ran queries to the speller.
- v. Removed punctuation, phrase quotes and stop words. Then transformed all queries to upper case letters (accents were preserved if present at initial query). Ran queries to the speller.
- vi. Removed punctuation, phrase quotes and stop words. Then transformed all queries to lower case letters without accents. Ran queries to the speller.
- vii. Removed punctuation, phrase quotes and stop words. Then transformed all queries to lower case letters (accents were preserved if present at initial query). Ran queries to the speller.



- viii. Removed punctuation, phrase quotes and stop words. Then removed all accents (case of the letters was preserved as in the initial query). Ran queries to the speller.

We submitted to Aspell 750 queries which are further broken down to 650 unique formulations. This means that the 650 queries carry unique formulation characteristics in terms of accents and letter cases. The study reveals that the overwhelming majority (98.7%) of the submitted queries to Aspell are treated in a consistent way regardless of the word format. We observe that both the presence or absence of accents and the case of letters do not affect Aspell’s consistency. The processed words with Aspell lead to the same results irrespective of their initial input form. Figure 1 presents some characteristic cases of the speller’s performance.

(No punctuation, No quotes, No Stop Words) +								
Initial Query	Initial Query Spelled	Initial Query (No punctuation, No quotes) Spelled	Initial Query (No punctuation, No quotes, No Stop Words) Spelled	Initial Query (Upper case, No accents) Spelled	Initial Query (Upper case, accents as initial) Spelled	Initial Query (Lower case, No accents) Spelled	Initial Query (Lower case, accents as initial) Spelled	Initial Query (No accents, case as initial) Spelled
λίπασμα	λίπασμα	λίπασμα	λίπασμα	ΛΙΠΑΣΜΑ	ΛΙΠΑΣΜΑ	λίπασμα	λίπασμα	λίπασμα
ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΛΩΡΙΔΑ	ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΛΩΡΙΔΑ	ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΛΩΡΙΔΑ	ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΛΩΡΙΔΑ	ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΛΩΡΙΔΑ	ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΛΩΡΙΔΑ	εθνικός δρυμός χλωρίδα	εθνικός δρυμός χλωρίδα	ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ ΧΛΩΡΙΔΑ
τρύπα του όζοντος	τρύπα του όζοντος	τρύπα του όζοντος	τρύπα όζοντος	ΤΡΥΠΑ ΟΖΟΝΤΟΣ	ΤΡΥΠΑ ΟΖΟΝΤΟΣ	τρύπα όζοντος	τρύπα όζοντος	τρύπα όζοντος
ΟΠΩΡΟΦΩΡ ΔΕΝΤΡ	ΟΠΩΡΟΦΩΡΑ ΔΕΝΤΡΑ	ΟΠΩΡΟΦΩΡΑ ΔΕΝΤΡΑ	ΟΠΩΡΟΦΩΡΑ ΔΕΝΤΡΑ	ΟΠΩΡΟΦΩΡΑ ΔΕΝΤΡΑ	ΟΠΩΡΟΦΩΡΑ ΔΕΝΤΡΑ	οπωροφόρα δέντρα	οπωροφόρα δέντρα	ΟΠΩΡΟΦΩΡΑ ΔΕΝΤΡΑ
Αιδήνός	Αιδήνός	Αιδήνός	Αιδήνός	ΑΙΔΗΝΟΣ	ΑΙΔΗΝΟΣ	Αιδήνός	Αιδήνός	Αιδήνός
γεωθερμική	γεωθερμική	γεωθερμική	γεωθερμική	ΓΕΩΘΕΡΜΙΚΗ	ΓΕΩΘΕΡΜΙΚΗ	γεωθερμική	γεωθερμική	γεωθερμική
γενετικ τροποποιη οργανισμ	γενετικά τροποποιη οργανισμ	γενετικά τροποποιη οργανισμ	γενετικά τροποποιη οργανισμ	ΓΕΝΕΤΙΚΑ ΤΡΟΠΟΠΟΙΩ ΟΡΓΑΝΙΣΜΟ	ΓΕΝΕΤΙΚΑ ΤΡΟΠΟΠΟΙΩ ΟΡΓΑΝΙΣΜΟ	γενετικά τροποποιη οργανισμ	γενετικά τροποποιη οργανισμ	γενετικά τροποποιη οργανισμ

Fig. 1. Results from speller after various transformations of the initial query word forms

The examples we provide here are representative of the most common cases we came across. They are presented here in order of appearance under the “Initial query” column (see Fig.1 above).

**λίπασμα:** Misspelled, small case word for “λίπασμα”, which stands for “fertiliser”.

**ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ-ΧΛΩΡΙΔΑ:** All caps, hyphenated words, which stand for “national parks-flora”. The hyphen here probably represents a subject subdivision.

**τρύπα του όζοντος:** Small case words which stands for “ozone hole”. The word “του” here is an article in singular genitive and is considered a stop word.

**ΟΠΩΡΟΦΩΡ ΔΕΝΤΡ:** Misspelled, all caps, truncated words for “ΟΠΩΡΟΦΩΡΟ ΔΕΝΤΡΟ” or “ΟΠΩΡΟΦΩΡΑ ΔΕΝΤΡΑ”, the first being in singular and the second in plural number. It stands for “FRUIT TREE” or “FRUIT TREES”.

**Αιδιψός:** Misspelled, Named entity, i.e. the name of the city "Αιδιψός".

**φεωθερμική:** Misspelled, small case word for "γεωθερμική", which stands for "geothermal".

**γενετικ τροποποιη οργανισμ:** Small case, all truncated words with the middle one being misspelled as well. The correct truncated form would be "γενετικ τροποποιημ οργανισμ". This phrase stands for "genetically modified organisms".

We observed some kind of inconsistency only in ten (10) of the overall submitted queries to Aspell. Actually, among them, there are seven (7) unique issues to deal with, since four (4) of these cases refer to the same query, namely *ΠΕΡΙΟΧΗ ΤΗΣ ΜΕΣΟΓΕΙΟΥ* (Area of the Mediterranean), which, along with one more, report to the same phenomenon, namely the capitalised genitive form of the female definitive article, i.e. *ΤΗΣ*. This seems to cause some kind of inconsistent results.

Figure 2 below depicts the cases which resulted to inconsistencies after the implementation of the spell checking and correcting. We observe that, in some cases, the input query causes problems to the speller when truncated words are involved (like in the last two cases) or in the case of unidentified Named Entities (like the third entry).

The results presented so far concern the consistency of the speller's behavior and refer to judgments on the words' morphology. The main criterion up to now was based on whether the speller treated all kinds of word forms the same way. This means that the critique concerned the last eight (8) columns of the figures we include in this text. There is still need to comment on the identified cases of semantic drifts of the initially submitted query after the queries were spell checked and corrected. This kind of examination refers to the comparison between the first column, which presents the initial query as submitted by the user, and the results on the following eight columns altogether.

In our study we judge the semantic consistency of the query transformations with manually examining the speller's outcome. We consider the words which change drastically and are easy to locate and clarify. We have recorded 61 queries, out of the 750, i.e. approximately 8.2%, that resulted to a semantic drift after implementing the spell checker. These 61 queries represent 48 unique cases. Eight (8) of these latter concern unidentified Named Entities, fourteen (14) have occurred from initially truncated words, while thirteen (13) are generated from queries with typing errors. The remaining queries concern various issues.

The table below presents, following the first column, the cases of the following submitted queries: *increase of earth temperature*, *temperature increase*, *Elati Pertouli* (two locations), *sea turtle*, *Area of the Mediterranean*, *river – forrest – lake – mountain – mount* (all truncated), *sea level clim(ate)?*.

