

PART II

COLLABORATING IN DIGITAL EPIGRAPHY

8. The EAGLE Data Aggregator: Data Quality Monitoring

*Andrea Mannocci, Vittore Casarosa, Paolo Manghi, Franco Zoppi**

Abstract

The EAGLE project aggregates epigraphy related content from about 20 different data providers, and makes its content available to both Europeana and to scholars. Data Quality monitoring is a key issue in Aggregative Data Infrastructures, where content is collected from a number of different sources with different data models and quality standards. This paper presents a Monitoring Framework for enabling the observation and monitoring of an aggregative infrastructure focusing on the description of the Data Flow and Dynamics Service, and exemplifying these concepts with a use case tailored to the characteristics of the EAGLE aggregation data flow.

An Infrastructure Quality Manager (IQM) is provided with a Web user interface (WebUI), allowing her to describe the data flows taking place in the infrastructure and to define monitoring scenarios. The scenarios will include the definition of sensors (pieces of software plugged into the data flow), which will provide observations of measured objects. The scenarios include also the definition of controls and analysers, which will store and process the observations received from the sensors and will verify if the values of the measured features comply with some expected behaviour over time.

A monitoring scenario for EAGLE has been defined and tested on simulated data (the monitoring framework is still under development) in order to monitor the “health” of different data collections involved in the EAGLE collection and transformation workflows.

Keywords: EAGLE, Aggregative Data Infrastructure, Data Quality, Metrics, Monitoring

* Andrea Mannocci, CNR-ISTI, Pisa, Italy. corresponding author Email andrea.mannocci@isti.cnr.it, Vittore Casarosa, CNR-ISTI, Pisa, Italy. Paolo Manghi, CNR-ISTI, Pisa, Italy. Franco Zoppi CNR-ISTI, Pisa, Italy.

8.1. Monitoring framework

The EAGLE project (fully described in Eagle 2013) aggregates epigraphy related content from about 20 different data providers (cultural institutions all over Europe), and makes its content available to both Europeana (Europeana 2015) (through an OAI-PMH interface) and to scholars and the general public through a Web portal.

Collecting data provided by a number of different sources, very often with different quality standards, presents the challenge of measuring the overall quality of the aggregated data, and how it compares with standards and objectives set by the aggregating institution.

We present here an extension to EAGLE that will take advantage of a *Monitoring Framework* (being developed in the context of another project (OpenAIRE 2015)), enabling the observation and monitoring of an Aggregative Data Infrastructure over time.

By modelling “data processing” as a manufacturing process involving data (Ballou et al. 1998), the Monitoring Framework provides tools for the automatic extraction of observations (numeric indicators about properties of the system) from the context where the data is being processed and stored, providing time series of indicators expressed in user-defined metrics.

The Monitoring Framework offers a *Data Flow and Dynamics Service* (DFDS for short) to the *Infrastructure Quality Manager* (IQM for short), a role that (hopefully) will become standard in data aggregation infrastructures. The DFDS enables the IQM to create one or more monitoring scenarios, each designed to monitor a particular functional area or aspect of the aggregation infrastructure. For example, an EAGLE monitoring scenario could deal with the workflow that builds the aggregated content to be used by the EAGLE Web portal and by Europeana’s OAI-PMH harvesters (Mannocci et al. 2014). After the collection and processing of (possibly heterogeneous) records coming from different content providers, the aggregator stores them into different environments, for different purposes. More precisely: *i*) the processed records are indexed as a full-text index (implemented by Apache Solr) to support search and browse queries from the EAGLE Web portal; *ii*) the processed records are stored also in a document store (implemented by MongoDB) to support OAI-PMH requests. In this case, for example, we might be interested in assessing (and keep assessing over time) whether the total number of artifacts indexed by Solr matches the number of artifacts delivered via OAI-PMH.

This could easily be accomplished by monitoring the number of records stored in the index and the number of records stored in the document store, and comparing their values. As another example, we might be interested in verifying whether the trend of a certain property of the monitored system complies with a given criterion (e.g. the number of records per content provider should be strictly increasing over time).

Using the Data Flow and Dynamics Service provided by the Monitoring Framework, it is possible to define *monitoring scenarios*, which define a conceptualization of the data flows taking place in the aggregating infrastructure. Similarly to what happens to goods in a manufacturing process, data collections and processes acting over that data can be observed thanks to specially devised *sensors*, which in our case are pieces of code providing numeric values about features of interest. The monitoring scenario can also define controls to verify if the values of such features comply with some expected behaviour over time.

With the concept of sensor, we refer to a piece of software capable of generating *observations* (a numeric value plus some contextual metadata) about a measured object. *Measured objects* can be of different granularity, e.g. single data units (datum) or data collections stored somewhere by the aggregating infrastructure. Each observation is expressed in just one specific metric, intended to measure a specific feature of a measured object (e.g. the number of publications present in the data collection stored in the Solr index). A sensor can generate observations in more than one metric, each one referring to a different feature of the measured object.

A sensor can in principle be plugged anywhere in the aggregating infrastructure; with the understanding that the implementation of the sensor and the point of the data flow where it will be placed are the responsibility of the IQM. When the workflows of the aggregating infrastructure are in execution, the sensors will be activated and will produce a stream of observations. The DFDS will separate the stream of observations into different streams, each one related to a specific metric, and will store them as points of time series. In this way, observations can be queried and examined either as charts or tabular data.

The interaction between the IQM and the Monitoring Framework is done through a user friendly user interface (WebUI) provided by the DFDS. In addition to providing easy access to the observation time series, the WebUI provides the facilities to define the overall Monitoring Scenario, which will include the sensors and the *controls*.

For the purpose of monitoring, the IQM can define *controls*, i.e. checks that can be scheduled to automatically verify the compliance of one or more metrics (and their observations) with a desired (or undesired) condition. A control uses an *analyser* for the comparison of observation values, such as, for example, values alignment, less or greater than, (strictly) monotonic increasing or decreasing values, threshold guards, thresholded peak or percentage variation, etc. For example, as mentioned before, we might be interested to check if the number of artefacts indexed by Solr is equal to the number of artefacts exported in OAI-PMH records, or if the total number of EAGLE objects provided by a content provider is steadily increasing over time, or, as a further example, if the “goodness” of an EAGLE record (the concept of goodness being defined by the IQM and implemented through sensors and controls) remains above a threshold of 0.8. The Data Flow and Dynamics Service offers some analysers out-of-the-box, but also enables the IQM to develop her own custom analysers.

Finally, the Data Flow and Dynamics Service enables the generation of an exhaustive *report* about the defined metrics and controls providing insights, via the WebUI, in a quick glance about key features and potential issues present in the infrastructure. Given a set of controls, the monitoring service also takes care of raising *alerts* and *notifications* informing the IQM about the status of the infrastructure and its operation.

8.2. Architecture

The Monitoring Framework (see Figure 1) is architected as a client-server application, where a core module (imported and used by the code of the aggregating infrastructure) plays the role of *client*. The infrastructure source code needs to be instrumented in order to put sensors in place and produce observations of the measured objects.

The implementation of a sensor provides the logic to produce a measurement; in general, a sensor is devised to probe a specific object type (thus the typing of the measured object is hardcoded), while its configuration can be defined via WebUI and retrieved at runtime (dynamically). Such a design opens up to configurability and extensibility of the sensor’s collection offered by the DFDS, which in any case comes with off-the-shelf implemented sensors.

The *server* component is a stand-alone web application that receives observations from sensors, stores those observations as time series and runs automatically user-defined controls over this corpus of data. The controls present the outcome of their activity as reports, which are made available to the IMQ via the WebUI. There is also an alert and notification service to warn the IMQ about potential anomalies or wrong operation in the aggregating infrastructure.

8.3. The EAGLE use case

In order to apply the monitoring framework described above to the EAGLE aggregator, we need to define first a monitoring scenario, based on the actual EAGLE data flow, which must include the following.

- Sensors to be used in terms of measured objects and metrics implemented (i.e. functions to apply in order to extract observations). Once deployed and running, a sensor dynamically retrieves its defined configuration from the server whenever it is needed. Metrics identified in this test case are reported in Section 3.1.
- Controls representing quality checks to be run against measurements obtained with certain metrics. The framework guides the IQM into the definition of a control according to the scenario defined so far. Controls identified in this test case are reported in Section 3.2.

8.3.1. Metrics

A monitoring scenario has been defined and tested on simulated data in order to monitor the “health” of different data collections involved in the EAGLE collection and transformation workflows. In particular, a first collection is stored in a full-text index (Apache Solr) serving search queries, and a second one is stored in a document store (MongoDB) serving data for OAI-PMH export.

In this same scenario, we are interested in monitoring also the collection workflow by inspecting every single native XML record flowing into the EAGLE infrastructure from content providers. Some useful metrics identified in EAGLE are described in Table 1.

MEASURED PROPERTY	METRIC
Total # of content providers joining the EAGLE infrastructure	Content providers
Total # of languages for translations	Languages
Total # of EAGLE records	Total records
The values of this metric track the percentage of vocabulary-compliant occurrences in the XML field containing the value for “material”, over the total amount of occurrences of that field. As an example, if the vocabulary defines the entries “aaa”, “bbb”, “ccc” and the vocabulary-controlled XML field uses “aaa” four times, “bbb” three times, and “xxx” three times, the metric yields 0.7 (i.e. 7 out of 10 occurrences match the vocabulary).	Voc:material compliance
Completeness of every single collected native XML record. The values of this metric track the percentage of non-empty XML fields among 5 user-defined fields (e.g. title, description, object type, date, material.). The value could be 0%, 20%, 40%, 60%, 80%, 100%, depending on how many non-empty values have been found.	Completeness

Tab. 8.1. Metrics implemented in EAGLE.

8.3.2. Controls

Given the metrics described in 3.1, Table 2 reports some controls defined over those metrics.

Metric	Control
Content providers	Check if the number of content providers indexed in Solr is monotonic increasing over time (considering the three last observations of the metric)
Content providers	Check whether the number of content providers indexed in Solr equals the number of OAI sets present in MongoDB (considering only the last observation of the metric)
Languages	Check if the number of modern languages present in translations indexed in Solr is steadily increasing over time (considering the two last observations of the metric)
Total records	Check whether the total number of EAGLE records (per content provider) is steadily increasing over time (considering only the three last observations)
Voc:material compliance	Check if such indicator, ranging from 0.0 to 1.0, is above 0.9 threshold (considering only the very last observation of the metric)
Completeness	Check if such indicator (actually its rolling average), ranging from 0.0 to 1.0, is above 0.8 threshold (considering only the very last observation of the metric average)

Tab. 8.2. Controls implemented in EAGLE.

8.3.3. Sample implementation

Three different sensors have been defined and implemented for this test on EAGLE: two collection sensors (one for Solr and one for MongoDB) and one single-datum sensor for XML record-by-record inspection. Once the three sensors have been placed in the EAGLE workflow implementation and the EAGLE infrastructure is running, they start to produce observations and deliver them to the server component of the monitoring framework.

As an example, we report in figures included in Section 5 simulated trends of the defined metrics and relative reports (i.e. the evaluation of controls defined over the metrics).

The metric about the “Total number of content providers” is reported in Figure 2; as expected, the number of content providers indexed is monotonically (each value is greater or equal to the previous one) increasing, as stated in the leftmost report, and the number of OAI sets present in the data collection stored in MongoDB equals the number of content providers indexed in Solr, as stated in the rightmost report. It is also interesting to notice how the two trends diverged in time back in October. In the simulated data we introduced an *ad-hoc* problem in October, related to the publication of EAGLE records into the OAI-PMH store, which the monitoring service succeeded to discover.

In Figure 3, we report the “Languages” metric; as expected its trend is strictly increasing over time indicating that the corpus of translations is expanding and that they are correctly integrated into the system.

Figure 4 shows the total number of EAGLE records for four content providers (CP1, CP2, CP3, CP4) and informs the IQM that everything behaves as expected; in fact, the four reports states that the trends of the metric are always increasing over time.

In Figure 5, the metric “Voc:material compliance” is depicted. The four trends show that the four content providers (CP1, CP2, CP3, CP4) are in general increasing the quality of their data by enforcing the use of correct values offered by the controlled vocabulary for materials. However, CP1 (in orange), with a 0.83 score, does not meet the threshold requirement (set to 0.9) indicated in the control, thus its report is marked in red notifying the issue.

