# EDB: Knowledge Technologies for Ancient Greek and Latin Epigraphy

Fabio Fumarola[1], Gianvito Pio[1], Antonio E. Felle[2], Donato Malerba[1], and Michelangelo Ceci[1]

[1] Dipartimento di Informatica, Università degli Studi di Bari "A. Moro"
via Orabona, 4 - 70126 Bari, Italy
{fabio.fumarola,gianvito.pio,donato.malerba,michelangelo.ceci}@uniba.it
[2] Dipartimento di Scienze dell'Antichità e del Tardo Antico,
Università degli Studi di Bari "A. Moro"
strada Torretta (Città Vecchia) - 70122 Bari, Italy
antonio.felle@uniba.it

**Abstract.** Classical Greek and Latin culture is the very foundation of the identity of modern Europe. Today, a variety of modern subjects and disciplines have their roots in the classical world: from philosophy to architecture, from geometry to law. However, only a small fraction of the total production of texts from ancient Greece and Rome has survived up to the present days, leaving many ample gaps in the historiographic records. Epigraphy, which is the study of inscriptions (epigraphs), aims at plug this gap. In particular, the goal of Epigraphy is to clarify the meanings of epigraphs, classifying their uses according to dates and cultural contexts, and drawing conclusions about the writing and the writers. Indeed, they are a kind of cultural heritage for which several research projects have recently been promoted for the purposes of preservation, storage, indexing and on-line usage. In this paper, we describe the system EDB (Epigraphic Database Bari) which stores about 30,000 Christian inscriptions of Rome, including those published in the *Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series* editions. EDB provides, in addition to the possibility of storing metadata, the possibility of *i)* supporting information retrieval through a thesaurus-based query engine, *ii)* supporting time-based analysis of epigraphs in order to detect and represent novelties, and *iii)* geo-referencing epigraphs by exploiting a spatial database.

**Keywords:** Epigraphy, Information Retrieval, Knowledge Bases, Novelty Detection, Spatial Databases.

## 1 Introduction

Many countries are nowadays interested in the valorization of the cultural heritage, since it is widely recognized that cultural heritage resources have significant implications for development (both as a knowledge basis and in terms of commercial exploitation). For this aim, many institutions which collect and

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 29–40, 2014.

**Fig. 1.** An example of epigraph

preserve cultural heritage have shown a great interest in the digitalization of their resources and in the exploitation of mechanisms to provide online access to digitalized products.

According to the definition reported in the 1972 UNESCO "World Heritage Convention" - Article 1 - cultural heritage refers to "monuments", "groups of buildings" and "sites" which are of outstanding universal value historically, artistically or scientifically. However, the concept of cultural heritage has recently assumed a broader connotation and includes, among other things, tangible, moveable objects such as works of art, artifacts, scientific specimens, photographs, books, manuscripts and recorded moving image and sound [2].

In the literature, several systems have been proposed for the analysis, also through knowledge technologies, of digital Cultural Heritage resources and metadata. Typically, they are developed in the context of research projects such as MASTER [13], MEMORIAL [4], D-SCRIBE [11][12], CULTURA [1], PROMISE [10], COLLATE [9] and CDLI (`http://cdli.ucla.edu`). They are either general purpose projects or focused on specific types of artifacts (e.g. COLLATE focuses on digitalized film archives of the Second World War) but none of them take into account the specific case of epigraphs.

Epigraphs are invaluable sources of information that provide us with a myriad of useful information of the past (an example of epigraph is shown in Figure 1). They play the role of "time capsules" for example by allowing us to shed light on otherwise undocumented historical events, or to gain new knowledge of local laws and customs, and even to determine the date and producer of a given piece of lead piping. Epigraphy also documents the evolution of languages and scripts, although indirectly. In some cases, such as that of the Rosetta Stone, it can provide those key insights that allow for the successful deciphering of an unknown script.