				(No punctuation, No quotes, No Stop Words) +				
Initial Query	Initial Query Spelled	Initial Query (No punctuation, No quotes) Spelled	Initial Query (No punctuation, No quotes, No Stop Words) Spelled	Initial Query (Upper case, No accents) Spelled	Initial Query (Upper case, accents as initial) Spelled	Initial Query (Lower case, No accents) Spelled	Initial Query (Lower case, accents as initial) Spelled	Initial Query (No accents, case as initial) Spelled
άνοδος θερμοκρασίας γης	άνοδος θερμοκρασίας γης	άνοδος θερμοκρασίας γης	άνοδος θερμοκρασίας γης	ΑΝΟΔΟΣ ΘΕΡΜΟΚΡΑΣΙΑΣ ΓΗ	ΑΝΟΔΟΣ ΘΕΡΜΟΚΡΑΣΙΑΣ ΓΗ	άνοδος θερμοκρασίας γη	άνοδος θερμοκρασίας γη	άνοδος θερμοκρασίας γης
ΑΥΞΗΣΗ ΤΗΣ ΘΕΡΜΟΚΡΑΣΙΑΣ	ΑΥΞΗΣΗ ΤΗ ΘΕΡΜΟΚΡΑΣΙΑΣ	ΑΥΞΗΣΗ ΤΗ ΘΕΡΜΟΚΡΑΣΙΑΣ	ΑΥΞΗΣΗ ΘΕΡΜΟΚΡΑΣΙΑΣ	ΑΥΞΗΣΗ ΘΕΡΜΟΚΡΑΣΙΑΣ	ΑΥΞΗΣΗ ΘΕΡΜΟΚΡΑΣΙΑΣ	αύξηση θερμοκρασίας	αύξηση θερμοκρασίας	ΑΥΞΗΣΗ ΘΕΡΜΟΚΡΑΣΙΑΣ
Ελάτη Πετρούλα	Ελάτη Πετρούλα	Ελάτη Πετρούλα	Ελάτη Πετρούλα	ΕΛΑΤΗ ΠΕΤΡΟΥΛΑ	ΕΛΑΤΗ ΠΕΡ ΤΟΥΛΑ	ελάτη Πετρούλα	ελάτη Πετρούλα	Ελάτη Πετρούλα
θαλάσσια χελώνα	θα λάσπια χελώνα	θα λάσπια χελώνα	θα λάσπια χελώνα	ΘΑΛΑΣΣΙΑ ΧΕΛΩΝΑ	ΘΑ ΛΑΣΙΑ ΧΕΛΩΝΑ	θαλάσσια χελώνα	θα λάσπια χελώνα	θαλάσσια χελώνα
ΠΕΡΙΟΧΗ ΤΗΣ ΜΕΣΟΓΕΙΟΥ	ΠΕΡΙΟΧΗ ΤΗ ΜΕΣΟΓΕΙΟΥ	ΠΕΡΙΟΧΗ ΤΗ ΜΕΣΟΓΕΙΟΥ	ΠΕΡΙΟΧΗ ΜΕΣΟΓΕΙΟΥ	ΠΕΡΙΟΧΗ ΜΕΣΟΓΕΙΟΥ	ΠΕΡΙΟΧΗ ΜΕΣΟΓΕΙΟΥ	περιοχή Μεσογείου	περιοχή Μεσογείου	ΠΕΡΙΟΧΗ ΜΕΣΟΓΕΙΟΥ
ποτάμ διας λιμν βοούν ρο	ποτάμ Δίας Λίμνη βοούν ρο	ποτάμ Δίας Λίμνη βοούν ρο	ποτάμ Δίας Λίμνη βοούν ρο	ΠΟΤΑΜΙ ΔΑΣΗ ΛΙΜΝΗ ΒΟΟΥΝ ΡΟ	ΠΟΤΑΜΙ ΔΑΣΗ ΛΙΜΝΗ ΒΟΟΥΝ ΡΟ	ποτάμ δάση Λίμνη βοούν ρο	ποτάμ δάση Λίμνη βοούν ρο	ποτάμ Δίας Λίμνη βοούν ρο
στάθμη θάλασσας	στάθμη θάλασσας	στάθμη θάλασσας	στάθμη θάλασσας	ΣΤΑΘΜΗ ΘΑΛΑΣΣΑ	ΣΤΑΘΜΗ ΘΑΛΑΣΣΑ	στάθμη θάλασσά	στάθμη θάλασσά	στάθμη θάλασσας

Fig. 2. Cases of the speller's inconsistent treatment

### 5.3 Lemmatising: Results and Considerations

For examining the outcome of the lemmatising process we performed the following checks.

- i. Ran the initial –not edited– queries' corpus to the lemmatiser.
- ii. Removed punctuation, phrase quotes and stop words. Ran queries to the lemmatiser.
- iii. Removed punctuation, phrase quotes and stop words. Ran queries to the speller. Ran queries to the lemmatiser.

After performing the first check, we observed that there is a semantic drift only in five (5) cases out of the 750 overall queries. These cases represent two issues. Three of the cases are Named Entities which were not recognised by the tool and were given other meanings. The remaining two refer to the same word, namely *όζων*, which is the misspelled *όζον* and was not disambiguated appropriately. Unfortunately, this is not the overall performance of the lemmatiser. All the misspelled words were not identified and, consequently, they remained unaltered after this step of the process. The same performance was recorded for the truncated words.

More specifically, during the first run of the initial queries to the lemmatiser we recorded 284 unknown words, i.e. 20.67%. Let us clarify at this point that the 750 queries, which we examined, represented 1,374 words, i.e. tokens. During the second run, for which we eliminated punctuation and quotes along with the removal of stop words, the report showed that 272 tokens remained unrecognised, i.e. 19.79%. These two steps revealed that there were a significant proportion of unidentified tokens during the lemmatisation process of the initial queries.

At this point, we conducted the last queries' run. We removed any allocated punctuation, phrase quotes and stop words and then ran the queries to the speller followed by their submission to the lemmatiser. The outcome of this last step is that 31 of the initial 750 queries remained unrecognised for the lemmatiser, i.e. 2.25%.

## 6 Discussion

As far as the performance of the speller is concerned, there is another judgment involved but not currently analysed, though it remains an open subject for our ongoing research. In some cases, although the characters of the words remained the same as in the initial query after the processing with the speller, there was, however, a change in the position of the accent, which occasionally dictates a change in the meaning of the word. We have already presented, in section 5.1 (example no ii), such cases. The problem becomes particularly important if an information system considers accents during search and retrieval. This is an aspect which we will present in the future in more detail since it requires deeper analysis taking into consideration the semantic, or even the pragmatic, aspects of the initial query.

An additional finding, in terms of what to expect when implementing language processing techniques, is the unexpected characters contained in the input, like in the case of the query *Μεσογειο*. The fact that the last two characters were in Latin script instead of Greek caused crash reports during the examination process.

Stopwords are the terms which appear too frequently in documents and thus their discriminatory value is low [8]. This is why most information systems disregard them in any of the techniques they implement. This is an aspect with which we, too, tried to deal with. In the process we followed, we observed that what even mattered was the step during which we chose to eliminate stop words. This happened because in some cases, either during the spelling or the lemmatisation of queries, stop words were retrieved as a result of the query transformation process. After various tests we concluded that when it comes to noun keyword searches it is safer to remove stop words as early as possible.

The problem of the semantic drift becomes more prominent in cases where the initial query was truncated. This requires the identification of the query intent in order to make the right judgment. In our study, we knew in advance of both the query intent of the user, because we created the information need, and we could make relatively safe judgments using the submitted queries of other users while searching for a specific topic.

## 7 Conclusions

The results presented in this study are part of our overall research in the field of query expansion using knowledge organisation systems, like thesauri. The research involves the query transformation based on the matching between query terms and terms from knowledge organisation systems in order to facilitate the users' task when searching for information. Usually searchers submit few words in the search box in order to get started with the information seeking process. This first encounter needs to be both engaging and effective for the searcher. This conceptualisation was the motive for combining natural language processing techniques with the process of expanding the user's initial query. At this phase of our study, we explored the performance of the tools implemented for shallow language processing motivated by the fact that users

tend to submit morphologically variant words which, in the case of a highly inflectional language like Greek, increases the input and the different cases a tool has to cover.

Among our findings is that the Aspell tool, for spell checking and correcting, manages to process in a consistent way the variant types of input in 98.7% of the cases. And we recorded some kind of semantic drift of the initial query intent in approximately 8.2% of the overall submitted queries, after implementing the spell checker.

The lemmatiser (ilsp\_nlp) performs extremely well for the words it identifies since only five cases are recorded, among the initial 750 queries we submitted to the tool, which led to semantic drift. However, the lemmatiser does not recognise either misspelled or truncated words which remained unchanged during this step of the process. Their percentage was essential to be ignored. Therefore, we proceeded to first spelling and then lemmatising the input queries. The processing of the data and the respective outcome led us to the conclusion that spelling prior to lemmatising is the appropriate sequence of implementing the specific shallow language processing.