Figure 6 reports the oscillations of the record-by-record metric “Completeness” (narrowed down to a hundred records sample), while Figure 7 reports the associated “rolling average” (i.e. each point is the updated average up to that instant). Again, as the last observation of metrics is equal to 0.52, the relative reports is marked in red, as the 0.8 threshold defined in the control is not met.

8.4. Conclusions and future work

We have presented here the possible application of the Monitoring Framework concepts to the EAGLE aggregating infrastructure. The Monitoring Framework is “work in progress” in another European project related to research data infrastructures (OpenAIRE, 2015).

The results presented here are based on simulated data, as the EAGLE infrastructure is not yet instrumented with the sensors needed for collecting observations, but we are planning to instrument it as soon as the development of the Monitoring Framework will reach the *beta* status.

EAGLE represents an ideal *testbed* for this monitoring technology, as the workflows are clean and well defined, the data collected also is well defined, given the mappings that have been defined between the different incoming records and the EAGLE data model.

Finally, when EAGLE will be equipped with the final Monitoring Framework, we expect that it will provide valuable data for ensuring that the epigraphy data made available at the EAGLE portal will be of the highest quality. It will also provide valuable feedback to the content providers, helping them to detect possible inconsistencies and lack of information in their data, in order to improve the quality of the data provided to EAGLE and also, even more important, the quality of the data that each content provider makes available to its users.

8.5. Figures

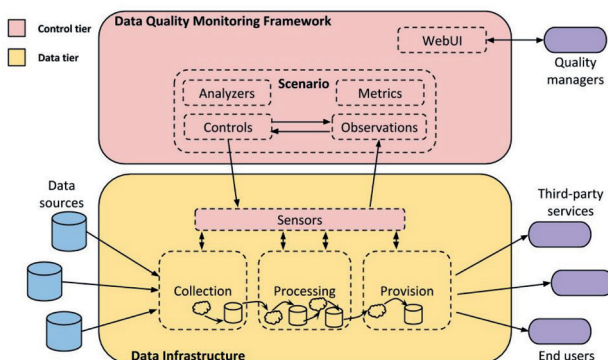


Fig. 8.1. The architectural overview of the monitoring framework.

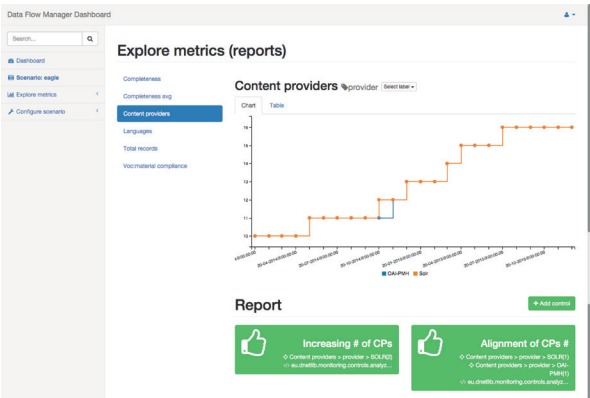


Fig. 8.2. The trend of the metric “Total number of content providers”.

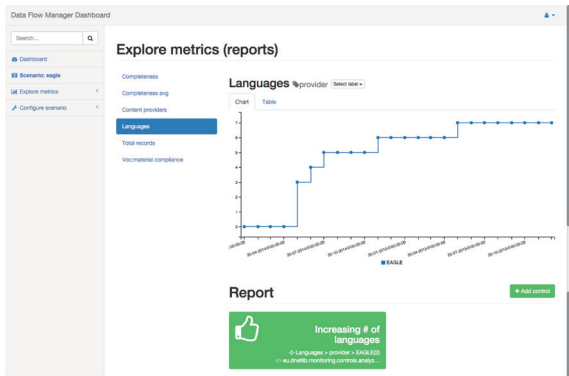


Fig. 8.3. The trend of the metric “Languages”.

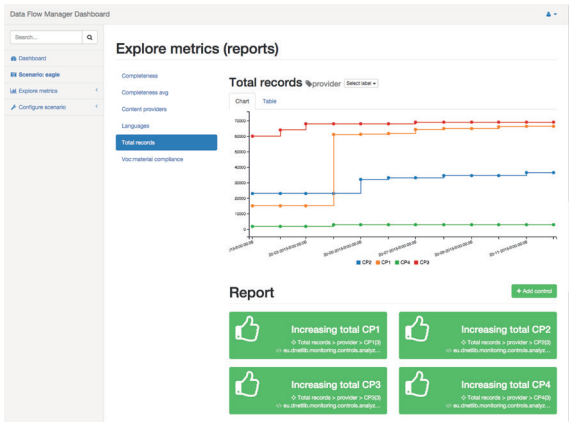


Fig. 8.4. The trend of the metric “Total records”.

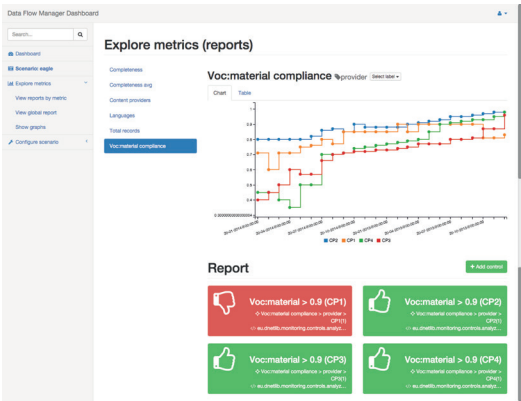


Fig. 8.5. The trend of the metric “Voc:material compliance”.

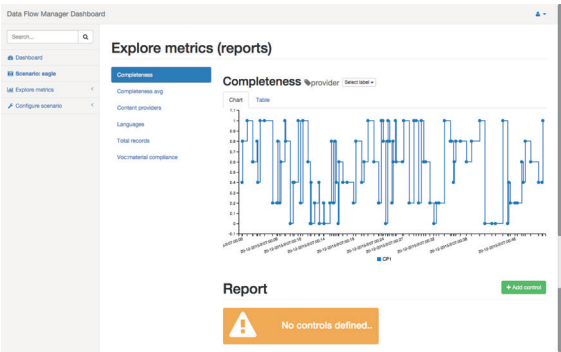


Fig. 8.6. The trend of the metric “Completeness”.

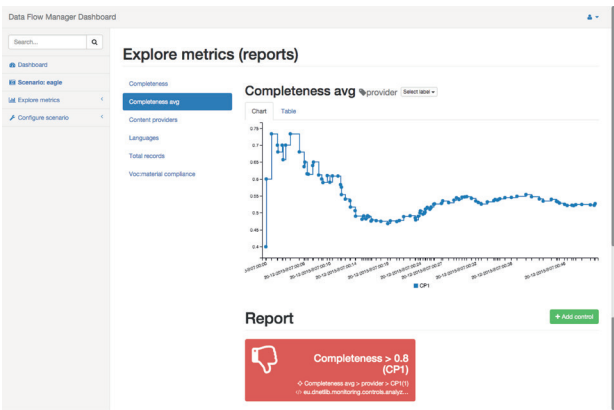


Fig. 8.7. The “rolling average” of the metric “Completeness”.

Acknowledgements

This work is partly funded by the EU EAGLE Best Practice Network project: Grant Agreement CIP 325122, call CIP-ICT-PSP-2012-6.

References

- BALLOU, DONALD, RICHARD WANG, HAROLD PAZER, and GIRI KUMAR TAYI. 1998. "Modelling information manufacturing systems to determine information product quality". *Management Science* 44(4): 462–484.
- EAGLE. 2013. "The EAGLE project". <http://www.eagle-network.eu>.
- EUROPEANA. 2015. "The Europeana project". <http://www.europeana.eu>.
- MANNOCCI, ANDREA, VITTORE CASAROSA, PAOLO MANGHI, AND FRANCO ZOPPI. 2014. "The Europeana network of ancient Greek and Latin epigraphy data infrastructure". In *Metadata and Semantics Research - 8th Research Conference, MTSR 2014, Karlsruhe, Germany*, 286–300. http://doi.org/10.1007/978-3-319-13674-5_27.
- OpenAIRE. 2015. "The OpenAIRE project". URL <http://www.openaire.eu>.

9. Searching inscriptions through the EAGLE Portal

*Claudio Prandoni, Antonella Fresa, Nicola Alfarano, Giuseppe Amato
Franaco Zoppi, Andrea Mannocci, Pietro Liuzzo, Nicola Alfarano
Francesco Mambrini, Silvia Orlandi, Raffaella Santucci**

Abstract

This paper describes the functionalities and the technical infrastructure of the EAGLE Portal, the main gateway into the EAGLE's massive epigraphic database. The portal can be accessed both by a human through a web browser and by external applications through a set of APIs. It is possible to perform full-text searches using a simple interface, or to launch more advanced queries, including the possibility to upload an image and search for inscriptions that are similar to the provided one. The seven controlled multilingual vocabularies that were created to help aligning the multilingual metadata of the inscriptions from the different content providers, have been also integrated in the search engine as well as all the translations that are available on the EAGLE MediaWiki. All this makes of the EAGLE Portal a very useful tool for the epigraphic community but also for single users and enthusiasts willing to contribute to the research in this field.

Keywords: Search engine, API, image recognition, controlled vocabulary, multilingual translations, advanced search, citizen participation

9.1. Introduction

The EAGLE Portal is the place where the content provided by the epigraphers' community is aggregated and stored and where it is made accessible to the users. It is composed by:

* Claudio Prandoni, Promoter Srl, Peccioli, Italy. Corresponding author. Email: prandoni@promoter.it; Antonella Fresa, Promoter Srl, Peccioli, Italy; Nicola Alfarano, R Gogate Srl, Ospedaletto, Italy; Giuseppe Amato, CNR-ISTI, Pisa, Italy; Franaco Zoppi, CNR-ISTI, Pisa, Italy. Andrea Mannocci, CNR-ISTI, Pisa, Italy; Pietro Liuzzo, Ruprecht-Karls-Universität Heidelberg; Nicola Alfarano, R Gogate Srl, Italy; Francesco Mambrini, Deutsches Archäologisches Institut, Berlin; Silvia Orlandi, Sapienza University of Rome; Raffaella Santucci Sapienza University of Rome.

- The Metadata Aggregation System (in the following referred also as Aggregator), which is where all the content is stored and indexed.
- A Graphical User Interface (GUI), which is the interface through which users can browse the content and interact with it.

The Aggregator relies on the D-NET software toolkit, developed within the DRIVER project and extended within the OpenAIRE, OpenAIREplus¹, EFG, EFG1914² and HOPE³ projects. D-NET is an open source, general-purpose software conceived to enable the realization and operation of Aggregative Data Infrastructures (ADIs) and to facilitate their evolution and maintenance over time. D-NET implements a service-oriented framework based on standards, namely Web Services with SOAP and REST APIs, where ADIs can be constructed in a LEGO-like approach, by selecting, customizing, and properly combining D-NET services as building blocks. The resulting ADIs are systems, which can be customized, further extended (e.g. by the integration of new services), and scaled (e.g. storage and index replicas can be maintained and deployed on remote nodes to tackle multiple concurrent accesses or very-large data size) at run-time (Amato et al. 2013).

The GUI exposes all the content stored in the Aggregator providing to the users the following functionalities (Prandoni et al. 2014).

- Search and browse the rich set of data made available by EAGLE content providers by using either a free text search or a more advanced interface, including faceted browsing through the integration of the EAGLE controlled vocabularies.
- Similarity search through the image recognition algorithm that have been integrated in the EAGLE Portal.
- Access to all the peer-reviewed translations of the epigraphic texts, in several European languages that have been produced and integrated in the EAGLE MediaWiki.
- Export to the user own PC the EpiDoc document of an object for further analysis and processing.
- Annotate and save relevant information in a user Personal Space (e.g. records of inscriptions, search results, queries), including content saved while using the Flagship Mobile Application.

¹ <http://www.openaire.eu/>.

² <http://www.europeanfilmgateway.eu/it>.

³ <http://www.peoplesheritage.eu/>.

- Access to the Flagship Storytelling Application to browse the existing epigraphy-related narratives and to create a new story.

The ingestion and curation of data is performed through a separate dedicated interface made available by the Aggregator. This interface includes a series of functionalities to support data ingestion and storage as well as for importing, indexing, enriching and managing the harvested metadata (Amato et al. 2015).

9.2. The EAGLE Collections

Classical Greek and Latin culture is at the very foundation of modern European identity. From philosophy to architecture, geometry to law, a variety of contemporary subjects and disciplines have their roots in the classical world. Only a small fraction of the total production of Greco-Roman texts has survived to the present day, leaving wide gaps in the historiographical record of an epoch that is immensely relevant to our modern day lives.