In recent years, the tendency to create epigraphic databases for storing both images of epigraphs and associated metadata, as well as for supporting retrieval functionalities has emerged [8]. Examples are [6] and, more recently, [3] which exploit the EPIDOC schema [5] to guarantee uniform representation of epigraphic metadata. In the particular case of [3], the authors propose the *Hispania*

*Epigraphica database* which allows the representation and the exploitation of semantic links between entities.

However, four main problems have, up to now, affected epigraphic databases:

1. Retrieval capabilities should take into account possible evolutions of the language, possibly due to the influence of other languages (e.g. in the Middle age) which lead to *aberrant* forms. Preserving retrieval effectiveness in these cases requires the consideration of a thesaurus which plays the role of background knowledge to be used for retrieval purposes.
2. Classic epigraphy has evolved into three strictly separate disciplines, i.e. Greek, Latin and Christian epigraphy, characterized by separate collections and corpuses used for reference, separate publications, separate populations of scholars and researchers. Indeed, different languages require different background knowledge to be exploited.
3. Data available in the databases can be used to extract knowledge through the application of data mining algorithms (following previous studies that use data mining algorithms for the analysis of digitalized cultural heritage resources [7]).
4. Epigraphs have traditionally been featured in non-geographical data bases. This results in the fact that several inscriptions that should be linked because of thematic or historic commonalities are scattered across multiple collections. Geo-referencing would help to overcome this limitation.

In this paper, we present EDB (Epigraphic Database Bari) which concentrates on the valorization of the huge Italian cultural heritage and, in particular, on the valorization of Christian inscriptions in Rome. EDB actually stores around 30,000 Christian inscriptions and provides an answer to the problems described before. It stores metadata, such as the type of support (e.g. marble), the approximate period, the engrave technique, the current position and the text. Moreover, in the retrieval phase it expands queries by exploiting a thesaurus which includes more than 3000 relationships between terms. It also integrates a data mining algorithm which faces a novelty detection task. In this way, it is possible to identify relevant (frequent) changes in the properties of the inscriptions over time. Finally, it embeds a spatial database used to geo-reference epigraphs.

The paper is organized as follows. In the next section, we describe the system architecture of EDB. In Section 3, we describe the application of EDB to the inscriptions published in the *Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series* editions. Finally, in Section 4, we report conclusions and delineate some future work.

## 2 System Architecture

The general architecture of the proposed system consists of several components, each of which is in charge of performing specific tasks, and of a set of different data sources (see Figure 2). A summarized description of each of them is reported in the following subsections.