Our findings seem consistent to previous works, like in the case reported in Lazarinis [7] where it is concluded that in the specific study, with the aid of authentic queries, it was made obvious that after the application of lemmatisation and stop word elimination more relevant documents were retrieved. The performance of the tools we examined and the sequence of steps we followed provide reasons to be optimistic that shallow language processing does benefit information retrieval as far as highly inflectional languages are concerned.

## 8 Future Considerations

One significant future consideration we have already planned for this ongoing study is to identify within the query corpus those phrases which should not be considered as random keywords during shallow language processing. This involves the kind of units which may carry a certain meaning when considered individually but as a phrase represent a different, distinguishing concept. We believe that this is an important aspect of our research which has to be addressed in conjunction with the lessons already learned.

In addition, the examination of more language processing tools and techniques is essential in order to provide both statistically and qualitatively confirmed results. The tools we have used so far perform well enough in relation to current data but our queries, though adequately representative of real user queries, they, for example, do not include verbs. Since our broader research interest involves the exploitation of language technologies against databases and knowledge organisation systems (KOS), the tools should be evaluated against this content, too.

Finally, we need to compare our results with similar techniques implemented in other languages, like Finnish, which is also highly inflectional.

**Acknowledgment.** This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program

"Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

## References

1. Jansen, B.J., Booth, D.L. & Spink, A., 2009. Patterns of query reformulation during Web searching. *Journal of the American Society for Information Science and Technology*, 60(7), pp.1358–1371.
2. Marchionini, G., 2006. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4), pp.41–46.
3. Ingwersen, P., 1994. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Dublin, Ireland: Springer-Verlag New York, Inc., pp. 101–110. Available at: <http://portal.acm.org/citation.cfm?id=188519> [Accessed July 2012].
4. schraefel, m.c., 2009. Building knowledge: what's beyond keyword search? *COMPUTER*. Available at: <http://eprints.soton.ac.uk/269963/> [Accessed September 2012].
5. Hearst, M., 2009. *Search user interfaces*, New York: Cambridge University Press.
6. Schafer, U., 2007. Integrating Deep and Shallow Natural Language Processing Components – Representations and Hybrid Architectures. Ph.D. thesis, Faculty of Mathematics and Computer Science, Saarland University, Saarbrücken, Germany. Available [http://scidok.sulb.uni-saarland.de/volltexte/2007/1326/pdf/Dissertation\\_1383\\_Schae\\_Ulri\\_2007.pdf](http://scidok.sulb.uni-saarland.de/volltexte/2007/1326/pdf/Dissertation_1383_Schae_Ulri_2007.pdf) (Accessed July 2012)
7. Lazarinis, F., 2007. Lemmatization and stopword elimination in Greek web searching. In *Proceedings of the 2007 Euro American conference on Telematics and information systems*. EATIS '07. New York, NY, USA: ACM, pp. 61:1–61:4. Available at: <http://doi.acm.org/10.1145/1352694.1352757> [Accessed July 2012].
8. Kettunen, K., Kunttu, T. & Järvelin, K., 2005. To stem or lemmatize a highly inflectional language in a probabilistic IR environment? *Journal of Documentation*, 61(4), pp.476–496.

# Visualising Television Heritage: The EUscreen Virtual Exhibitions and the Linked Open Data Pilot

Johan Oomen<sup>1</sup>, Vassilis Tzouvaras<sup>2</sup>, Kati Hyyppä<sup>3</sup>, and Daniel Ockeloën<sup>4</sup>

<sup>1</sup> Nederlands Instituut voor, Beeld en Geluid, Hilversum, the Netherlands

joomen@beeldengeluid.nl

<sup>2</sup> National Technical University of Athens, Greece

tzouvaras@image.ntua.gr

<sup>3</sup> Aalto University, Helsinki, Finland

kati.hyyppa@aalto.fi

<sup>4</sup> Noterik BV, Amsterdam, The Netherlands

daniel@xs4all.nl

**Abstract.** The EUscreen initiative represents the European television archives and acts as a domain aggregator for Europeana, Europe's digital library, which provides access to over 20 million digitized cultural objects. The main motivation for the initiative is to provide unified access to a representative collection of television programs, secondary sources and articles, and in this way to allow students, scholars and the general public to study the history of television in its wider context. This paper explores the EUscreen activities related to [1] novel ways to present curated content in virtual exhibitions and [2] publishing EUscreen metadata as Linked Open Data. Regarding the latter, it is demonstrated how available metadata can be enriched and visualized using a timeline interface.

**Keywords.** Visualization , Linked Open Data, TV on the Web, Metadata Interoperability, Europeana

## 1. Introduction

The main motivation of the EUscreen initiative is to provide a unified online access to a representative collection of television programs, secondary sources and articles, allowing students, scholars and the general public to explore European audiovisual cultural heritage.

The multidisciplinary nature of the EUscreen project is mirrored in the composition of the socio-technical nature of the consortium; comprising of 20 collection owners, technical enablers, legal experts, educational technologists and media historians from 20 countries. EUscreen represents all major European television archives and acts as one of the key domain aggregators providing content to Europeana, the platform connecting millions and millions of cultural artifacts across Europe.

The goals of the project are to (i) develop a state-of-the-art workflow for content ingestion, (ii) define content selection and IPR management methodology (35.000 items will be made available), and (iii) design and implement a front-end that accommodates requirements from several user groups.

Close cooperation between the different stakeholders in the consortium is essential in order to reach these goals. For example, the content selection policy needs to take in to account, not only the available content, but also the needs of the user groups as well as copyright issues. The workflow also needs to accommodate to the existing metadata structures, support aggregation by Europeana<sup>1</sup> and provide support for multilingual access. In this paper we

---

<sup>1</sup> <http://www.europeana.eu>

demonstrate the EUscreen metadata workflow focusing on the metadata ingestion, the creation of the Linked Open Data and an application that uses the data.

The paper is organized as follows. The first section highlights the EUscreen project, its main objectives and metadata ingestion workflow. After discussing the related work, Section three the front-end and the design of the EUscreen portal. Section four discusses the design and development of the tools for creating virtual exhibitions as a novel way to navigate through the collection in a curated fashion. Section five illustrates the procedure followed for the publication of metadata as Linked Open Data. Finally, the timeline application that used the LOD, is presented.

## 2. Related Work

The most closely related project to EUscreen is the European Film Gateway project. Rooted in the film archiving domain, EFG provides links to cinema history related materials and contributes this information to Europeana. References to moving images on the portal point to local websites where the materials can be watched in varying formats and qualities.

Moving image archives are continuously looking for ways to bring their content online without foregoing the uniqueness of their materials, and to strengthen the reach of their collections. The dominating power in online video, YouTube, is far from a perfect fit for that vision. Amongst other reasons for this is its lack of support for extensive metadata models, lack of contextual information, lack of flexibility of integration with external platforms (ingest, export), lack of advanced methods to prohibit illegal downloading. Video sharing sites such as YouTube are leading the way in introducing new interaction mechanisms.

In the course of 2011, different European platforms have sprung up to let users gain access to audiovisual heritage materials. The Swedish Filmarkivet launched in February 2011 with 300 short films, non-fiction films, newsreels and commercials: films that ‘reflect the transformation of Swedish society over the last century’. Another initiative that combines national historical documents with online video sources is the Dutch website Film in Nederland that was launched in 2011.

MAVISE holds key profile and contact data for around 29 national TV markets, 8.918 television channels, 6.060 companies. This database was developed by the Strasbourg-based European Audiovisual Observatory at the request of the DG Communication of the European Commission and launched in 2008. MAVISE is a tool for the television industry as parts of the database will be freely accessible to the general public. MAVISE does, however, not provide access to items in respective collections.

Multiple archives are host to educational websites specifically tailored to the national educational requirements. Eight different museums, institutions and audiovisual archives in the Netherlands, for example, developed Les 2.0, which launched in September 2010 with extensive facilities such as video editing and presentation tools.

Film and audiovisual archives have been experimenting with using the web as a means to establish a communication between their collections and their users. By giving viewers access to archival content, they can subsequently add to that collection by adding both stories and content. Wir Waren so Frei was a project run in 2009 that amassed amateur footage from the fall of the Berlin wall. The website Oorlog in Blik gathered both professional and amateur film recordings to visualise World War II.