The collections held by the EAGLE partners have been assembled with the two-fold criterion of historical-cultural significance and strong thematic unity. They feature a great variety of inscriptions written in Greek, Latin and other ancient languages, providing scholars with an authoritative resource by which to verify the reliability of historical reconstructions. Additionally, they equip the broad public with a way to understand and easily appreciate interesting and geographically dispersed inscriptions.

The EAGLE collections come from a wide range of digital repositories of epigraphic content. The following paragraphs illustrate these repositories.

Arachne⁴ is the central object-database of the German Archaeological Institute (DAI). Since 2001, Arachne has been integrating negative archives of ancient sculpture that went beyond the specialised documentation retained in Cologne itself, such as the negatives of ancient sculpture of the German Archaeological Institute in Rome and the historic glass plate negative collections. Besides this larger project, many additional activities are going on different levels, for example the online preparations for the “Corpus der Antiken Sarkophagreliefs”.

⁴ <http://arachne.uni-koeln.de>.

Archaia Kypriaki Grammateia Digital Corpus (AKGDC) is a searchable digital library based on the 6 volumes of Archaia Kypriaki Grammateia (Ancient Cypriot Literature), published by the A. G. Lev-entis Foundation between 1995-2008. The AKGDC/ STARC collection of Cypriot inscriptions are dated from the 5th century BC to the 5th century AD and are mostly funerary or dedicatory epigrams in elegiac couplet or hexameter.

Epigraphic Database Bari (EDB)⁵ is a project specialized in epigraphic documents of Christian patronage (third to eighth centuries, AD), including those contained in the *Inscriptiones Christianae Urbis Romae, nova series*, voll. I-X, in civitate Vaticana 1922-1992 (= ICVR), and those edited in other bibliographical seats and/or not contained in the ICVR. Currently the inscriptions present in the EDB (counting those already online and those awaiting definitive approval) amount to 40,000 items ca, though this number is obviously increasing continually.

The task of the Epigraphic Database Heidelberg (EDH)⁶ is the systematic entry of ancient Latin and bilingual (usually Latin and Greek) inscriptions into a complex database. The Epigraphic Text Database is the heart of EDH and contains 65,000 inscriptions at present. Almost all of the records present texts, which have already either been edited in the monumental Inscription corpora or published, revised and discussed in thousands of scholarly articles.

Epigraphic Database Rome (EDR)⁷ was launched in 1999 as an experimental project aimed at creating a unified database for ancient epigraphy. EDR is part of the international federation of Epigraphic Databases called Electronic Archive of Greek and Latin Epigraphy. The federation's purpose is to collect all published Greek and Latin inscriptions up to the 7th century A.D. considering their best existing editions, also enclosing, when possible, a number of additional important data and/or images. As part of the federation, EDR aims at collecting the whole epigraphy of Rome and of the Italian peninsula including Sardinia and Sicily, with the exception of Christian inscriptions (under EDB jurisdiction).

⁵ <http://www.edb.uniba.it>.

⁶ <http://edh-www.adw.uni-heidelberg.de>.

⁷ <http://www.edr-edr.it>.

Hispania Epigraphica Online⁸ was created in 2002 when an EU grant enabled a joint research project between the Archivo Epigráfico de Hispania and Mag. K. Schaller, who was developing computing applications for archaeological purposes. The focus of the collection is the rich epigraphic patrimony of Portugal and Spain, mainly written in Latin, but with some small pockets of Greek, Semitic and Iberian inscriptions.

Petrae database⁹ is a system for the recording of Latin and Greek inscriptions developed at the Institut Ausonius, which collects epigraphic texts from the various regions in which its researchers and collaborators are active. Each record produced has the text of an inscription in both an uppercase and lowercase version, accompanied by metadata on all aspects of the monument, including support, fragments, epigraphic fields and text elements (dating, paleography, critical apparatus, translation, notes).

The Last Statues of Antiquity¹⁰ is a project funded in 2009 by the 'Arts and Humanities Research Council' in Oxford. The project collects and analyses all the evidence for new, newly dedicated, or newly reworked statuary in the period *circa* 284–650. The two main results of the project are a major database, with over 2600 individual entries, and a book, published in late 2012, discussing in print the entire phenomenon of the late-antique statue habit.

The picture database VBI ERAT LUPA¹¹ contains stone monuments (sculptures, reliefs, inscriptions, architectural pieces etc.). The project's scope ranges from prehistoric stone monuments to around the time of Justinian (500 AD). Due to its inception in Vienna, most project data at the moment is from the mid- and south-eastern European region.

Finally, thanks to the great effort in the networking activities, many other partners joined the EAGLE Best Practice Network and made available their collections that are currently being uploaded in the EAGLE Portal¹².

⁸ <http://eda-bea.es>.

⁹ <http://petrae.tge-adonis.fr>.

¹⁰ <http://laststatues.classics.ox.ac.uk>.

¹¹ <http://www.ubi-erat-lupa.org>.

¹² For a full list of the EAGLE Associated Partners see <http://www.eagle-network.eu/about/partners/>.

9.3. The EAGLE inscription search engine

The EAGLE Inscriptions Search Engine is accessible through the main horizontal navigation bar of the EAGLE Portal¹³. It represents the core functionality of the portal, through which the entry of keywords and phrases produce matches from EAGLE's massive epigraphic database.

The EAGLE Portal makes available a "simple search", an "advanced search" where the user can specify the values of a number of fields in order to make a more accurate search, and an "image search" where the user can upload an image and search for inscriptions that look similar to the provided one (Prandoni et al. 2014).

The objects that a user can search in the EAGLE Portal belong to three different categories, in accordance with the EAGLE conceptual model (Sicilia et al. 2015):

- "Artefacts", which contain all the information that is somehow related to the physical carrier of the inscription.
- "Texts", which contain all the information that is textual in nature.
- "Images", which contain all the information that is visual in nature.

9.3.1. How to search inscriptions

The simple search user interface is very straightforward. The text entered in the query box is used to make a full text search in all the fields of all the EAGLE objects in the category determined when making the query. By default, the search will be done on all the objects in the "Artefact" category. Two more tabs on the result list (labelled "Texts" and "Images") allow the user to perform the same query choosing a different category.

Using the advanced search interface, a user can specify values for a number of fields, in order to have more accurate results: findspot, bibliography, text of the inscription, type of inscription, decoration, object type, material, type of writing, state of preservation (Fig. 1).

In the fields having a controlled vocabulary, the user is allowed to enter only values coming from the vocabulary. For this purpose, those fields have a "drop-down menu" which displays all the defined values for that field in the "preferred label", i.e. the text string that has been indicated in the vocabulary as the preferred one for display, regardless of the language in which the string is defined.

¹³ <http://www.eagle-network.eu/basic-search/>.

Finally, by choosing the Image Search option, a user can exploit the image recognition algorithm that has been integrated in the EAGLE Portal and that allows to search for images that are similar to the one a user has uploaded, ranked in decreasing order of similarity (Fig. 2).

The screenshot shows the 'europeana eagle project' logo at the top left. A navigation bar includes links: HOME, SEARCH INSCRIPTIONS, COLLECTIONS, RESOURCES, NEWS, ABOUT, and CONTACTS. Below this is a sub-header 'HOME » ADVANCED SEARCH'. The main section is titled 'SEARCH INSCRIPTIONS'. On the left, there is a 'LOGIN' section with fields for 'Username' and 'Password', a 'Log In' button, and a 'Remember Me' checkbox. Below login is a 'Register' button and a 'Greek keyboard' checkbox. A sidebar on the left lists search categories: BASIC SEARCH, ADVANCED SEARCH, IMAGE SEARCH, and ARCHIVES. Under 'ADVANCED SEARCH', various filters are listed with checkboxes: Modern findspot, Ancient findspot (checked), Detailed findspot, Location, Bibliography (checked), Text of the inscription (checked), Type of inscription (checked), Decoration, Object type (checked), Material, Type of writing, and State of preservation. At the bottom of the sidebar is an 'Update search form' button. The main search area contains a 'Text of the inscription' text box, an 'Object type' dropdown menu, an 'Ancient findspot' text box, a 'Type of inscription' dropdown menu, and a 'Bibliography' text box. Below these are checkboxes for 'Only with image' and 'Only with translation', and two date range boxes: 'Not Before - Year' and 'Not After - Year'. A 'Search' button is at the bottom right of the search area.

Fig. 9.1. Search for inscriptions in the EAGLE Portal.

Content Based Image Retrieval (CBIR) is becoming an effective way for searching digital libraries, as the amount of available image data is constantly increasing. CBIR applications are increasingly becoming useful in accessing cultural heritage information, as a complement to metadata based search. In fact, in some cases metadata associated with images do not describe the content with sufficient details to satisfy the user queries, or metadata are completely missing. For example, images containing reproductions of works of art contain a lot of implicit information that is not generally captured in manually or automatically generated metadata (Amato et al., 2013).

The EAGLE Image Retrieval System (IRS) receives in input the image of an epigraph and provides effective and efficient image similarity search and image content recognition of epigraphs.

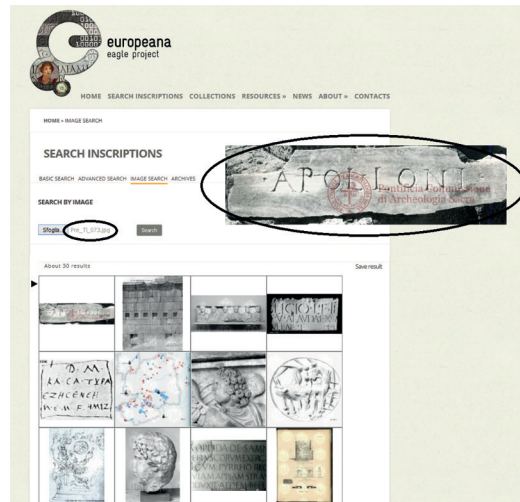


Fig. 9.2. Image search in EAGLE.

The images provided by the EAGLE content providers and collected in the Aggregator have been processed by a set of IRS components (Image Feature Extractor, Image Indexer, CBIR Index) to build the index that allows a fast and efficient similarity search during the query phase.

In addition, for those epigraphs where a set of images is available (training set), each training set has been processed to extract the main features characterizing the epigraph. The training sets and the characterizing features are the base for building another index, to be used by the image recognizer in order to decide if a received image (during the query phase) can be classified as belonging to one of the existing sets or not. In this way, in many cases, the recognizer is able to properly recognize the content of a query image even if the image given in the query was never stored in the database.

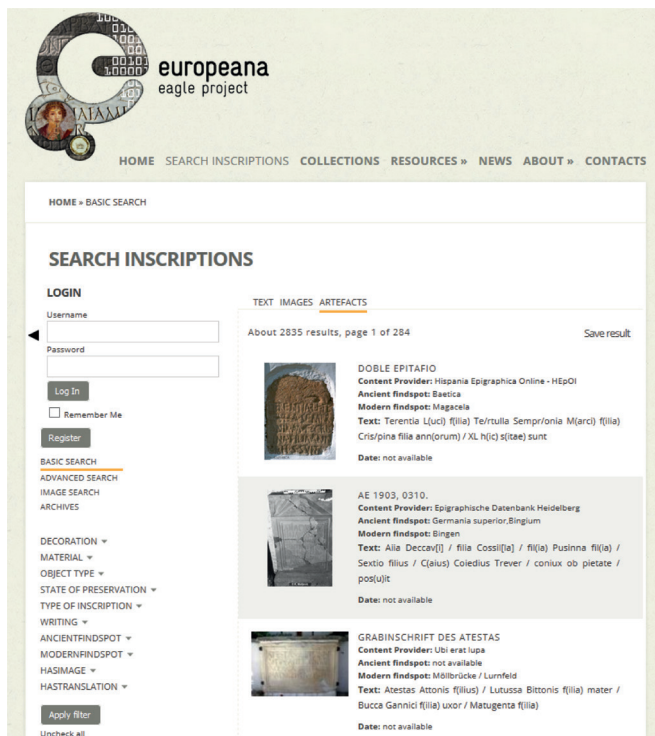


Fig. 9.3. Results list - Artefacts.

9.3.2. Filtering and browsing results

After having submitted a query, the user is presented with a paginated list of results containing some basic information (Fig. 3).