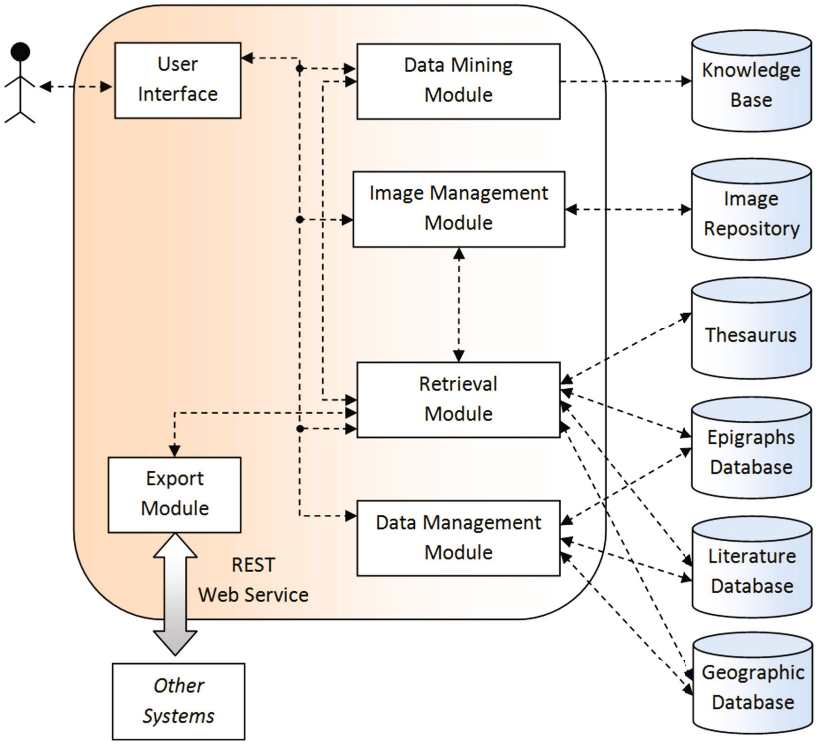



**Fig. 2.** System architecture

### 2.1 Data Sources

Since the main goal of the system is to store and retrieve data about epigraphs, the main data source is the *Epigraph Database*, which, together with the text of the epigraphs, stores a set of metadata about dating, original context, current location, related literature, etc.

The *Literature Database* stores metadata about scientific papers in which epigraphs have been studied. These metadata can include, but are not limited to, the authors, the journal, the year of publication, the citations, etc. Each epigraph stored in the Epigraph Database can be associated to one or more scientific papers stored in this database.

The *Knowledge Base* stores the knowledge extracted by the Data Mining Module, i.e. patterns of interest in the form of rules, clusters, etc.

The *Image Repository* stores photos of the epigraphs. This repository can be either internal or external, as well as an integration of internally produced and external resources.

The *Thesaurus* data source is useful to support retrieval tasks. In particular, it gives the possibility to enhance and/or expand the user's query to better match data stored in the epigraph database. This aspect is detailed later in the Section 2.3.

The *Geographic Database* stores data about geographic positions. It can be either an internal or an external resource (e.g. a web service). One or more geographic positions can be associated to each epigraph, which can represent either the locations where the epigraph (or an its fragment) was found or the position where the epigraph is currently located.

### 2.2 System Components

The *User Interface* component allows users to interact with the system. In particular, this component is in charge of translating each user action to a request to other system components and to properly show the returned results. Although in Figure 2 a single actor is reported, it is noteworthy that different types of users may interact with the system. In particular:

- **Administrators**, which access the system to manage users and services;
- **Compilers/Epigraphists**, which are the domain experts in charge of inserting, editing and deleting data about epigraphs to/in/from the main database, as well as of managing the image repository and the literature related to the epigraphs;
- **Web users**, which are mainly interested in retrieving information about archived epigraphs according to many different filtering criteria;
- **Data Analysts**, whose goal is to analyze data in order to extract valuable knowledge from them.

The *Retrieval Module* exposes an interface to query the Epigraph Database. This is the central component of the system, through which users, as well as other components of the system, can access data to perform their own tasks.

In particular, the main goal of this component is to retrieve epigraphs which satisfy a given set of filtering criteria. Such filters can be defined on the epigraphs' text, as well as on their metadata, which include the dating, the related literature and the geographic position about their original context or about their current location. Therefore, this component has to access directly to the Epigraph Database, to the Literature Database and to the Geographic Database. Furthermore, since performing the retrieval on the basis of the epigraph's text could be tricky when aberrant forms are present (see Section 2.3), this component also exploits the Thesaurus data source.

The *Export Module* offers a service to other systems that need to access to information about epigraphs, acting as a bridge between them and the Retrieval Module.

The *Data Management Module* performs the tasks which are mainly related to the activity of compilers. In particular, this component allows the users to insert, edit and delete data about the epigraphs and manage all their aspects, such as the related literature and the geographic positions associated to the original context or to the current location of conservation.

The *Image Management Module* allows the users (mainly compilers) to manage, i.e. insert into, editing or deleting from, the image repository. Part of the

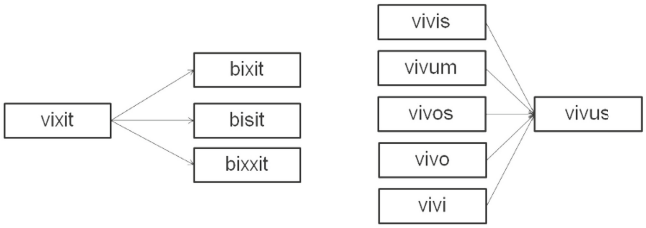



**Fig. 3.** On the left, an example of aberrant forms of the term "vixit". On the right, a set of terms which can be mapped to the same term "vivus".

image repository is obtained via web queries to the *Pontificia Commissione di Archeleogia sacra* (PCAS) web site[1].