A more recent, overseas example to put the power of online crowd sourced archiving to good use, was the Understanding 9/11 project. Launched in August, 2011, it is an online library of ‘news coverage of the events of 9/11/2001 and their aftermath as presented by U.S. and international broadcasters’, with 3.000 hours of footage from a time span of seven days from 20 different international TV news channels. Additionally, select academic analysis is brought online with the collection for further contextualisation. An attempt to preserve audiovisual history and actuality as it happened, was set into motion by the ‘Archive-It’ team, who created the Occupy Wall Street collection<sup>8</sup> as an archival approach to independent media reporting.

This online back-and-forth between users and institutions fits into a wider trend in museums to move catalogues and exhibitions online and to use open source cataloguing platforms for the creation of online exhibitions. Examples of this open technology are platforms like CollectiveAccess, that was used for the Wir waren so Frei initiative, OpenCollections and Omeka, which forms the basis for the Europeana virtual exhibitions.



### 3. EUscreen Content and Workflow

In collaboration with leading television historians EUscreen has defined a content selection policy [1], divided into three strands:

Historical Topics: 14 important topics in the European history in the 20<sup>th</sup> Century (70% of content)

Comparative Virtual Exhibitions: two specially devised topics that explore particular aspects of European history (10% of content complemented with contextualizing texts and other items)

Content Provider Virtual Exhibitions: Each content provider showcases interesting audiovisual content, supporting it with other related digital materials. The workflow for publishing the metadata related to the video items consists of a number of consecutive steps: the metadata ingestion, their transformation to a common reference schema, enrichment and finally publication as Linked Data. This is illustrated in Figure 1.

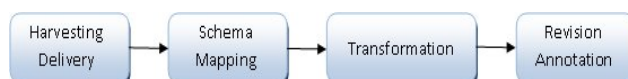


Fig 1. The ingestion workflow.

As the 20+ content providers are using different management systems and in turn different types of metadata, the need arises for interoperability. In order to achieve semantic interoperability with external web applications, a harvesting schema was implemented based on EBUCore [2], which is an established standard in the area of audiovisual metadata. An extensive evaluation of alternative standards in this area (including MPEG7, DCMI, TV Anytime) has been conducted [3] before choosing the EBUCore.

EBUCore has been purposefully designed as a minimum list of attributes that describe audio and video resources for a wide range of broadcasting applications including archives, exchange and publication. It is also a metadata schema with well-defined syntax and semantics for easier implementation. It is based on the Dublin Core to maximize interoperability with the community of Dublin Core users. EBUCore expands the list of elements originally defined in EBU Tech 3293-2001 for radio archives, also based on Dublin Core.

The MINT platform<sup>2</sup> (Metadata Interoperability Services) is used for the ingestion and transformation of the metadata. MINT is a web based platform for assisting the mapping of provider's existing metadata to the proposed metadata model. EUscreen also uses MINT for enriching metadata elements, for instance to add translations of television programme titles and to add terms from the IPTC thesaurus, as a means to support multilingual retrieval.

The final step, after the transformation of the content's metadata, is the publication of the metadata as Linked Open Data [4]. Linked data aims to make data accessible, not only to humans, but also to software agents, building a semantic layer to improve and enrich their interaction.

Content Providers use the back-end tools, while the different audiences can access the content (including versatile metadata) through the EUscreen portal, which serves as the front-end, providing different ways to explore the content (Sections 2 and 3) such as advanced search and curated exhibitions. Section 4 provides more detail regarding the project's efforts related to LOD.

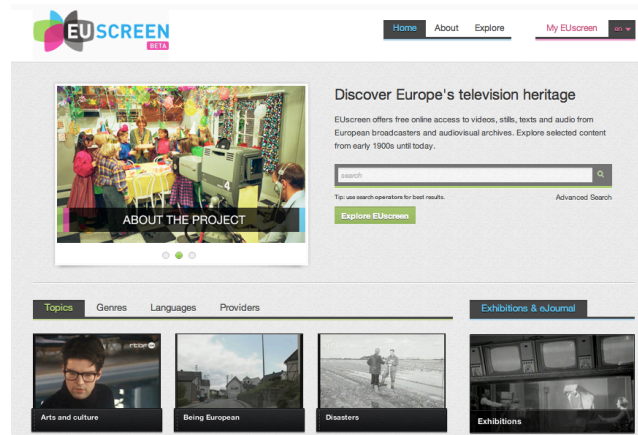
### 4. The Front-end Design

The metadata links to the media files, stored on distributed servers maintained by the project partners. Strict rules were put in place to guarantee a uniform play-out behavior of the files.

Representatives of the four primary user groups, i.e. secondary education, academic research, the general public and the cultural heritage domain were consulted in order to define the user requirements and the design of the EUscreen portal. The main challenge for the portal's front-end (Figure 2.) is to include advanced features for specific use cases without overwhelming the users with a complex interface.

---

<sup>2</sup> <http://mint.image.ece.ntua.gr>



**Fig. 2.** EUSCREEN homepage.

*Implementation of the front-end services is not done in the traditional way using server-side programming language like php, java or asp. Instead, it consists of 'server-less' front-end APIs where a JavaScript/Flash proxy system handles the communication with the back-end services. The resulting front-end system can be 'installed' on any plain HTML web server without additional server-side technologies. This means it can be hosted and moved to any location, or multiple locations, and that partners (Content Providers in particular) can use the APIs to integrate parts of the functionality in their own intranet and online systems using simple 'embed' mechanisms.*

## 4.1 The Virtual Exhibitions

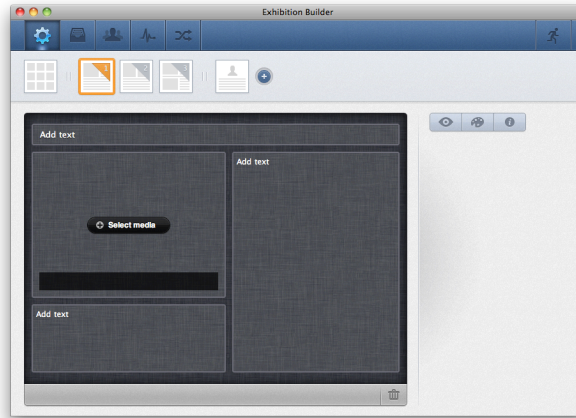
The design and implementation of the so-called Virtual Exhibitions is ongoing since summer 2011, and led by Noterik<sup>3</sup> and Aalto University School of Arts, Design and Architecture<sup>4</sup>. The idea is to provide an online tool, which allows users to create online presentations on a theme of their interest using videos and other media from EUSCREEN.

An important part of designing the exhibition tools have been design workshops, where programmers and interface designers have worked hands-on in close collaboration with scholars that curated the first exhibitions. This collaboration and further testing and feedback from the EUSCREEN content providers has enabled taking into account the needs and different perspectives regarding audiovisual storytelling. As a result, the Exhibition Builder tool has been developed, which provides a WYSIWYG style interface for creating online exhibitions (Figure 3.). With this tool, it is possible to select different layouts, add media from EUSCREEN, and to enrich it further with, for example, texts, links to external websites as well as direct links to parts of videos. The exhibitions will be published on the EUSCREEN portal and can be also embedded on external websites (see Figure 4. for a current example).

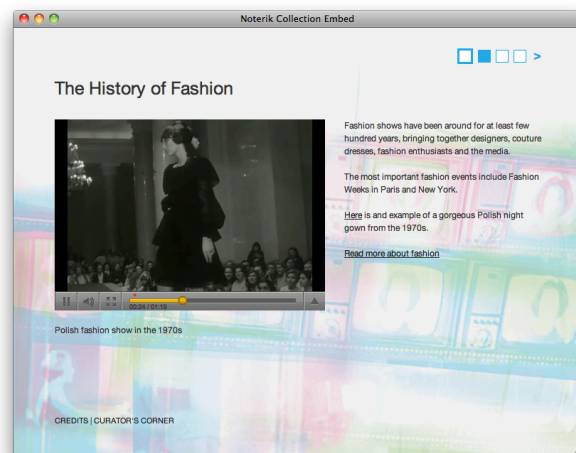
The exhibitions demonstrate the added value of curating audiovisual content from an online resource such as EUSCREEN. The first batch of exhibitions, as created by the consortium members, will be released in August 2012.