By using the facets displayed on the left side panel, the user has the possibility to refine the query by applying some filters based on the fields that are associated with a controlled vocabulary or based on the fact whether the inscription has an image and/or a translation.

Regardless of the type of query performed by the user and of the category in which the query was done, clicking on one of the items in the result page will display a summary of all the information available for that item, with links to get further details (Fig. 4).



Fig. 9.4. Object details.

The text in big characters on top of the page is the Title of the object. Below the title there is the local ID of the object and after that there is a line collecting all the clickable items of the summary page: Bibliography, which opens a text box with the bibliography associated with the object; Translations, which displays all the translations available for the inscription of the object; Original source, which links the page of the object in the database of the Content Provider who provided it; Save, which allows to save the object in the user's personal space; Export, which allows to download locally the EpiDoc document describing the object. Finally, all available pictures are displayed, together with the remaining descriptive metadata.

Duplicates identification has been performed through the Trismegistos platform¹⁴. By standardizing publication references, collection information, material, provenance and date, the overlap of the various databases has been mapped to identify records describing the same inscription. Afterwards a unique numeric identifier, the Trismegistos ID, has been assigned to each document and spread across the partner databases for inclusion in the metadata record to be uploaded in the EAGLE Portal.

¹⁴ <http://www.trismegistos.org/>.

If more than one instance of an object is available, the other instances appear on the summary page as clickable “tabs”, each tab being labelled with the local ID of the other instances of the same object.

The seven controlled multilingual vocabularies that have been created (Type of Inscription, Object Type, Material, Execution, Decoration, State of Preservation, Dating Criteria) to help aligning the multilingual metadata of the inscriptions from the different content providers, have been also integrated in the EAGLE search engine. Every time that a term which belongs to one of the vocabularies is displayed, it is automatically linked to the specific page describing the term, where further connections, synonyms and translations can be explored. The EAGLE vocabularies are accessible also as Linked Open Data¹⁵.

Finally, all the peer-reviewed translations in several European languages that are available, have been integrated in the EAGLE Portal too through a RESTful API native of Wikibase, the MediaWiki extension used for this purpose¹⁶. For each inscription it is possible to view the existing translation, to request a new one and to contribute by connecting to the EAGLE MediaWiki¹⁷.

9.3.3. Annotating and saving objects

A registered user, during a “web” or a “mobile” session, has the capability of annotating and saving (some of) the information that is being provided by the EAGLE system.

Depending on the information the user is looking at, hitting the save button will save two types of data in the personal space:

- A query and its results.
- Detailed information about an inscription.

The data saved by the local user on the EAGLE server is stored internally in a relational database.

A registered user logged in at the EAGLE Portal can access his/her saved data and perform some simple operations on it:

- Display of the saved queries or objects.
- Modify the textual annotations associated with the saved item.
- Delete the saved item from her personal space.

¹⁵ <http://www.eagle-network.eu/resources/vocabularies/>.

¹⁷ <http://www.eagle-network.eu/wiki>.

- View the saved item.
- Import the data that he/she saved during a mobile session.

It has to be noted that, when a user views one of the saved items, the object is displayed exactly as he/she saved it. The saved data might be different from the data that can be retrieved by issuing the same query at the time of editing, due to changes in the data stored in the EAGLE database. This allows the user to make comparisons between the information that was available when he/she first saved the object and what has been added afterwards.

9.4. Conclusion

This paper presented the main functionalities of the EAGLE Portal, giving some highlights of the content that is available and of the technical infrastructure that allows to search and browse them.

EAGLE aims to build a multi-lingual online collection of millions of digitised items from European museums, libraries, archives and multi-media collections, which deal with inscriptions from the Greek and Roman World. The aim of the network is to make available the vast majority of the surviving inscriptions of the Greco-Roman world, completed with the essential information about them, enriched through the use of seven multilingual vocabularies, and complemented with a series of peer-reviewed translations in several European languages, which are notoriously unavailable for inscriptions.

All this makes of the EAGLE Portal a very useful tool for the epigraphic community but also for single users and enthusiasts willing to contribute to the research in this field.

The participation of Europe's citizens in scientific research represents an important opportunity for improving European competitiveness, because of the value that citizens can add in specific areas of research. In particular, the participation of citizens in the research on CH and humanities has the potential to play an important role in the development of the European Research Area, and can take the lead in the discovery of new directions of cross-disciplinary research.

In this framework, the CIVIC EPISTEMOLOGIES¹⁸ project developed a Roadmap for the use of e-Infrastructures to support the participation

¹⁸ <http://www.civic-epistemologies.eu>.

of European citizens in CH practices and humanities research, where such engagement has a twofold benefit for culture: to be enriched by the citizens' contributions and to become more widely used and exploited (also, for example, with the participation of creative industries) (fresa\s\do5(r)oadmap\s\do5(2)015). The future development of the EAGLE initiative will refer to the CIVIC EPISTEMOLOGIES Roadmap for investigating how to put in place more advanced services supporting citizen science in epigraphic research.

The EAGLE Portal is only an example of application that makes use of the APIs provided by the Aggregator to access the EAGLE collections. Other examples are the Flagship Mobile Application¹⁹, that allows to access the inscriptions' database through a mobile device and to fully exploit the image recognition features integrated in the EAGLE Portal, and the Flagship Storytelling Application²⁰, that allows a user to create stories starting from the content available in EAGLE, Europeana and many other data sources. Other applications may be created based on the APIs provided by the Aggregator, which uses a SOLR-base search engine through which it is possible to search and browse the full set of content²¹.

Furthermore, APIs are provided to register a new user, login and perform an image-based search using the EAGLE image recognition algorithm.

References

- AMATO, GIUSEPPE, PAOLO BOLLETTIERI, FABRIZIO FALCHI, PAOLO MANGHI, AND ANDREA MANNOCCI. 2015. "Second Release of the AIM Infrastructure Specification". Deliverable D4.2, EAGLE.
- AMATO, GIUSEPPE, PAOLO BOLLETTIERI, CLAUDIO GENNARO, PAOLO MANGHI, AND ANDREA MANNOCCI. 2013. "AIM Infrastructure Specification". Deliverable D4.1, EAGLE.
- PRANDONI, CLAUDIO, VITTORE CASAROSA, AND NICOLA ALFARANO. 2014. "EAGLE Portal". Deliverable D5.2, EAGLE.
- SICILIA, MIGUEL-ANGEL, JOAQUÍN L. GÓMEZ-PANTOJA, MARÍA JOSÉ RUBIO FUENTES, EYDEL RIVERO RUIZ, PAOLO MANGHI, ANDREA MANNOCCI, AND FRANCO ZOPPI. 2015. "EAGLE metadata model specification – Second Release". Deliverable D3.1.2, EAGLE.

¹⁹ <http://www.eagle-network.eu/resources/flagship-mobile-app/>.

²⁰ <http://www.eagle-network.eu/stories/>.

²¹ For further information see http://wiki.apache.org/solr/#Search_and_Indexing.

10. Mapping Epigraphic Databases to EpiDoc

*Liuzzo Pietro**

Abstract

The Europeana network of Ancient Greek and Latin Epigraphy brings together most repositories of ancient epigraphic material and aims to provide historians not just with a useful research tool, but a curated online edition which has high quality content as well as high quality data. In this paper some of the up-conversion, alignment and enrichment tasks are presented.

Keywords: EpiDoc, XML, Vocabularies, Linked Data, Data harmonization

10.1. Introduction

The Europeana network of Ancient Greek and Latin Epigraphy¹ brings together most repositories of ancient epigraphic material and aims to provide historians not just with a useful research tool, but a curated online edition which has high quality contents as well as high quality data. Towards this end many steps are needed as the databases on which many institutions have worked for decades cannot be just discarded. Epigraphy enjoys a special position among digital resources as most transcriptions of existing classical inscriptions have been digitized. Many duplicates exist and a lot of the content lacks metadata almost entirely, but the work continues with dedication so that among the 10 % of digitised European Cultural Heritage, Ancient Epigraphy can certainly claim to play a strong role.² While the long term aim for epigraphy is on

* Ruprecht-Karls-Universität. Heidelberg. Email: pietro.liuzzo@zaw.uni-heidelberg.de.

¹ Orlandi et al. 2014b, Orlandi et al. 2014a, Liuzzo et al. 2014 <http://www.eagle-network.eu/>.

² The data is taken from the Europeana 2020 strategy report available at <http://strategy2020.europeana.eu/>.

the one hand to keep up with new finds and studies and on the other to have a common and flexible entry point and backend on which to work together, this is not easy to accomplish without a careful process which involves a gradual transition from independent databases (the best option at the time when the EAGLE databases were created) to a common online data entry and editing system working on XML files in EpiDoc, as it has now become feasible. In this paper some of the mapping and content harmonization efforts undertaken by the EAGLE BPN to achieve this step towards a common epigraphic resource, will be presented as a case in which this apparently technical task involved deep revision of contents related to the discipline and required discussions and collaborative efforts by the working groups of digital epigraphists, encoders, epigraphists and developers, coming from several different background and experiences. The result of this work are the automated conversion to TEI EpiDoc XML of 90% of the text features in the EAGLE databases and the EAGLE Vocabularies, the largest existing controlled and aligned vocabularies for classical epigraphy.³

10.2. Why XML

Choosing XML for an online publication is almost obvious given the current stage of development of the editorial methodology in epigraphy. However there is room for a brief explanation of why this is preferable on a large scale to a relational database. I would like to suggest a comparison here. Coffee is produced by filtering water through a powder more or less thick. This can be accomplished by pushing water through the coffee (as in most systems, from a mocha to an espresso machine) or coffee through the water (as in a french press, where the coffee floats and is pushed down). There is pretty much the same difference when we compare the action of entering data into a database, i.e. a rigid structure of information, and marking up a text, adding tags to the text itself. The first process forces data through a grid however well thought off, while in the second case structure is added on the contents without altering them, but just enhancing their potential so that the richness and values of both contents and structure empower each other. There is just more taste and much more freedom of encoding to the level desired and to the complexity one wants.

³ <http://www.eagle-network.eu/resources/vocabularies/>.

An additional advantage is the level of interoperability. An XML file is virtually software independent, it is not an Access, FileMaker or whatever database format. A third reason is that an XML file can host a complete textual description with semantic markup and still have the structure of a normal text edition, while the same source can be used to produce a database or an index for interrogation purposes. For example the same markup of an abbreviation `<expan><abbr>C</abbr><ex>aio</ex></expan>` can be used to produce a database of abbreviation marks and expansions, to output in a diplomatic edition only the character actually on the stone or to print the expanded abbreviation in a critical edition.⁴ To print a database to a book would be at least more cumbersome. XML provides also a better service as an archival format being explicit on the values it uses to describe contents, moreover if these follow a schema of agreed value of such elements. This said producing and sharing XML files means that more quality editions can be produced, printed editions can be produced and all sort of outputs can be supported with flexibility; it means that data is safe from software developments and support; it means you don't have to put information into a rigid table which is then hard to restructure, but that you can annotate your text as much as you want thus adding layers and layers of "possible databases" of relevant information to be extracted and used. The text is what one uses to describe situations and complexities, the markup allows also the machine to know and retrieve that information as in a database: you don't need a field or an element to state what the orientation of your text is, you can simply state and describe as you would do in a normal publication and if you want to reuse information about the text orientation then you can add elements and attributes to do so. XML allows not to worry too much on "what goes where" or about "where do I put that info" because it is not a database, so that the editor can instead focus on "what is that" and on accurately describing it. The TEI specification which is EpiDoc allows the specific epigraphic content to be described fully for what concerns the text of inscriptions, drawing on more than a decade long experimenting and development of this standard thanks to a vast and active international community. The simple fact of using the same markup to describe our texts allows us to study them together and to produce and run software for more than just one corpus of inscriptions.

⁴ <http://www.stoa.org/epidoc/gl/latest/trans-abbrevmark.html>.

10.3. EAGLE mappings and metadata modeling

The EAGLE BPN has taken 18 different content providers and has come to the decision to take the first steps to bring hundreds of thousands of ancient Greek and Roman inscriptions to TEI. This required 14 mappings of local database metadata models to TEI/EpiDoc, as well as the elaboration of XSL Transformations⁵ for up-conversion⁶ and the alignment of the text markup. Although this was thought to be a trivial task it turned out that whereas databases claimed to use the same conventions, a considerable amount of differentiation took place due to the data structure and to data entry procedures as well as to policies and internal decisions.