The *Data Mining Module* allows the users (mainly data analysts) to execute data mining algorithms on the available data, in order to discover valuable knowledge, which is then stored in the Knowledge Base. This component accesses data through the Retrieval Module. An example of data mining task which can be applied to epigraphs is reported in Section 2.4.

### 2.3 Dealing with Aberrant Forms through a Thesaurus

Among all the possibilities that a user can exploit to find an epigraph of interest, the text-based search on the inscription is one of the most straightforward way provided by the Retrieval Module.

However, in this particular domain, a simple matching strategy between the query and the text of the inscriptions stored in the database can easily fail, since *i)* the same term could have changed in its phonetic or orthography over time (aberrant forms) and *ii)* many different terms semantically map to the same concept (see Figure 3).

In this context, it is useful to design a thesaurus containing the relations between aberrant forms and normal forms as well as the sets of terms which map to the same concepts. In this way, the quality of the results returned by the Retrieval Module can be substantially improved, since it can easily map each term contained in the query and in text of the inscriptions to its normal form (and/or to the term which express the general concept), before verifying the matching between the query and the inscription.

### 2.4 Novelty Detection

In this subsection we report one of the possible data mining tasks, called *novelty detection*, which can be applied on data about epigraphs. In particular, the main goal of this task is to identify emerging patterns, that is, patterns which show relevant changes (in frequency) over time. Therefore, this task is strongly related to the temporal dimension associated to the epigraphs.

[1] `http://pcas.xdams.net/pcas-web/home.html`

Since epigraphs can usually be associated to an historical period (dating), even to a single year or to a definite time interval, it could be interesting to identify how social and cultural changes over time have affected the epigraphs, in the phonetics or in the orthography as well as in the used materials or in the executing techniques.

This task requires to deal with several pre-processing issues. In particular:

- Identification of the *reference objects* and of the *task-relevant objects*. In this case, target objects are clearly the epigraphs, while the task relevant objects are the materials, the executing techniques, the different kinds of writing, etc.
- Definition of proper time intervals (or time windows), which consists in the identification of an adequate number of intervals and in the choice of the discretization method to apply (e.g. equal width, equal frequency, clustering-based discretization).
- Feature selection, that is, identification of features of interest among all the available ones, in order to focus the algorithm only on the relevant data.

In the following we report two examples of possible emerging patterns, expressed as a list of logical predicates, which could be discovered by applying novelty detection algorithms to data about epigraphs.

$$[250, 349] \rightarrow [350, 399] : epigraph(E), transcription(E, T), term(T, \text{“vixit”}) \quad (1)$$

$$[250, 349] \rightarrow [350, 399] : epigraph(E), pertinence\_area(E, \text{“Via Appia”}) \quad (2)$$

In the example (1), the discovered pattern describes a relevant increase of the number of epigraphs containing the term “vixit” in their transcription, in the time interval [350, 399] with respect to the time interval [250, 349]. The pattern in the example (2) emphasizes an increased amount of epigraphs in the area of “Via Appia”, in the same period.

The discovered patterns can be ranked according to some measures of relevance, such as the *growth rate*, which represents the relative variation of the support of the pattern in the considered time intervals.

It is noteworthy that the discovered patterns can suggest the researchers some relevant aspects that are worth to be deeply investigated, since they can describe social and cultural changes otherwise difficult to identify in the huge amount of available data.