<sup>3</sup> [www.noterik.nl/](http://www.noterik.nl/)

<sup>4</sup> <http://arts.aalto.fi/en/>



**Fig. 3.** The Exhibition Builder tool



**Fig. 4.** Example of an exhibition

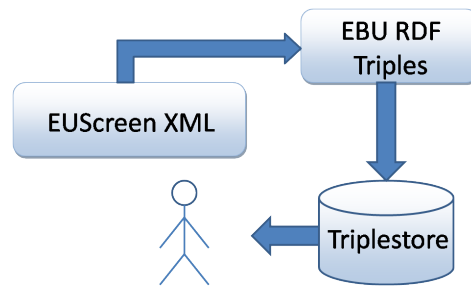
## 5. The Linked Open Data Pilot

Following the metadata format harmonisation, a Linked Open Data publication procedure has been established. This required the conversion of the harvested metadata to RDF using an expressive data model. The RDF representation of EBUcore (<http://tech.ebu.ch/lang/en/MetadataEbuCore>) was used. Finally, internal and external linking to the EUscreen content has been performed and the resulting repository was made accessible through a SPARQL query endpoint. Once available, SPARQL endpoints will need to be masked by a GUI (graphical User Interface) hiding the complexity of SPARQL queries. The GUI will provide a framework for search specific to a domain of application or user profile (e.g. public access point vs. professionals or academics)

Early 2012 EUscreen launched its so-called Linked Open Data Pilot, based on the Resource Description Framework (RDF) representation of the data, and structured it using the EBU Core ontology. Aim of the pilot is to demonstrate the added value of adhering to LOD standards in the cultural heritage domain.

The data (available under the CC0 license) served by [lod.euscreen.eu/resource/](http://lod.euscreen.eu/resource/) is generated from the metadata EUscreen has harvested using the MINT ingestion platform.

More specifically, the procedure followed is illustrated in the figure 5.



**Fig. 5.** EUScreen data transformation.

We have implemented a mapping of all the harvesting schemas' elements to classes and properties of the EBU Core ontology. The EBUCore ontology is an RDF representation of the EBU Class Conceptual Data Model (CCDM). CCDM defines a structured set of audiovisual classes (e.g. groups of resources, media resources, parts, media objects but also locations, events, persons and organizations). The EBUCore ontology also defines the semantic relationships (objectProperties) between these classes as well as properties (dataProperties) characterizing these classes. A lot of the knowledge gathered in the EBU CCDM and EBUCore RDF was used to develop the W3C Media Annotation ontology<sup>5</sup> (W3C MAWG). Reciprocally, EBUCore RDF has implemented in a subsequent version the RDF modeling options chosen by W3C MAWG.

Therefore every provider's instance that has been transformed according to the harvesting schema, is also transformed to an RDF file. This file is then stored in a triplestore repository from where the users can consume data using SPARQL.

In addition to the original data harvested by EUScreen from its data providers, the data served at [lod.euscreen.eu/resource/](http://lod.euscreen.eu/resource/) includes:

internal links to other videos that are also served by [lod.euscreen.eu/resource/](http://lod.euscreen.eu/resource/)

semantic enrichment done by NTUA, connecting EUScreen videos to structured representations of places and languages.

In particular, the EUScreen content has been enriched by its linking to external data sources like the DBpedia, Eurostat, Freebase and NY Times allowing for more expressive search and retrieval.

The store<sup>6</sup> is serving data over the HTTP protocol, using established linked data recipes, most notably the use of HTTP URIs as identifiers and entry points into data. The data can also be consumed through the SPARQL endpoint or by using the web interface of the 4store repository.

## 5.1 The EUScreen Timeline

As a showcase, EUScreen created a visualization that uses the LOD end-point. This visualization offers new possibilities for scholars and researchers to examine the EUScreen collection on a timeline (Figure 6, and also accessible online<sup>7</sup>).

Above the timeline itself, a group of links is displayed, representing all of the genres and topics defined for EUScreen. Using these links, a user can filter items in the timeline, in the way that only items of the selected genres and topics are shown. When no genres or topics are selected, all items are shown.

In order to implement this demonstrator the main task was to interface with the LOD end-point hosted as defined in the Section above. To visualize the collection data in a timeline, the (JavaScript-based) SIMILE timeline<sup>8</sup> widget is used in combination with AJAX and a Django<sup>9</sup> server application of which the latter takes care of the actual communication with the SPARQL end-point to request the metadata.

<sup>5</sup> [www.w3.org/2008/WebVideo/Annotations/](http://www.w3.org/2008/WebVideo/Annotations/)

<sup>6</sup> [lod.euscreen.eu/](http://lod.euscreen.eu/)

<sup>7</sup> [http://blog.euscreen.eu/?page\\_id=3117](http://blog.euscreen.eu/?page_id=3117)

<sup>8</sup> <http://www.simile-widgets.org/timeline/>

<sup>9</sup> <https://www.djangoproject.com/>

## 6. Conclusions and future work

In this paper, we have summarized the main technical results of the EUScreen project so far, and showed initial results in the areas of visualization and Linked Open Data. In future work, we will work on:

advanced metadata and ontology enrichment to be provided through entity extraction (e.g. names, places) from existing metadata.

Assessing potential of time-based cross-linking with content on Europeana.

In parallel to project-based activities, we hope that the developer community takes advantage of this unique and rich resource of European television heritage. EUScreen strongly believes that providing access to their content by exploiting LOD will be an important driver for innovation.

In relation to the virtual exhibitions, we have identified two areas of work that future work will be focused on:

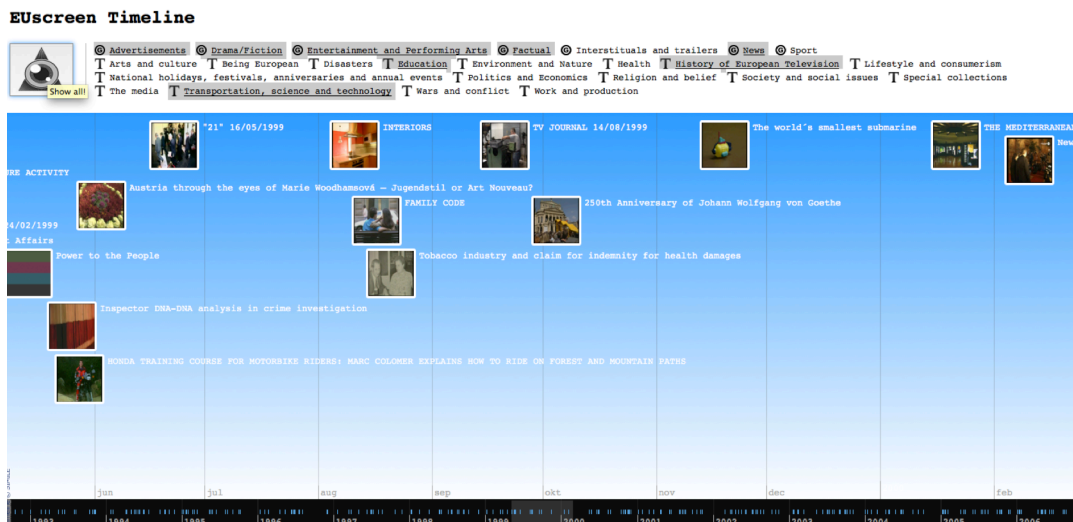
Exception-based IPR. Allowing content owners to set the rights of reuse in flexible ways per collection, video or even parts of video's.

Content-based Editing. Editing video not based on time but by adding time-based entity information and concepts so it would be possible to shift the moment of editing from production moment to consuming moment and allow for multiple on montages still reflecting the wishes/intent of the creators.

This work will be executed in the context of the three-year project EUScreenXL, starting start in 2013. We anticipate these alternative item discovery methodologies will attract new user-groups to engage with the rich body of content offered by the archives represented in the EUScreen consortium.