The main harmonization task undertaken has been to align the XML format of data provided for aggregation and ingestion, to the TEI specification EpiDoc.⁷ This well established TEI schema, broadly used in many high quality projects, allows for a very easy alignment, for the production of an XML file compliant with international standards and for high flexibility for integration of the vocabularies and places gazetteer in use.

Moving from a database to a marked up text is a partially mechanical operation which involves a theoretical jump. As coffee is better when water goes through coffee powder rather than when coffee goes through the water (as in a french press), a database has fields to be filled, while marking up a text is a descriptive activity which is attached to information whichever textual form it takes.⁸ Mapping from a database to XML forces into the XML a structure and a logic which is that of a database, whereas the freedom and flexibility achieved with XML are yet to be actually realized and exploited.

In order to offer a complete and critically structured endpoint to the user on the side of data, to describe inscriptions and their representations, EAGLE considered beside TEI also CIDOC CRM, studying and providing a full EpiDoc to CIDOC-CRM mapping.⁹

⁵ <https://github.com/EAGLE-BPN>.

⁶ This is the terminology used by Kay 2008, 906 to describe transforming without explicit structure in data which has it. In this case it is a transformation from a less explicit database structure to a fully explicit and interoperable one.

⁷ <http://www.stoa.org/epidoc/gl/latest/>.

⁸ See above.

⁹ Still under development, within the ARIADNE project. <http://www.ariadne-infrastructure.eu/>. The need for a working group and a variant declension of CIDOC for epigraphy have been highlighted during the Nicosia EAGLE meeting in March 2015.

This would have enabled a further full description in a different logic, that of the web of data, but the attempts made proved the need for more specific efforts.

There are several mappings involved in the aggregation work of EAGLE. Content providers need to map to a common Eagle Metadata Format, and the data produced will then be mapped to the Europeana Data Model¹⁰ in order to be aggregated in Europeana, for a wider dissemination with a special eye for the general user.

10.4. EAGLE metadata model and harmonization

The occasion of a mapping work allows also for other tasks of curation to be performed. EAGLE Members had in some case the text of a same inscription with three parallel different types of encoding conventions applied.

Content Curation: Transcriptions

<p>5</p> <p>D(is) M(anibus) s(acrum). L(ucio) Silicio Niconi filio, qui vixit an(nis) XXII, mens(ibus) VI, die= b(us) XX L(ucius) Silicius Ni= con pater fecit.</p>	<p>D(is) M(anibus) s(acrum) / L (ucio) Silicio Niconi / filio qui vixit an(nis) / XXII mens (ibus) VI die/b(us) XX L (ucius) Silicius Ni/con pater fecit</p>	<p>D M S L SILICIO NICONI FILIO QVI VIXIT AN XXII MENS VI DIE B XX L SILICIVS NI CON PATER FECIT</p>
---	--	--

```

<div type="edition" xml:lang="la">
  <head>Text</head>
  <ab><b>5</b></ab><exp><abbr>D</abbr><ac>le</ac></exp> <exp><abbr>M</abbr><ac>anibus</ac></exp>
  <exp><abbr>s</abbr><ac>acrum</ac></exp><b> </b><exp><abbr>L</abbr><ac>ucio</ac></exp> Silicio Niconi<b>
  n="3">filio qui vixit <exp><abbr>an</abbr><ac>nies</ac></exp><b> n="4">XXII <exp><abbr>mens</abbr><ac>ibus</ac></exp>
  <exp>VI die<b> break="no" n="5"><exp><abbr>b</abbr><ac>us</ac></exp> XX L</exp> <exp><abbr>Silicius Ni</abbr><ac>con pater fecit</ac>
  </div>

```

5

D(is) M(anibus) s(acrum)
L(ucio) Silicio Niconi
filio qui vixit an(nis)
XXII mens(ibus) VI die=
b(us) XX L(ucius) Silicius Ni=
con pater fecit

Fig. 10.1. Different Texts before and after harmonization though up-conversion.

In the context of the mapping to EpiDoc of the metadata, the text also underwent transformation with tools which have been developed to support the alignment and harmonization of data from content providers to international standards for what concerns digital editions of inscriptions.

Given the template described in Part III, and ANNEX II of the EAGLE Metadata Schema (Sicilia et al. 2015), an XSL Transformation converts from string epigraphic texts in marked up TEI-EPIDOC XML, following the EpiDoc¹¹ guidelines (Elliott et al. 2007-2016).

¹⁰ Sicilia et al. 2015.

¹¹ <http://www.stoa.org/epidoc/gl/latest/>

These XSLT:¹²

1. allow the conversion of epigraphic texts with various encodings and conventions from string to Epidoc markup and valid against the The EpiDoc RelaxNG schema.
2. Populate appropriate elements with available common URI from the EAGLE vocabularies¹³.

This export set up will also guarantee that contents are kept aligned to the EpiDoc guidelines at all stages guaranteeing an effort free alignment to these international conventions for partners who can continue to apply local conventions for editing. But I would like to give more details on the steps of this process as an example of how it was possible to extract semantics, entities, and patterns from these text while aligning metadata format to an internationally recognized standard.

10.4.1. Step 1

Each project uses different conventions and therefore the regular expressions used to match particular situations are different. The process of mark up of the string text in `div[@type="edition"]` is accomplished in several steps to guarantee consistency and precision.

The `textstructure.xml` looks for marker of different sections and tokenize them to apply the same XSLT to each section of the text which needs to be contained by an `<ab>` element. If there is only one part it applies following instructions to that only.

Each section of text is then processed by the `brackets.xml`. Normalizing brackets is important for the following steps and splits individual semantic values. The notation `[ort 3]`, which would mean that a supplied text is followed by a gap of three letters, is divided into `[ort][3]`.

The normalized string which results from this process is then passed to the `up-conversion.xml` which works using a specific operation to search for regular expressions patterns (`xsl:analyze-string`) and substitute them with correct xml elements.

Running the transformation on the pattern `<E=F>` will return the following result:

```
<choice><corr>E</corr><sic>F</sic></choice>
```

¹² Based on `Chetc.txt` (by Hugh Cayless, Elli Mylonas, Gabriel Bodard and Tom Elliott) and further support from the Epidoc Collaborative (especially from Gabriel Bodard).

¹³ <http://www.eagle-network.eu/resources/vocabularies/>.

The result of this template is then passed to a further template which gives consistent numbers (insertnumbers.xsl). Empty lines do not need to have numbers, so Xpath is used to evaluate where to put a 0 as value of the @n in the <lb> element. Starting from this

```
-----] / e[t?] Q(---) Bl(a)e[sus?] / contub/ernalis / eius / d(e) s(uo) l(ibens)
l(aetus) d(edit)
```

The result of this processes is then the following¹⁴

```
<ab>
<lb n="0"/><gap reason="lost" extent="unknown" unit="line"/>
<lb n="1"/><e<supplied reason="lost" cert="low">t</supplied>
<abbr>Q</abbr>
Bl<expan><ex>a</ex><abbr>e</abbr></expan><supplied
reason="lost" cert="low">sus</supplied>
<lb n="2"/>contub
<lb break="no" n="3"/>ernalis
<lb n="4"/>eius
<lb n="5"/><expan><abbr>d</abbr><ex>e</ex></expan>
<expan><abbr>s</abbr><ex>uo</ex></expan>
<expan><abbr>l</abbr><ex>ibens</ex></expan>
<expan><abbr>l</abbr><ex>aetus</ex></expan>
<expan><abbr>d</abbr><ex>edit</ex></expan>
</ab>
```

10.4.2. Step 2

On the elements which contain information such as Object Type, Material, Execution Technique, etc. which are typically handled with a controlled vocabulary a series of XSL transformations is passed, one for each vocabulary related to that element to match the content of the element with the vocabulary entry into the SKOS vocabulary stored in git, regularly updated and published as a self standing resource on the EAGLE portal.¹⁵ What follows is a brief description of such vocabularies and their development.

¹⁴ The amount of feature supported is much higher and some example of complex text successfully converted can be seen in this presentation online.

¹⁵ <http://www.eagle-network.eu/resources/vocabularies/>.

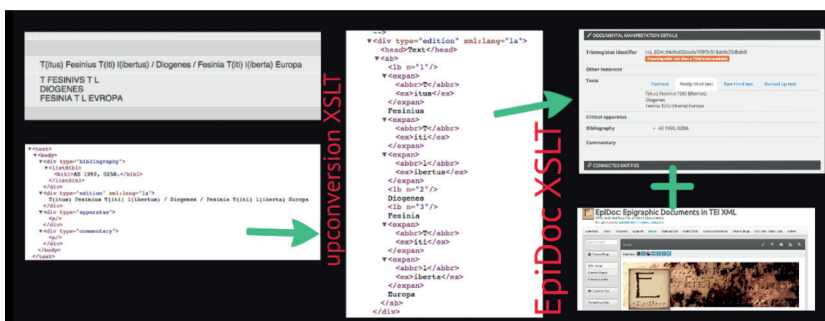


Fig. 10.2. Stages of transformation of a text: database, template, up-conversion, checking and display with EpiDoc XSLT.

10.5. Classification problems: the EAGLE Vocabularies

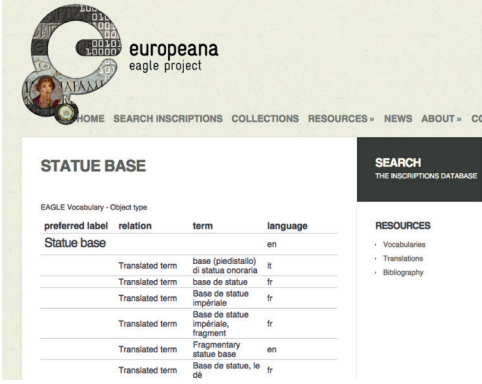
As Di Stefano Manzella¹⁶ clearly explained classification is no easy issue in any field: epigraphy is no exception to this rule. Traditionally the CIL VI (Rome) classification has been used as a reference, as this typology has served as a model for all epigraphic production in the Roman Empire. There are nevertheless new glossaries and classification curated by http://cil.bbaw.de/cil_en/dateien/glossar.php#auswahlglossarCIL, which retain the scientifically selective constraints of a formal classification, together with the benefits of this.

Problems are various, and include also the use of terms across vocabularies and the doubts which might be generated by archaeological chance.¹⁷

What looks like an **altar** could be also the **base** of a statue, for example. Or perhaps it could have served two different function has an object of which we might or not have any archeological, contextual or textual trace.

¹⁶ Di Stefano Manzella 1987, 109.

¹⁷ See Piso 2001, XI-XII, following a current of studies which has its main in G. Susini and J.-N. Bonneville.



STATUE BASE

EAGLE Vocabulary - Object type

preferred label	relation	term	language
Statue base			en
Translated term		base (pedistallo)	it
Translated term		di statua onoraria	fr
Translated term		base de statue	fr
Translated term		Base de statue impériale	fr
Translated term		Base de statue impériale, fragment	fr
Translated term		Fragmentary statue base	en
Translated term		Base de statue, le	fr
Translated term		de	

SEARCH
THE INSCRIPTIONS DATABASE

RESOURCES

- Vocabularies
- Translations
- Bibliography

Fig. 10.3. Statue Base.

This is also the case when we deal with techniques of execution of an inscription. Those can also be multiple and an inscription can be for example both a graffito and painted. And this is just not to mention the extreme complications of having a lot of independent vocabularies, each one for its own sake and small scope purpose. This is often the best option for correct attribution as local habits do vary. But within the framework of harmonization activities, it is important to refer to connected and interrelated definition of terms (not univocal!) and to allow for multiple values to coexist. This is one best practice in the definition of vocabularies which is very important for the nature of content in question. This is also the reason why the EAGLE decided to provide not just definitions of all main terms but also examples from different areas and in different state of preservation, and as far as possible also bibliographic reference to authoritative sources.