## 3 Application to Roman Inscriptions

EDB (Epigraphic Database Bari)[2] is an online, freely accessible, database hosted by the University of Bari. It includes about 30000 Christian inscriptions of Rome,

[2] `http://www.edb.uniba.it`

including inscriptions published in the *Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series* (ICVR) editions. The *ICVR* editions started in 1922 with the first volume and is going to end with the eleventh volume in the next years.

Similar initiatives are EDR (Epigraphic Database Roma)[3], EDH (Epigraphic Database Heidelbergh)[4] and HE (Hispania Epigraphica)[5]. However, they actually do not offer the possibility of performing complex text-based search, also with the help of the thesaurus. Furthermore, data about epigraphs are not entirely reported in each of these databases, since they may focus on different specific aspects. The project EAGLE (Europeana network of Ancient Greek and Latin Epigraphy), indeed, aims at the creation of a federation of these databases (including EDB), in order to allow the researcher to retrieve data about epigraphs in an integrated way from all the databases.

Currently, EDB includes the text and the metadata of around 30000 inscriptions, discovered, classified and enriched with semantic metadata by Carlo Carletti and his team. EDB makes freely available over the web the inscriptions

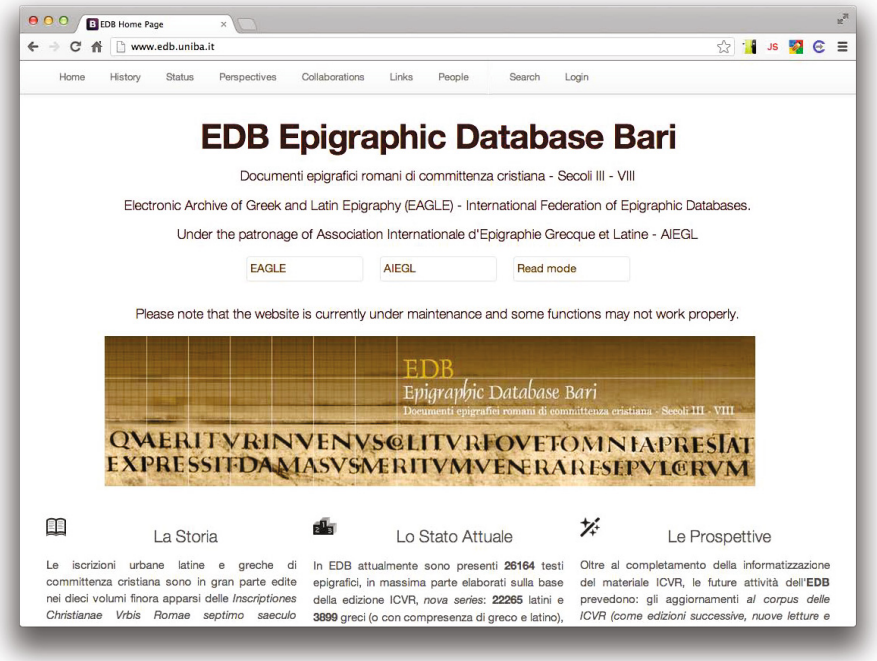
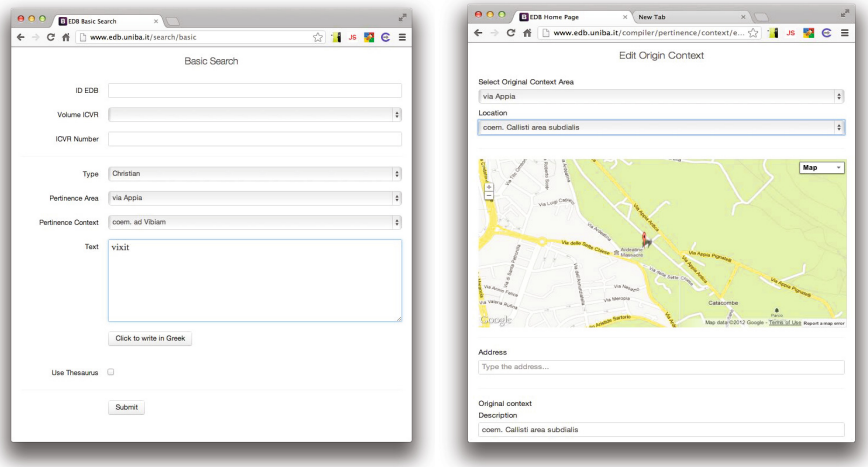