## References

1. Kaye, L. (2011) "Content Selection and Metadata Handbook", Consulted August 12, 2012. Available: [http://blog.euscreen.eu/wp-content/uploads/2010/10/Content-Selection-and-Metadata-Handbook\\_public.pdf](http://blog.euscreen.eu/wp-content/uploads/2010/10/Content-Selection-and-Metadata-Handbook_public.pdf)
2. Evain, Jean Pierre (ed.) (2009) "EBU Core Metadata Set EBU", Consulted August 12, 2012. Available: [http://tech.ebu.ch/docs/tech/tech3293v1\\_3.pdf](http://tech.ebu.ch/docs/tech/tech3293v1_3.pdf)
3. Schreiber, G (2010). "Metadata Models, Interoperability Gaps, and Extensions to Preservation Metadata Standards"
4. Buerger, T. & Evain, J-P. & Champin P-A. (2011), "W3C Media Annotation Working Group RDF ontology". Consulted August 12, 2012. Available: <http://www.w3.org/ns/ma-ont.rdf> (see also <http://www.w3.org/TR/2011/PR-mediaont-10-20111129/>).





# Providing Suitable Information Accesses for New Users of Digital Libraries – Case Study DMG-Lib

Rike Brecht, Torsten Brix, Ulf Döring, Sascha Falke, Michael Reesing

Ilmenau University of Technology, Ilmenau, Germany  
{r.brecht, torsten.brix, ulf.doering, sascha.falke,  
michael.reessing}@tu-ilmenau.de

**Abstract.** In today's information society large collections of data arise by the digitization of various content sources. For example, the DMG-Lib, a digital library to collect, preserve and present the knowledge of mechanism and gear science, contains a wide range of digitized information resources in very heterogeneous media types. A main issue of all these collections is to make their contents available for new users in an adequate way. In case of DMG-Lib, the resources are enriched with various additional information like animations and simulations. Combined with innovative multimedia applications and a semantic information retrieval environment, the DMG-Lib provides an efficient access to this knowledge space of mechanism and gear science, for both experts and laymen.

**Keywords:** Information retrieval, Exploration strategies, Digital libraries

## 1 Introduction

Digital libraries around the world are many and varied. The growth in digital libraries including collections of e-books, pictures, audio and video sequences has been phenomenal in recent years [6]. Digital libraries in practice very often include multimedia content but just a few try to link and enrich the documents in order to support information retrieval and users' orientation in the multimedia information space. Furthermore the common storage method for knowledge, mainly in static texts and images, does not comply with requirements concerning an efficient and quick information retrieval.

The Digital Mechanism and Gear Library (DMG-Lib) is a scientific digital library with the objective of collection, integration, preservation, systematization and adequate presentation of information in the field of mechanism and machine science [1, 2]. Currently, the DMG-Lib provides more than 40.000 items of various types like documents, images, movies, interactive animations, CAx models, etc. and the amount of items is still growing with the thinkMOTION project [7].

To offer users a wide variety of opportunities for retrieval and utilization beyond the standard keyword-based search paradigm, the digitized sources are extensively post-processed and enriched as well as cross-linked with various information e.g.

animations, metadata, references and constraint based models. This huge amount of available heterogeneous information sources in the DMG-Lib calls for the implementation of an efficient, uniform and user-centered information retrieval, specific concepts to enrich and cross-link the information as well as adequate tools to work with the information presented.

## **2 Accessibility of Information**

The worldwide information about mechanisms and machines is mostly scattered, difficult to access and does not comply with today's requirements concerning a rapid information retrieval. Most important problems are the amount of information originating from different sources like books, technical reports, functional models or mechanical drawings and the efficiency to retrieve and work with this information.

The DMG-Lib is going to digitize the different information sources so that this still important knowledge becomes again accessible for the public and to offer users a wide variety of opportunities for retrieval and use. In order to meet the users' needs the digital library design is based on the requirements of different user groups like engineers, scientists, teachers, students, librarians, historians and others.

According to the Usability Engineering Lifecycle developed by Deborah J. Mayhew [3] a requirement analysis including expert interviews and user focus groups have been carried out to develop a conceptual model of the DMG-Lib portal. For a user adequate design and implementation the usability of the portal and its modules was evaluated analytically by usability experts as well as empirically by users in several stages of development.

## **3 Suitable Information Accesses for New Users**

The DMG-lib provides access to the mechanism and gear technology for different user groups with specific requirements each on the design of the online portal. For example, engineers are looking for a task specific mechanism, scientists are researching for literature, students are looking for teaching materials, and laymen expect a comprehensible introduction to the subject.

Therefore, different information accesses to the collected information within the portal are offered. Keyword-based searches on all sources or source-specific are possible as well as browsing in the database. Especially for new users of the DMG-Lib a timeline, a virtual museum, hypervideos, history maps, interactive e-books with animations and topic maps are offered, to give them an insight and overview of the diversity of the collection.

### **3.1 Timeline**

The timeline application gives users a multimedia overview of important persons, inventions and publications in the historical development of mechanism and gear



science. Users can switch directly to corresponding information resources in the library, for example available books of selected persons. The timeline offers a chronologically order of information and a new opportunity to explore the collection.

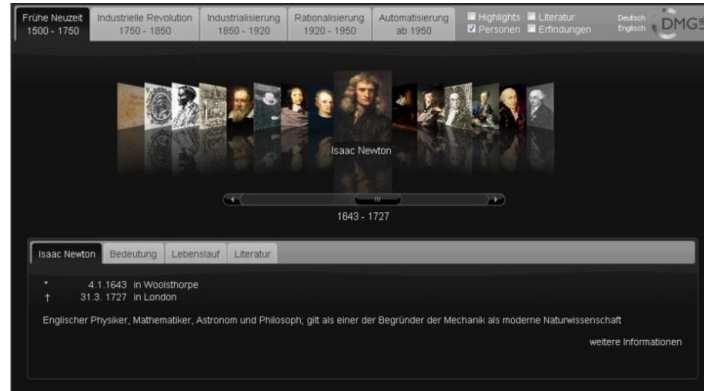


Fig. 1. Timeline application within the DMG-Lib

### 3.2 Virtual Museum

The virtual museum of the DMG-Lib highlights outstanding items. Thematically different rooms inform about the most important inventions, the development of technology and significant persons of mechanism and gear science. So, users explore rare and bizarre mechanism models or discover mechanisms hidden in daily life.



Fig. 2. Virtual museum within the DMG-Lib

### 3.3 Hypervideo

A hypervideo is a video stream with embedded hyperlinks on different areas which allow to navigate in the video or to switch to other applications. With the concept of hyper videos an access for new users from the everyday view is created to discover and experience the collection of the DMG-Lib.

The user discovers everyday objects within the videos and takes notice of contained mechanisms by additional information included as hyperlinks. These mechanisms are not obvious in most devices and their principles of operation are little or not known to the user. Therefore, new users find an entrance into the world of mechanism and gear science from a very practical point of view through the hyperlinks to the heterogeneous information sources in the DMG-lib.

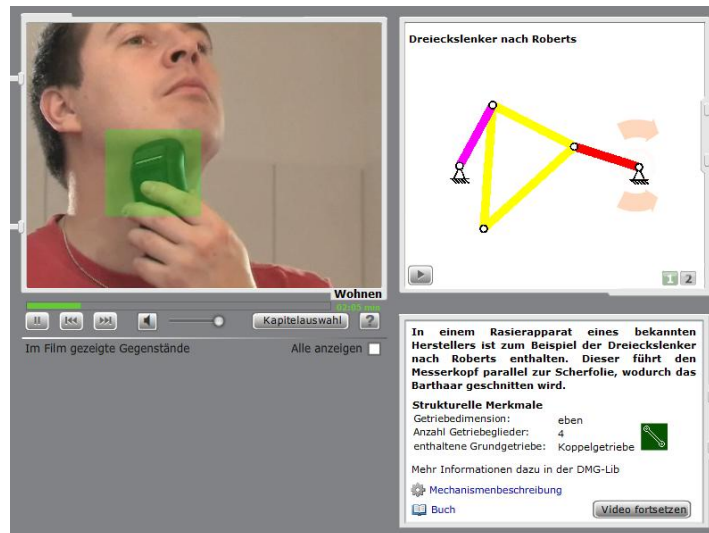


Fig. 3. Hypervideo within the DMG-Lib

### 3.4 History Maps

The collection of DMG-Lib contains several items described with time and location-based metadata. For example, biographies contain place and date of birth as well as important stages of life career. Mechanism descriptions contain information on date of manufacturing or the current location of collections. Metadata belonging to text documents include the publication date and place. These metadata can be presented in several maps. For example, DMG-Lib history map illustrates a person's road of live (Fig. 4) or shows the locations of all mechanism collections in Germany. The map can also be used to locate search results within the map in addition to a result list.