The EAGLE Vocabularies are the following

- <http://www.eagle-network.eu/voc/typeins.html> Type of Inscription
- <http://www.eagle-network.eu/voc/objtyp.html> Object Type
- <http://www.eagle-network.eu/voc/material.html> Material
- <http://www.eagle-network.eu/voc/writing.html> Writing and Execution
- <http://www.eagle-network.eu/voc/decor.html> Decoration
- <http://www.eagle-network.eu/voc/statepreserv.html> State of Preservation
- <http://www.eagle-network.eu/voc/dates.html> Dating Criteria

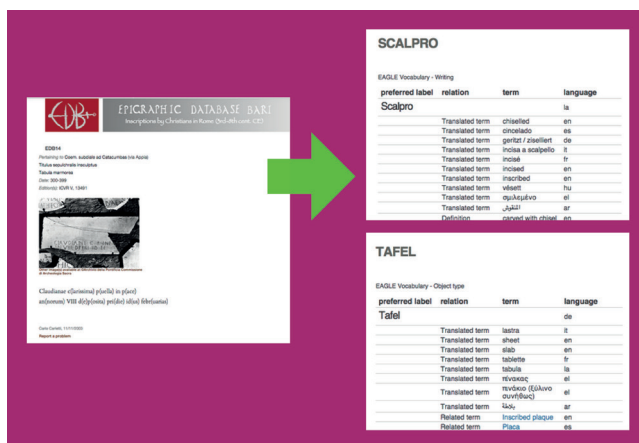


Fig. 10.4. EAGLE Vocabularies used by the Epigraphic Database Bari.

The EAGLE vocabularies are maintained in a git repository where they are manually updated and transformed to a readable versions (with another XSLT transformation) published on the EAGLE website as a self standing resource linked to several others. A user is thus able to search for a type of object and read a definition as well as searching for more information on Wikipedia, Wikidata, Wikisource, etc. including the partners' websites. The Vocabularies perform thus both a technical function and are at the same time a living resource which can be read and used independently.

10.6. Linked Open Data

The Europeana LOD practice for metadata recommends the adoption of machine readable vocabularies. Within the EAGLE BPN Linked Open Data practices and approaches have been taken on board for a process which will bring current material some steps closer to that practice.¹⁸ Addressing existing problems in classification and the publication of a machine readable vocabulary of values (controlled vocabulary) is one of these steps:¹⁹ it involves identifying and matching values as well as giving them stable identifiers and making of them self standing resources where possible.

¹⁸ Bizer et al. 2009.

¹⁹ Harper 2012 for the definition. See also Isaac et al. 2012.

In facts the many problems and complexities which Di Stefano Manzella (1987) underlined and exposed can find a solution using the LOD approach.²⁰ Both the limits of choices and hierarchical organization can be bypassed using unique identifiers and relations among those. Polyhierarchic structures as well as the presence of many possible denomination with specific and self standing definition allows for both precision, consistency and sustainability over time of this approach.²¹

For example, an inscription can easily be classified both as *magistrati populi romani* and as *decuriones municipales*; a document can be classed both under the *fasti* and under the individual mentioned in the text. These major advantages can be exploited when the effort of publishing open machine readable material is undertaken and a document can be then classified chronologically and alphabetically without the need of an authorial choice.²²

Nevertheless, the major possible advantage is that there is no evident need, with a LOD approach, to distinguish among Greek and Latin inscriptions: they can find their place together in a Linked Data edition and benefit of other efforts in this direction.

10.7. Crosswalking EpiDoc to EDH with XSLT

The proof of concept of usability of the XML source file is given by the reverse function being possible, i.e. to transform the XML files into a database data structure. A team of people from the EAGLE BPN²³ and Scott Vanderbilt, curator of the online edition of the Roman Inscriptions of Britain,²⁴ worked on mapping the EpiDoc contents of that project to the EDH database model,²⁵ facing a series of problems and challenges, which brought an interesting discussion, also on principles of “up-conversion” and cross-walking, which should be as often as possible multi-directional.²⁶

²⁰ Bizer et al. .2009.

²¹ Harpring 2010.

²² I would like to thank prof. Piso for input on this point, given during the work of the Working Group.

²³ Gabriel Bodard, Pietro Liuzzo, James Cowey, Frank Grieshaber, Brigitte Gräf and Francisca Feraudi Gruénais.

²⁴ <http://romaninscriptionsofbritain.org/>.

²⁵ The work was largely based on the previous exercise of this kind, made by Gabriel Bodard and James Cowey for the Inscriptions of Roman Tripolitania.

²⁶ The files are available here: https://raw.githubusercontent.com/EAGLE-BPN/Epidoc2EDH/master/rib_to_edh-2.xsl.

Challenges included

- alignment to EDH internal vocabularies
- alignment to EDH conventions for bibliographic references
- use of internal references and retrieval of key based information
- matching of existing items
- formatting to EDH conventions differentiated based on the type of content extracted (text or names)

The result was successful and proves once more if needed that an XML source format is a preferable option whatever the variety of desired output are.

10.8. Conclusion

I have presented in this paper the effort of a large consortium of people and institutions belonging to the same field of interest, that of Ancient Roman and Greek Epigraphy. Mappings and harmonization of data have been for the EAGLE project not just a way to allow machines to do more, but an intellectual effort of revision of contents and establishment of best practices. This work prepares further steps which hopefully will take place soon, in which one repository of inscriptions will be accessible to all to edit, contribute and download data from several websites and interfaces. In such environments related contents based on descriptive URIs, as well as places, personal names and other semantic information will be possible and meaningful and perhaps in a not too far future classicists will have a complete and functioning network of Linked Ancient World Data, including texts, manuscripts, papyri, coins and inscriptions.

References

- BIZER, CHRISTIAN, TOM HEATH, AND TIM BERNERS-LEE. 2009. "Linked Data - The Story So Far" 5(3): 1–22. URL <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>. This is a freely available preprint of the article later published in IJSWIS.
- DI STEFANO MANZELLA, IVAN. 1987. *Mestiere di epigrafista: guida alla schedatura del materiale epigrafico lapideo*. Rome: Edizioni Quasar.
- ELLIOTT, TOM, GABRIEL BODARD, ELLY MILONAS, SIMONA STOYANOVA, CHARLOTTE TUPMAN, AND SCOTT VANDERBILT. 2007-2016. "EpiDoc Guidelines: Ancient documents in TEI XML". <http://www.stoa.org/epidoc/gl/latest/>.

- ISAAC, ANTOINE, ROBINA CLAYPHAN, AND BERNHARD HASLHOFFER. 2012. "Europeana: moving to Linked Open Data" 24 (2/3): 34–40.
- KAY, MICHAEL. 2008. *XSLT 2.0 and XPath 2.0 Programmer's Reference*, 4th edition. Birmingham: Wrox.
- LIUZZO, PIETRO MARIA, ANDREA ZANNI, LUCA MARTINELLI, LORENZO LOSA, AND PIETRO DE NICOLAO. 2014. "The EAGLE Mediawiki. A fully collaborative database for academics, data engineers and the general public." In *Information Technologies for Epigraphy and Cultural Heritage*, edited by Silvia Orlandi, Vittore Casarosa, Raffaella Santucci, and Pietro Maria Liuzzo 187–200. Rome: Sapienza Università Editrice.
- ORLANDI, SILVIA, LUCA MARCO CARLO GIBERTI, AND RAFFAELLA SANTUCCI. 2014a. "EAGLE: Europeana Network of Ancient Greek and Latin Epigraphy. Making the Ancient Inscriptions Accessible." <http://lexicon.cnr.it/index.php/LP/article/view/408>.
- ORLANDI, SILVIA, RAFFAELLA SANTUCCI, VITTORE CASAROSA, AND PIETRO MARIA LIUZZO, eds. 2014b. *Information technologies for epigraphy and cultural heritage. Proceedings of the first EAGLE international conference*. Rome: Sapienza Università Editrice.
- PISO, IOAN. 2001. *Inscriptions d'Apulum*, volume III of *Inscriptions de la Dacie Romaine*. Paris: De Boccard.
- SICILIA, MIGUEL-ÁNGEL., JOAQUÍN L. GÓMEZ-PANTOJA, MARÍA JOSÉ RUBIO FUENTES, EYDEL RIVERO RUÍZ, PAOLO MANGHI, ANDREA MANNOCCI, FRANCO ZOPPI. 2015. "EAGLE metadata model specification – Second Release". Deliverable D3.1.2, EAGLE.

11. Trismegistos Places, a geographical index for all Latin inscriptions

*Herbert Verreth**

Abstract

The Trismegistos database has recently created a geographical index for all Latin inscriptions. For the moment we have 67.820 geographical references attested in Latin documentary texts, but this rough starting material still has to be refined. This paper describes how we undertook this task, which problems we encountered while doing so, and the choices we made for the presentation of the material.

Keywords: Imperium Romanum, Latin, epigraphy, topography, geography, Trismegistos

The Trismegistos (TM) database (<http://www.trismegistos.org>) of the Leuven in Belgium gathers the metadata for all documentary and literary texts from Egypt and the Ancient World in general written in whatever language or script between 800 BC and 800 AD. In this respect we try to collaborate as much as possible with other scientific databases all over the world. We started some ten years ago with the papyrological material from Egypt (both Greek and Egyptian), and the last few years – through the collaboration with EAGLE – we have also been incorporating Latin epigraphical texts from the whole Roman world. For the moment TM shows the metadata for 618.917 texts, 439.858 of them written in Latin or containing Latin passages, but the number keeps on growing. Our direct EAGLE partners Epigraphic Database Heidelberg (EDH), Epigraphic Database Roma (EDR), Hispania Epigraphica Online Database (HEp) and Epigraphic Database Bari (EDB) mainly focus on inscriptions on stone and other important texts from the regions they cover, but they usually omit

* Trismegistos Project, Leuven, Belgium. Email: Herbert.Verreth@kuleuven.be.

the inscriptions and stamps on *instrumentum domesticum* or other minor materials. Since TM also wanted to incorporate these texts, we were glad to find them in the Epigraphik-Datenbank Clauss-Slaby (EDCS), which contains virtually all published Latin inscriptions, to fill in the missing gaps, a process which is still going on.

Trismegistos is a relational database created within the computer program Filemaker (with as latest version Filemaker Pro 14). Its contents are uploaded weekly to an online MySQL / PHP environment. Attached to the main TM Text file are numerous other files, such as Collections, Archives, People, Names and Places. These files automatically copy the relevant information for every card from the main file, so that no double work is needed. With the help of the Papyrological Navigator (PN) (<http://www.papyri.info>) we developed a PHP environment with the assistance of a programmer (Jeroen Clarysse), to tag all words starting with a capital occurring in the published papyrological texts, which yielded in the end a full index of all personal names and toponyms in every Greek and Latin papyrus. This process is nowadays called Named-Entity Recognition [NER]¹. These TM files within People and Places are freely accessible online and can be looked up, questioned and investigated in a number of ways. But that is a different story. Here we want to focus on our project to do the same for all published Latin inscriptions.

The main credits for the whole set-up of this new project go to Mark Depauw, the Trismegistos director. The general idea remained the tagging of all words starting with a capital, but this time Depauw tried a new approach, which was completely imbedded within Filemaker. Since the Roman 'tria nomina' hardly occur in the Greek papyri, also a new onomastical structure had to be devised for the automatic recognition. As a test case we choose the full text corpus of the EDCS. All these texts were 'cut up' in capital clusters, i.e. strings of consecutive words all starting with a capital. Words as *filius*, *nepos* and *libertus* were added to the string so that in most cases the full identification of a person could be grouped together, e.g. *Quintus Caecilius Quinti filius Quirina Mustacus* (TM 332260), *Caio Annioleno Cai filio Arnensi Karthaginiensi Galliano* (TM 349961) or *Maesiaes Cai libertae Chrysidis* (TM 244384). A minor disadvantage of this corpus was the capitalization of the first word of every inscription (which is not done in the PN),

¹ For more information on NER, Broux and Depauw 2014, 304-313.

which yields quite a significant number of mere nouns in the group of expected personal names and toponyms. Also other words starting with a capital were not our prime goal: names of gods, religious festivals, months, army units, ships, animals, mythological persons and Roman numbers. Excluded (for the moment) are also the names of the emperors and the members of the imperial family, often occurring in a complicated titulature which is not easy to standardise. In the end this yielded 898.134 capital cluster cards. From our previous projects we already had a fairly elaborated reference corpus of Roman personal names, which was now expanded and used by Depauw to match every word in the capital clusters with the names in the reference corpus. If there was a match, the case of the ending was added, e.g.