**Fig. 4.** Screenshot of the EDB's main page

[3] http://www.edr-edr.it
[4] http://edh-www.adw.uni-heidelberg.de
[5] http://eda-bea.es




(a) Screenshot of the search web page. (b) Screenshot of the geographic position for the context *coem. Callisti area subdialis*.

**Fig. 5.** Examples of functional screenshots of the EDB web interface

discovered over 25 years of archeological studies inside the Roma's catacombs. EDB allows access to an unique cultural heritage which is in its major part not accessible by public visitors.

Figure 4 shows the main page of the EDB web site. In order to maximize the user experience, EDB is implemented using HTML5 and CSS3. This allows users to fully access the information stored in EDB using a common web browser, a tablet or a smartphone. Moreover, all the functionalities of the web applications are exposed through a *restful* interface [14].

The main functionality offered by EDB is in the epigraph search. In Figure 5a the basic search web page is presented. Epigraphs can be retrieved on the basis of their EDB identifier, their ICVR volume and number, by the religious identity of the epigraph (Christian, Jewish and Pagan), by the area and context of pertinence, as well as by specifying a textual query and using the thesaurus. The text for a query can be written using latin words and greek words by enabling an automatic greek inputter inserted in the query web page. For example (see Figure 5a) if the user searches for epigraphs of type *Christian*, which are discovered on the *via Appia*, in the context *coem. ad Vibiam* and such that their text contains *vixit*, EDB currently retrieves 11 matching inscriptions (see Figure 6). This figure shows the text of the inscriptions and the related metadata.


**Fig. 6.** Screenshot of a query result page of EDB

Figure 7 shows information stored for the epigraph *14387*. In particular, it shows information on (from the top left corner) the ICVR volume, the reference literature, the area of pertinence and its context, the place where the inscription is physically stored, the support and the technique used, the editing time (which is estimated by archeologists), a link to the image of the inscription on the PCAS web site and, finally, its text.

Moreover, thanks to the spatial database, all the inscriptions stored in EDB are geo-referenced. Figure 5b shows as example the geographic position of the *coem. Callisti area subdialis* situated in the *via Appia*.

All the stored information allow us to evaluate the effort made by the archeologist and by EDB ecosystem to add valuable information to the text of the epigraph. This is the added value of the EDB project.

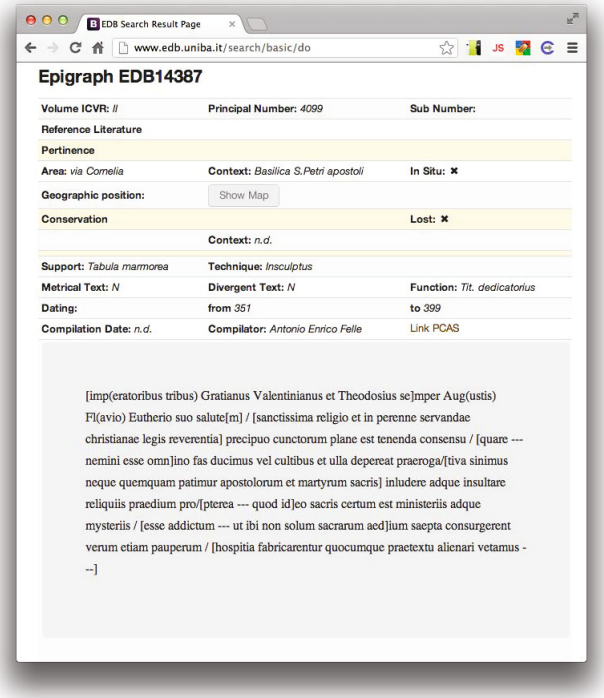



**Fig. 7.** Screenshot of the epigraph EDB-14387

## 4 Conclusions

This paper presents the system EDB which supports epigraphists in storing, managing and retrieving information on a large repository of Italian inscriptions found in the area of Rome. Peculiarities of the proposed system are in the possibility of managing, in addition to metadata, geo-spatial information and references which cite the specific epigraph. EDB also supports time-based analysis of epigraphs which aims at detecting and representing changes of inscriptions in their properties/metadata over time.