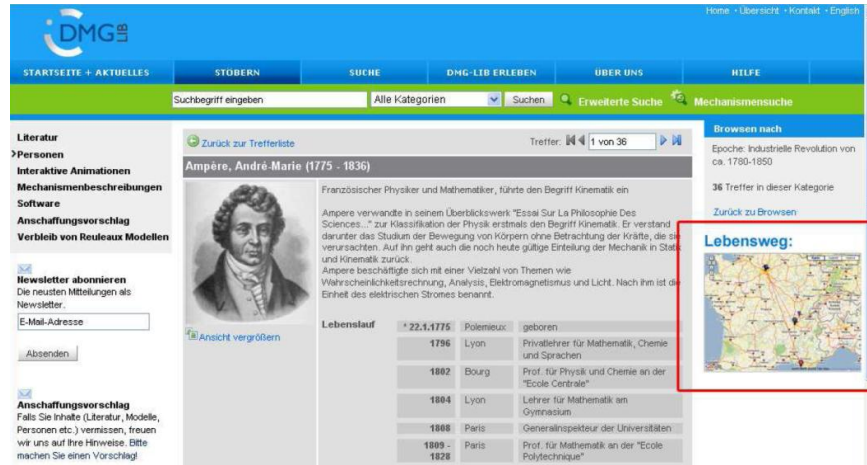


Fig. 4. History map within the DMG-Lib

### 3.5 Interactive e-books with animations

In the DMG-Lib thousands of e-books are available which contain many figures of mechanisms. To show the movement of the mechanisms selected figures are animated. The animations are integrated in the e-books as overlays. Users can read the text, watch the animation simultaneously and interact with it. Thus, the users have the possibility to gain a better understanding of the mechanism.

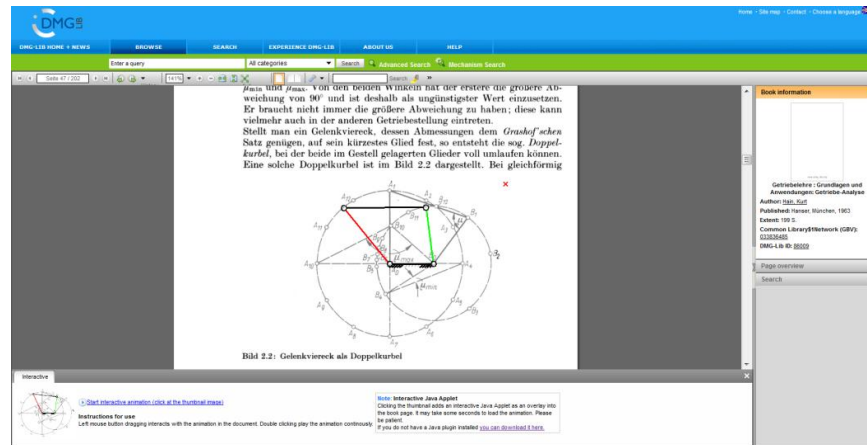


Fig. 5. Interactive e-book with animations within the DMG-Lib

### 3.6 Topic Maps

A further field of research is the retrieval in heterogeneous information resources using different mechanism and gear hierarchies like the structural system of Reuleaux [5]. Visualization and efficient navigation over these different categories of gears can help users to get a systematic overview over the huge amount of existing mechanism and gear constructions. The stored knowledge can be generalized and explicit modeled in a semantic meta-layer, and visualized with the help of topic maps [4]. Hereby, different hierarchies and relations can be modeled, valid contexts, alternative names and other relevant semantic information can be included, and all relevant information resources available in the DMG-Lib can be linked. This enables users to decide which structuring system he wants to use for navigation.

## 4 Conclusion

In this paper the various information accesses for different user groups to the DMG-Lib, a digital and interactive library for mechanism and gear science, are presented. Based on the vast amount of available heterogeneous information resources in the library and the extensive enrichment, the DMG-Lib is able to provide an effective retrieval as well as various utilization options for users.

The DMG-Lib supports engineers and scientists in finding ideas for design solutions, detailed structural and functional descriptions, adequate technical terms, historic publications in different languages, etc. Additionally, alternative information accesses help new users to explore the heterogeneous information resources in the library and to discover interests in mechanical engineering.

The DMG-Lib envisions itself as an example of a modern knowledge space aiming at one of the key tasks in today's information society – satisfying user's needs for getting an effective, efficient and custom-tailored access to required information.

## References

1. Brix, T., Brecht, R., Henkel, V., Reeßing, M.: DMG-Lib – An Open Access Digital Library as an Exploration Tool for Historians and Engineers. In: Koetsier, T., Ceccarelli, M. (eds.) *History of Mechanism and Machine Science*. LNCS, vol. 15 part 6, pp. 579-592. Springer Netherlands (2012). DOI: 10.1007/978-94-007-4132-4\_40
2. Digital Mechanism and Gear Library, <http://www.dmg-lib.org>
3. Mayhew, D.J.: *The Usability Engineering Lifecycle: A Practitioner's Handbook for User Interface Design*. Morgan Kaufmann, San Francisco (2008)
4. Park, J., Hunting, S.: *XML Topic Maps: Creating and Using Topic Maps for the Web*. Addison-Wesley, Boston (2003)
5. Reuleaux, F.: *Lehrbuch der Kinematik*. Vieweg, Braunschweig (1875)
6. Tedd, L.A., Large, J.A.: *Digital Libraries: Principles and Practice in a Global Environment*. Saur, München (2005)
7. thinkMOTION – DMG-Lib goes Europeana, <http://www.thinkmotion.eu>

# Exploring Pelagios: A Visual Browser for Geo-Tagged Datasets

Rainer Simon<sup>1</sup>, Elton Barker<sup>2</sup>, and Leif Isaksen<sup>3</sup>

<sup>1</sup> AIT - Austrian Institute of Technology, Donau-City-Str. 1,  
1220 Vienna, Austria

`rainer.simon@ait.ac.at`

<sup>2</sup> The Open University,  
Milton Keynes, United Kingdom

`e.t.e.barker@open.ac.uk`

<sup>3</sup> University of Southampton,  
Southampton, United Kingdom

`l.isaksen@soton.ac.uk`

**Abstract.** When dealing with networked information spaces, there are two challenges a data navigation system must address: first, it needs to provide the user with a quick high-level overview of the nature and structure of the data. Second, it must offer mechanisms to easily *drill down* and navigate to specific detail information. In this demonstration, we present the *Pelagios Graph Explorer*, a prototype visual exploration tool for place metadata which is derived from ancient world research datasets brought together by the Pelagios Project. Furthermore, we report on ongoing work on the Pelagios API. The API enables external re-use of the Pelagios place metadata, and provides a basis for future development of - and experimentation with - alternative navigation tools for the networked place metadata from the Pelagios project.

**Keywords:** Linked Data, Visual Browsing, Information Exploration, Digital Libraries, Digital Humanities, Geoinformation

## 1 Introduction

On-line resources that reference ancient places are multiplying rapidly, bringing huge potential for the researcher - provided that they can be found. But users currently have no way of easily navigating between such resources, or of comparing their contents. The Pelagios<sup>4</sup> consortium is a growing international collective of ancient world projects that aims to address the problems of discovery and reuse with the twin aims of helping digital humanists to make their data more discoverable, and of empowering real-world users (scholars and the general public) to find information about particular ancient places and visualize it in meaningful ways. While the project focuses on the ancient world, the methodology and tools developed will be of interest to anyone working with data

<sup>4</sup> <http://pelagios-project.blogspot.co.uk>

containing spatial references. The Pelagios collaboration intentionally includes partners maintaining a wide range of different document types, including texts, maps and databases<sup>5</sup>. In doing so we take some of the first steps required to build a Geospatial Semantic Web for the Humanities [4].

## 2 Connecting Data Through Place

Pelagios aims to create links between data entities on the Web based on the places they refer to. It does not impose any restrictions on the type and format of data entities themselves; nor on the metadata schemas or media types used. Instead, it provides a lightweight framework of principles and requirements with which to express place references as additional metadata.