Caio	Annioleno	Cai	filio	Arnensi	Karthaginiensi	Galliano
dat	dat	gen	filius	tribus	origo	dat

EDCS id temporary	16305	Tit, lexid 349961	take_last_with_next	next_word	famini d	349961
Cluster001	7	Caio Annielino Cai filio Arnensi Karthaginiensi		previous_word		match_twotextids 1
Numples, Cluster001			check check	georef		
Clusterwith001	C(aio) Annielino C(a) (filio) Arn(ensi)					
Clusterline001	1 1 1 1 2 2 3	1 1 1 1 2 2 3	line_number	1 1 1 1 2 2 3		
Cluster_no	1	Cluster_no_plus1	2	transfer_plus21		
less 21 capitals	y88					
cluster_interpretation		GEO OK	exported	OK_voorlopig	yes	
	Word01	Caio	Word01_namevcase	316154	Word01_case	dat
	Word02	Annielino	Word02_namevcase	440673	Word02_case	dat
	Word03	Cai	Word03_namevcase	316163	Word03_case	gen
	Word04	filio	Word04_namevcase	++filio	Word04_case	filius
7242	Word05	Arnensi	Word05_namevcase		Word05_case	tribus
484	Word06	Karthaginiensi	Word06_namevcase		Word06_case	Arn(ensi)
	Word07	Galliano	Word07_namevcase	430100	Word07_case	Karthagin(i)ensi
	Word08		Word08_namevcase		Word08_case	Galliano
	Word09		Word09_namevcase		Word09_case	
	Word10		Word10_namevcase		Word10_case	
	Word11		Word11_namevcase		Word11_case	
	Word12		Word12_namevcase		Word12_case	
	Word13		Word13_namevcase		Word13_case	
	Word14		Word14_namevcase		Word14_case	
	Word15		Word15_namevcase		Word15_case	
7726	Text	C(aio) Annielino C(a) (filio) / Arn(ensi) Karthagini(i)ensi Galliano / fiam(i)n(i) divi Titi / civitas Uccula / decreto Afror (um) / positu	Text without brackets	C(aio) Annielino Cai filio Arnensi i Karthaginiensi Galliano fiamini divi Titi civitas Uccula decreto i Afrorum positu		
TM_provenance	Tunisia, Africa - Uccula		TM_century			
EDCS province	Africa proconsularis		comment			
EDCS place	Henchr Durat / Uccula					

Fig. 11.1. Named Entity Recognition with FileMaker Pro 14.

The technical details of this complicated process are better discussed by Depauw himself at some other occasion. On the basis of this matching all capital cluster cards were split up in two groups: (1) 'yes', this card contains a personal name [454.183], and (2) 'no', this card does not contain a personal name [443.951].

Within the second group also other labels have been added, e.g. army [23.201], god [15.663], emperor [41.944], which will be useful for future research.

This is also the phase where the toponyms come in (which can occur both in the first and in the second group). Within the EAGLE project we created already a fairly large reference corpus for toponyms from the Roman empire, but this corpus was enlarged by entering the toponyms occurring in the *Itinerarium provinciarum Antonini Augusti* [3.434] and the *Tabula Peutingeriana* [3.287]. The TM Geo file now contains 46.707 toponyms from all over Egypt and the Roman empire, both ancient and modern, which cover most of the places where ancient texts have been found, and most of the toponyms mentioned in Egyptian papyrological and Latin epigraphical sources². We played with the idea of automatic matching, like we did for the personal names, but except for the case of the relatively straightforward tribus names, this yielded no satisfactory results. A lot of toponyms resemble personal names (e.g. *Florentia*, *Venusia*, (*Fundus*) *Bassianus*), and the automated identification of strings such as *Colonia Ulpia Traiana Augusta Fidelis Lepcis Magna* (TM 198383) or *Municipium Augustum Hipponiensium Regionum* (TM 200133) seemed too cumbersome. In the end, we settled for plan B, which shows that in ‘Digital Humanities’, the human component is still essential: we had to go through all 900.000 capital cluster cards manually, identifying the toponyms in every cluster, adding the corresponding TM Geo number and – whenever necessary – correcting the indications for yes/no and the automatic identification of the cases of the personal names. It was six months of tedious work, resulting in 61.139 capital clusters with at least one toponym. No doubt some toponyms escaped our attention, but we do hope to have identified the majority of the names involved. In this process, however, we also encountered some set backs, especially in the longer texts and in the more complicated wooden or metal tablets: in the beta version on which we worked, the creation of the capital cluster strings was not always so perfect as we had hoped for, and also the line numbers automatically assigned to each cluster string have sometimes gone astray. Mark Depauw is developing a new and improved version, especially in preparation for the much larger batch of personal names, where it is virtually impossible to manually correct everything that has gone wrong.

² For more information on TM Geo, Verreth, 2016.

Due to these problems I guess that we now have about only 80 to 90% of the toponyms occurring in all Latin inscriptions, but on the whole we are quite pleased with the result and in due time the remaining toponyms no doubt will find their way into the database also.

Phase 1, the identification of toponyms in the capital cluster strings, was finished in the beginning of July 2015, and we have now almost completed phase 2, the incorporation of the capital cluster file into the 'real' TM Georef file. For every place listed in TM Geo we try to list all the ancient text references where that place is attested. The file of these geographical references (TM Georef) is directly linked with the main TM Text file, so that every reference automatically receives a chronological and a geographical context. Every toponym found in the capital cluster file is exported to a separate Georef card. When a toponym exists out of several consecutive elements, like the *colonia* and *municipium* examples mentioned before, they are automatically grouped on one card. Twofold toponyms such as '*Bithynia et Pontus*', which cannot belong to the same capital cluster string because of the intermediary '*et*', are exported double and then afterwards joined manually. For the moment TM Georef lists 67.820 geographical references attested in Latin documentary texts and 10.474 references attested in Latin literary texts (but except for Egypt the latter have not yet systematically been entered). The Latin references make out almost 40% of the total of 196.794 Georef cards.

Fig. 11.2. The Georef Card.

In this phase we also start comparing the reading of the toponym in EDCS with the readings of the same passage in EDH, EDR, HEP and EDB, which can all be shown simultaneously in the Filemaker database.

In theory it is possible to automatically look for differences in readings among all these databases, but because each of them has its own approach, there will be so many small differences in line numbering, punctuation, the use of uncertainty dots and the way unconventional spelling in an inscription is indicated, that we doubt that there will be many exact matches. We therefore think that it will not be worth the enormous amount of work that it would involve. Human observation is again the answer and we do hope that our partners and the users of the geographical index will point out to us any mistakes we have made or obsolete readings we have kept. For the online version we have to talk with our partners whether they want to have their texts also shown on the TM page (like TM does for the texts from PN) or not. For the Open Access CC-0 texts in the Europeana EAGLE portal, this will in any case be implemented in the future. Anyway it is always possible to put a direct link on the Georef page to every partner that has the text in its corpus.

Since the users of TM must be able to look for specific spelling variants of each toponyms, all these references are presented the way they are on the stone, with as little additions or emendations as possible, except of course for any abbreviations at the end of the word; e.g. *T(h)ra[c(um)]* becomes *Tra[c(um)]*, and *Rom(a)e* becomes *Rome*. In another field the standardized nominative case is given, without brackets or uncertainty dots; e.g. *Trax* and *Roma*.

Fig. 11.3. Philadelpheia.

For every text in TM we try to give a reference to the most authoritative edition, where the user can find the best and most up to date reading and interpretation of that text. As authoritative editions we preferably use CIL, *Année épigraphique* and more recent major editions such as RIB, ILAlg, ICUR or I. Alex. Imp. Any corrections to the reading of the toponym with regard to that edition are to be listed in the field Bibliography, while the obsolete reading is recorded in the field Note. If the correction comes from one of the online full text databases, we add a reference to the number of the text in that database.

A major problem is the dating of the texts. Unfortunately not every edition provides a date for each inscription. Even if the scholar who publishes the text, has a fairly good idea of the century or range of time to which the text might belong, it is not always mentioned explicitly in the edition. For every Latin text for which TM did not receive a date from its partners, we added the broad range of 199 BC till AD 799, hoping that this dating will become more refined in the future.

The third and final phase involves the context of the toponym. In the field Detail we give a plain translation of the immediate phrase to which the toponym belongs. The translation should be a standardised as possible, with *termini technici* preferably added in Latin, so that the users can easily search for them in the database; e.g. 'Tiberius Iulius Martialis son of Tiberius of (the tribus) Claudia from Savaria, soldier (miles) of legio XV Apollinaris', 'praefectus Aegypti', 'cohors I Tungro-rum'. If the place is explicitly ascribed to a *provincia* or a region, this *provincia* is listed in the field 'Administrative situation'. If a town is explicitly called an *oppidum*, a *vicus* or a *civitas*, this information is listed in the field Status. By adding this information in searchable fields we hope that the user can start asking quite specific questions; e.g. the first and last attestation of a place in the sources; the periods in which a town was called a *colonia* or a *municipium*; the places in which the '*ala I Thracum Mauretana*' has been attested.

TM is a relational database, which implies that it is possible to get to the information from different angles. If someone is studying a certain text, he can get a list of all toponyms mentioned in that text. On the other hand, if someone is examining a certain place, he can find the list of all attestations for that place, in the order that he wants. In some cases a scholar can have very specific questions, which are difficult to search through the online TM search interface; it is quite well possible,

however, that these questions can be easily answered in the more complex Filemaker structure we have at our disposal; just contact us and we will try to solve the problem for you.

We are aware that this is a very succinct presentation of the new and exciting developments in Trismegistos Places, but everybody interested is always more than welcome to ask for more information. Please feel free to provide us with any addenda or corrigenda to the database you might have.

References

- BROUX, YANNE AND MARK DEPAUW. 2014. "Developing Onomastic Gazetteers and Prosopographies for the Ancient World Through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People." In *Social Informatics. SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014*, edited by L. M. Aiello and D. McFarland, 304–313. Berlin: Springer.
- VERRETH, HERBERT. 2016. "Topography of Egypt online". In *Proceedings of the 27th International Congress of Papyrology in Warsaw*. Forthcoming.

12. Augmenting the Workspace of Epigraphists. An interaction design study

Angelos Barmpoutis, Eleni Bozia***

Abstract

This paper presents the results of an interaction design study that focuses on the use of natural user interfaces for professionals in the fields of epigraphy and archaeology. This study proposes solutions for utilizing the sensors that can be found in popular handheld devices, such as tablets and smart phones, in order to naturally perform common tasks from the typical workflow of epigraphists. The developed interface allows the users to naturally hold digitized inscriptions, interact with them in order to relight or manipulate them as if they were real physical objects, and interact with metadata or other multi-modal data, such as text and images.

Keywords: Mobile Applications, Interaction Design, Natural User Interfaces, 3D models, Archaeology

12.1. Introduction

The technological advances in the last decade have equipped the general public with several handheld electronic commodities that changed significantly daily routine in a personal and professional level and contributed to the user's quality of life not only in developed countries but also in developing economies in Africa and Asia (Osman 2011). Handheld devices, such as tablets and smart phones, not only connect the users with tremendous amount of information through the internet, but also offer interfaces for natural user interaction that enable non-technology-oriented populations to use computers intuitively.

* Digital Worlds Institute, University of Florida.

** Department of Classics, University of Florida. Corresponding author. Email: bozia@ufl.edu

In the fields of epigraphy and archaeology, the areas of digital epigraphy and computational archaeology have benefited from the use of several of the sensors available in handheld devices. Crowd-sourcing of photographic data and geo-spatial information, augmented-reality navigation in archaeological spaces and museums, and 3D scanning of historical artifacts, using smart phones and tablet computers, are few of the exemplar applications of handheld sensors in epigraphy and archaeology. One common component in all the aforementioned applications is the ability to record tridimensional data either in the form of geo-spatial coordinates, or in the form of local 3D point coordinates needed for augmented-reality interaction, or for the construction of triangular meshes of 3D models.

There are several examples in literature that present 3D digitization projects that have been undertaken by museums including the Epigraphic Museum of Athens (Papadaki et al. 2015; Sullivan 2011), Museo Arqueológico Nacional de Madrid (Ramírez-Sánchez et al. 2014), Museo Nazionale Romano di Palazzo Altemps (Barmpoutis et al. 2015), Museo Geologico Giovanni Capellini di Bologna (Abate and Fanti 2014), National Museums Liverpool (Cooper et al. 2007), Smithsonian Institution (Wachowiak and Karas 2009), and several other museums and institutes (Gonizzi Barsanti and Guidi 2013; Landon and Seales 2006; Levoy et al. 2000).

Several novel methods for scanning, processing, and analyzing 3D models of inscriptions have been developed, including methods for text extraction from inscriptions (Aswatha et al. 2014; Sullivan 2011), accurate 3D scanning of inscriptions (Papadaki et al. 2015), visualization of inscriptions (Bozia et al. 2014), as well as 3D applications for other archaeological artifacts (Babeu 2011; Esteban and Schmitt 2004; Malzbender et al. 2001; Pollefeys et al. 2001). Comparative studies of 3D scanning methods for cultural heritage can be found in (Pavlidis et al. 2007) and (Böhler and Marbs 2004).