By means of its web interface, it also allows (possibly non-expert) web-users to define search queries and retrieve all the necessary information. Search queries are automatically expanded by the system, according to a thesaurus, in order to consider aberrant forms.

The effectiveness of EDB is proved by its current and extensive use by a team of epigraphists which elaborate a set inscriptions discovered over 25 years of archeological research and studies inside the Roma's catacombs.

**Acknowledgments.** The authors thank Carlo Carletti, first scientific responsible of the EDB project since 1988, Antonella Daniela Agostinelli and Anita Rocco as well as past and present collaborators: Cristina Grisanzio, Ruggero

Lombardi, Filippo Piazzolla, Marida Pierno, Miriam Ramunni, Domenico Schiraldi and Carolina Ventura.

This work partially fulfills the research objectives of the PON 02_00563_3470993 project "VINCENTE - A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems" funded by the Italian Ministry of University and Research (MIUR).

## References

1. Agosti, M., Benfante, L., Orio, N.: A contribution for the dissemination of cultural heritage content to a wider public. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 195–206. Springer, Heidelberg (2013)
2. Altamura, O., Berardi, M., Ceci, M., Malerba, D., Varlaro, A.: Using colour information to understand censorship cards of film archives. International Journal on Document Analysis and Recognition 9(2-4), 281–297 (2007)
3. Álvarez, F.-L., Gómez-Pantoja, J.-L., Barriocanal, E.G.: From relational databases to linked data in epigraphy: Hispania epigraphica online. In: Barriocanal, et al. (eds.) [5], pp. 225–233
4. Antonacopoulos, A., Karatzas, D.: Document image analysis for World War II personal records. In: 1st International Workshop on Document Image Analysis for Libraries (DIAL), pp. 336–341 (2004)
5. García-Barriocanal, E., Cebeci, Z., Okur, M.C., Öztürk, A. (eds.): MTSR 2011. CCIS, vol. 240. Springer, Heidelberg (2011)
6. Bodard, G.: The inscriptions of aphrodisias as electronic publication: A user's perspective and a proposed paradigm. Digital Medievalist 4 (2008)
7. Ceci, M., Berardi, M., Malerba, D.: Relational data mining and ilp for document image understanding. Applied Artificial Intelligence 21(4&5), 317–342 (2007)
8. Feraudi-Gruénais, F.: Latin on Stone: Epigraphy and Databases. Lexington Book (2010)
9. Frommholz, I., Brocks, H., Thiel, U., Neuhold, E.J., Iannone, L., Semeraro, G., Berardi, M., Ceci, M.: Document-centered collaboration for scholars in the humanities – the COLLATE system. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 434–445. Springer, Heidelberg (2003)
10. Gäde, M., Ferro, N., Paramita, M.L.: CHiC 2011 - Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
11. Gatos, B., Ntzios, K., Pratikakis, I., Petridis, S., Konidaris, T., Perantonis, S.J.: A segmentation-free recognition technique to assist old greek handwritten manuscript OCR. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 63–74. Springer, Heidelberg (2004)
12. Gatos, B., Pratikakis, I., Perantonis, S.J.: An adaptive binarization technique for low quality historical documents. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 102–113. Springer, Heidelberg (2004)
13. Le Bourgeois, F., Kaileh, H.: Automatic metadata retrieval from ancient manuscripts. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 75–89. Springer, Heidelberg (2004)
14. Richardson, L., Ruby, S.: RESTful web services. O'Reilly Media, Incorporated (2007)