The first, and essential, requirement is that when partners refer to a place, they do so according to a shared reference system: in our case, this reference system is the *Pleiades Gazetteer of the Ancient World*<sup>6</sup>, which provides Uniform Resource Identifiers (URIs) for more than 34.000 places in the ancient world. By expressing each place reference in the form of a Pleiades URI, the otherwise disjointed datasets are joined up to an interconnected graph, with Pleiades as the central backbone.

The second requirement is a syntactical requirement: partners need to publish their resulting place metadata online according to a common data representation format, and common vocabulary terms. The representation format used in Pelagios is RDF<sup>7</sup>; the vocabulary used to model place references is the *Open Annotation* vocabulary [5]. The idea behind considering a place reference an “annotation” is that the reference should not be considered certain, but rather that a statement is being made about some kind of relation between the data and the place. Additional descriptive metadata about the dataset as a whole should be published using the *Vocabulary of Interlinked Datasets (VoID)*<sup>8</sup>. This information will typically include things such as textual descriptions of the data, or publisher and license information.

It is important to note that the additional place metadata required by Pelagios can be completely separate from the origin data itself. This approach is sometimes referred to as “standoff markup” [8], and helps to avoid data management problems which would arise if participants were required to adapt or extend their internal metadata schema for Pelagios.

<sup>5</sup> At time of writing, a growing list of Pelagios partners included: Arachne, The British Museum, CLAROS, Fasti Online, Google Ancient Places, Inscriptions of Israel/Palestine, Meketre, Nomisma, OCRE, Open Context, Oracc, Papyri.info, Perseus, Ports Antiques, Ptolemy Machine, Regnum Francorum Online, SPQR, The Ure museum. For an up-to-date list, see: <http://pelagios-project.blogspot.co.uk/p/about-pelagios.html>

<sup>6</sup> <http://pleiades.stoa.org>

<sup>7</sup> <http://www.w3.org/RDF/>

<sup>8</sup> <http://www.w3.org/TR/void/>

### 3 Exploring the Pelagios Network

From a conceptual point of view, the Pelagios network consists of so-called *datasets*. Datasets are collections of *place references*: statements that express the fact that a particular data object (e.g. a record in an archaeological database, a segment of text in an online book archive, an image in a collection of digitized photographs, etc.) bears some kind of relation to a particular place. Datasets may also be structured into *subsets*, and carry additional metadata such as a title, a description, or license information. In implementation terms, the datasets are defined through the *VOID* metadata.

It is important to stress that Pelagios datasets do not hold the data itself (or copies of it). They only consist of the place references plus a minimum of metadata about the data related to the place. Implementation-wise, this information is expressed via the annotations.

#### 3.1 Pelagios Graph Explorer

The *Pelagios Graph Explorer* [7] is a research prototype that allows users to explore the Pelagios network of place references using visual metaphors. Simple specific queries the Graph Explorer can answer are, for instance:

- Which datasets reference a particular place?
- Which places are referenced in a particular dataset?
- What is the geographical footprint of a particular dataset?

Results are presented in combinations of map- and graph-based views. Views are synchronized and reveal more information as the user interacts with them. For example, hovering the mouse over a dataset in the graph view will show additional details about it, and overlay its geographical footprint on the map.

The key goal of the Graph Explorer, however, is not only to provide answers to questions that relate to single specific places or datasets. Instead, the Graph Explorer aims at, first, providing users with a sense of the overall “connectivity” within the Pelagios network of data and places and, then, enabling them to drill down into specific details. The user interface separates this into two different modes of exploration: one mode is dedicated to exploring *connections between places, as established through data*. The other mode is dedicated to the “inverse” task of exploring *connections between data, as established through places*.

Fig. 1 shows a screenshot of the first exploration mode. In this mode, the user enters names of places in the search field. The Graph Explorer will produce an interactive network visualization that shows how these places are connected to each other through different datasets. In the graph view, datasets are represented as bubbles. Bubbles are colour-coded according to their root dataset: i.e. datasets that are subsets (or sub-subsets, etc.) of the same parent set share the same colour. Bubble size represents the total amount of place references contained in the dataset: the bigger the bubble, the larger the dataset. Likewise, the size of the connecting lines between places and datasets is scaled according to the number



Fig. 1. Exploring Places Related Through Data.

of references to that particular place, in that specific dataset. As a consequence, a connecting line that has a width close (or equal) to the size of the dataset bubble will indicate a case where most (or all) of the references in the dataset are to this one place, and only few (or no) other places are referenced in this dataset. This way, users can quickly identify datasets that are primarily about this place, versus datasets where the place is referenced in isolated cases only. If, in addition to a relatively wide line, the bubble is small (representing a dataset with a low number of place references total), users will furthermore know that they are looking at a very specific, small set of data – such as a single page in a book – primarily about this place; and that this dataset could be worth looking at in more detail. A click on any dataset will bring up a list of the place references and accompanying metadata (including links to the actual origin data) in the *Data View* side panel on the left.

Fig. 2 shows a screenshot of the second exploration mode. In this mode, users can explore all datasets in the graph, using a similar representation. Datasets are again visualized as colour- and size-coded bubbles. Double-clicking on a dataset reveals its subsets, allowing the user to navigate through the dataset hierarchy. Similar to the first exploration mode, hovering over a dataset shows additional information about it, as well as its geographical footprint on the map. The user can select a dataset with a single click to get detail information about the places referenced in it. To provide the user with a sense of how different datasets correspond to (or overlap with) each other in terms of the places they reference, it is possible to select multiple datasets in the graph view. The view will generate connecting lines between the selected sets which are scaled according





The Pelagios API is currently undergoing a major upgrade to offer increased performance and enhanced scalability. A publicly accessible beta version is available at <http://pelagios.dme.ait.ac.at/api>.

## 4 Future Work

In its present state, the Pelagios Graph Explorer is a research prototype. It has undergone a first round of user evaluation [6]. Further work has been carried out with regard to tailoring user interfaces built on top of the Graph Explorer or the API towards the needs of specific audiences. An important criterion in this regard has been the distinction between *super users* – Website owners who build, own or maintain place-related web pages, web sites or blogs in the ancient history domain; and *end users* – e.g. ancient history researchers, teachers and students with expertise in the ancient history domain [3]. Ongoing and future work on the user-facing aspects of Pelagios will focus on refinements in this regard.

Furthermore, Pelagios partners and third parties (such as the *Open Context*<sup>9</sup> portal for primary field research data from archaeology and related disciplines, or the *Ancient World Linked Data* JavaScript library project *awld.js*<sup>10</sup>) are starting to make use of the Pelagios API in their own applications. This real-world exposure will be crucial to gather valuable feedback on the use and utility of our toolset, and to guide its future development.

## References

1. Bizer, C., Cyganiak, R., Heath, T.: How to publish Linked Data on the Web (2007), <http://www4.wiwiw.fu-berlin.de/bizer/pub/LinkedDataTutorial/>
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems* 5(3), 1–22 (2009)
3. Culver, J.: Pelagios project blog: Developing widgets for pelagios (March 2012), <http://pelagios-project.blogspot.co.at/2012/03/developing-widgets-for-pelagios.html>
4. Harris, T.M., Rouse, L.J., Bergeron, S.: The Geospatial Semantic Web, Pareto GIS, and the Humanities. In: Bodenhamer, D.J., Corrigan, J., Harris, T.M. (eds.) *The Spatial Humanities: GIS and the Future of Humanities Scholarship*, pp. 124–142. Indiana University Press, Bloomington, IN (2012)
5. Haslhofer, B., Sanderson, R., Simon, R., Sompel, H.: Open annotations on multimedia web resources. *Multimedia Tools and Applications* pp. 1–21 (2012)
6. Ridge, M.: Pelagios project blog: Pelagios usability testing results (September 2011), <http://pelagios-project.blogspot.co.at/2011/09/draft-results.html>
7. Simon, R.: Pelagios project blog: The pelagios graph explorer: A first look (August 2011), <http://pelagios-project.blogspot.co.at/2011/08/pelagios-graph-explorer-first-look.html>
8. Thompson, H.S., McKelvie, D.: Hyperlink semantics for standoff markup of read-only documents. In: *Proceedings of SGML Europe 97: The next decade Pushing the Envelope*. p. 227229 (1997)

<sup>9</sup> <http://opencontext.org>

<sup>10</sup> <http://isawnyu.github.com/awld-js/>