The aforementioned examples show that the use of 3D technologies in epigraphy and archaeology has been a well-studied topic over the past two decades. However, there is a notable disconnect between the research on these technologies and the actual use in the professional epigraphic and archaeological practice, as it has been hard for non technology-oriented audiences to handle and manipulate tridimensional data, using conventional computer equipment.

Furthermore, without mechanisms for proper user interaction, a 3D model that is projected on a 2D screen is not significantly advantageous compared to a set of 2D photographs.

The recent advances on Natural User Interfaces (NUI) along with their marketing as low-cost general-purpose devices (smart phones and tablets) have created a nurturing environment for integrating them in cultural heritage applications. Popular low-cost NUIs, such as touch screens, marker-less position trackers, motion sensors, and head-mounted displays have been recently studied and employed by museums as mechanisms for multi-sensory virtual experiences (Ujitoko and Hirota 2015; Soile et al. 2013; Ikei et al. 2015).

This paper tries to fill the gap between the 3D technologies and their actual professional application in the field of epigraphy by proposing innovative uses of NUIs specially designed to serve epigraphists. This is, to the best of our knowledge, the first systematic interaction design study in the field of epigraphy. This study proposes solutions for utilizing the sensors that can be found in popular handheld devices to naturally perform common tasks from the typical workflow of epigraphists. The developed interface allows the users to naturally hold digitized inscriptions and interact with them in order to relight or manipulate them, as if they were real physical objects, and also interact with metadata or multi-modal data, such as text and images.

12.2. Understanding the workflow of epigraphists

Understanding the users is one of the integral steps of interaction design, which is an iterative process during which representative users interact with preliminary designs and provide useful feedback (Preece et al. 2015). For the purposes of this study, our team interacted with early adopters of our prototype system, who were epigraphists and conservation specialists from Cornell University, the University of California, Berkeley, the University of Lyon 2, the Berlin-Brandenburg Academy of Sciences and Humanities, the U.K. National Archives, and the University of Florida. The goals of our interaction were twofold: a) to study the various forms of physical interaction that epigraphists have with an inscription as a real physical object and b) to expose epigraphists to a digital interface that imitates their interaction routine, using digital replicas of physical objects.

The first part of our study revealed 3 common types of interaction with the inscriptions as physical objects:

1. Change of point of view: Observation of the inscription from different viewing angles assists epigraphists understand better the shape of the inscribed letterforms.
2. Change of lighting conditions: Relighting the inscription by introducing artificial shadows or additional light sources from different angles may reveal details that were not legible in the original lighting conditions.
3. Magnification of inscribed details: Close observation of an inscribed region of interest, with or without artificial magnification, may assist epigraphists in assessing weathered fragments and make a better informed decision regarding the deciphering of the original text.

It should be noted that in addition to the above 3 types of interaction, there are two additional interactions that are special cases of I and II. More specifically, the physical object can be either portable (such as a small fragment of stone or other material) or not (when the inscription is on an inscription bearer). In the case of a handheld object, interactions I and II involve manual movement of the inscription with respect to the fixed observer (case I) or the fixed light source (case II), while in the case of large rigid objects the observer and the light source move with respect to the fixed inscription.

According to the above analysis, in the case of digitized inscriptions a NUI should provide the means for an epigraphist to “hold” the virtual object, “move” the point of view with respect to the virtual object, “manipulate” the virtual object with respect to the virtual light source, and “focus” on details of interest. The next section presents a NUI-based interaction design that proposes natural solutions to the aforementioned forms of interaction that seamlessly imitate the typical workflow of epigraphists.

12.3. Natural User Interface design for epigraphy

Natural User Interfaces consist of sensors that track the natural behavior of users and provide a natural form of interactivity with computers and other electronic devices.

The common forms of NUI sensors are: pressure sensors for sensing touch gestures (e.g. touch screens and touch pads), motion sensors for sensing user-initiated changes in the orientation and acceleration of the device (e.g. accelerometer, gyroscope, and compass), and position sensors for tracking changes in the relative position of the user with respect to the device, such as body motions (e.g. Microsoft's Kinect), finger motions (e.g. Occipital's Leap Motion), eye movements, and others.

An optimal interaction design solution should be intuitive, minimalistic, and non intrusive (Preece et al. 2015). Therefore, in order to design interaction for epigraphy one should choose devices that are easily accessible by epigraphists and do not interfere with their workspace (e.g. avoid introducing new devices or external sensors). All forms of interaction described in Sec. 1.2 can be implemented, using motion and pressure sensors, which can be easily found in tablet computers or smart phones. In both types of handheld devices the virtual object can also be assumed handheld, without loss of generality, in order to generate a multi-sensory experience for the user (i.e. holding the device = holding the digital inscription). Hence, NUI design is possible by utilizing accessible devices and without the use of external sensors as it is described in details in the following sections.

12.3.1. Natural interactive relighting of 3D models

In order to achieve natural interactive relighting of an inscription, the system should imitate the process of relighting a handheld physical object (such as a paper cast of inscription) by reorienting the object with respect to the light of the environment. Without loss of generality, we can assume that the default virtual lighting source is located on the ceiling, right above the device, which is also a very intuitive choice as it is the most probable real-world lighting condition. Under this assumption, a gyroscope, a sensor that tracks the orientation of the device with respect to the gravitational vector, is enough to track the slope of the device with respect to the virtual light. The top row of Fig. 1 shows the approximated real-world orientation of a tablet computer as it was estimated using the gyroscope of the device. The orientation is updated in real-time as the user moves the device.

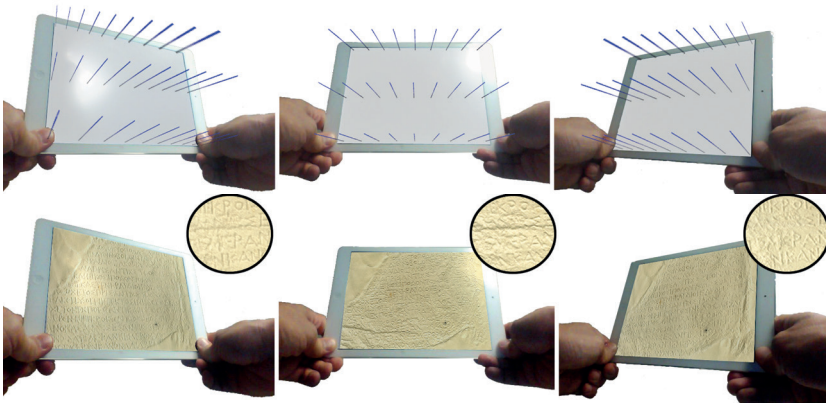


Fig. 12.1. Top row: Illustration of interactive manipulation of the virtual lighting by moving the device. The figures show the corresponding field of normal vectors in the 3D space as computed using the gyroscope of the device. Bottom row: Demonstration of interactive relighting of a 3D digitized inscription. Different virtual lighting angles reveal different inscribed details.

The estimated orientation of the device can be used in order to re-light the depicted 3D model of the inscription, using the angle of the device with the direction of the virtual light source. The bottom row of Fig. 1 demonstrates interactive relighting of a digitized paper cast. Epigraphists can relight an inscription by reorienting the tablet as if it were a real physical object. This process matches perfectly with the physical interaction of epigraphists with real inscriptions and can be extrapolated intuitively to the case of 3D models of large inscriptions that were not handheld in the real world (see Fig. 3).

12.3.2. Natural interactive manipulation of 3D models

Another important form of interaction in the epigraphic routine is the change of the point of view in order to understand better the structural details of the inscribed letterforms. Assuming that the model of the inscription is parallel to the screen of the device, the change of the point of view involves only change of the perspective projection of the digital object without any virtual rotation. In such case the rotation of the object is equivalent to the physical rotation of the device without any virtual rotation of the object.

Fig. 2 shows 15 different projections of the same virtual cube that correspond to the change of perspective caused by moving the observer's head parallel to the screen. Note that all cubes are parallel to each other.

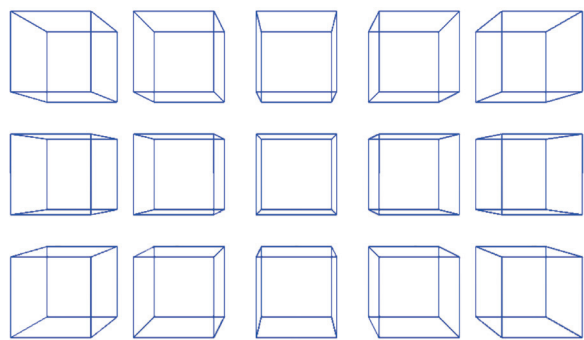


Fig. 12.2. Visualization of same-size boxes using 15 different perspective projections with the same FOV angle and different cropping parameters. None of the boxes is rotated in the space.

In the case of a tablet computer, the change of perspective can be implemented using the accelerometer of the device, which senses non-gravitational accelerations in the 3D space. With the logical assumptions that: a) the tablet is initially facing the user, and b) the user’s eyes remain in a relatively fixed position in the 3D space (otherwise an eye-tracker should be required), the change of perspective can be realistically achieved by naturally reorienting the tablet as shown in the top row of Fig. 3. The superimposed boxes in this figure were estimated, using the accelerometer reading of the device for three different orientations of the tablet.

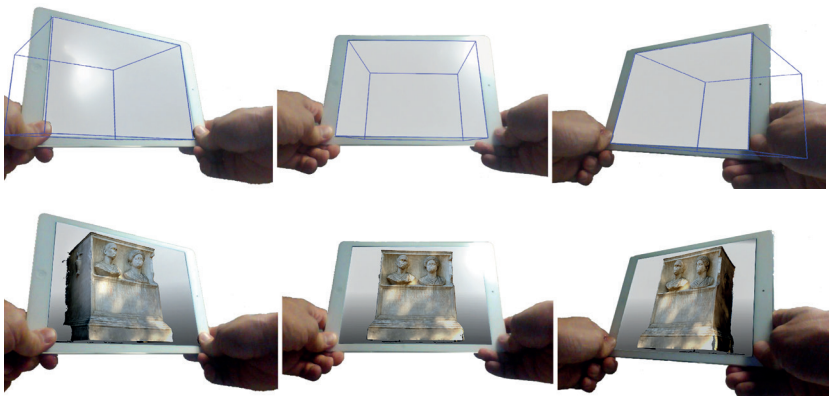


Fig. 12.3. Top row: Illustration of interactive manipulation of the perspective projection by moving the device and using the “fixed eye” assumption. Bottom row: Demonstration of interactive visual inspection of a 3D digitized inscription along with the inscription bearer. The user can view the object from different perspectives, using natural motions.

Interactive manipulation of a 3D digitized inscription bearer is shown in the bottom row of Fig. 3. Different sides of the bearer can be observed by reorienting the tablet naturally. In this example a large inscription model with its bearer was chosen in order to demonstrate that the proposed interaction design remains intuitive independently of the scale. It should also be noted that the interactive manipulation of the perspective can be optionally performed simultaneously with the interactive relighting as shown in Fig. 1 (bottom) in order to perform a more realistic interaction that causes relighting and change of point of view at the same time.

12.3.3. “Touching” the metadata: Interacting with multi-modal data

The forms of interactions presented in 1.3.1 and 1.3.2 involved only the motion sensors of a tablet computer without utilizing the pressure sensors of the screen of the device. The commonly used touch gestures (such as 2-finger twist to rotate, 2-finger pinch to zoom, and 2-finger translate to move) can be employed in order to enhance the proposed NUI design. In addition to the aforementioned gestures, a tap gesture could activate regions of interest with additional modalities of information, such as text, images, and metadata. The user can interactively browse the different forms of data by using intuitive touch gestures as shown in Fig. 4. This set of 2D interactions along with the 3D NUI design presented earlier can compose an intuitive yet powerful workspace for an epigraphist who can now perform digitally several parts of the epigraphic workflow.



Fig. 12.4. Screenshot of our interactive environment. The user interacts with the 3D object, using touch gestures and selects one of the regions of interests. This action initiates other data tools, such as the image viewer or the edge filter as shown in this example.

12.4. Conclusions and future directions

In this pilot study, a complete set of natural user interactions was designed based on the physical interactions of epigraphists with real inscriptions. The proposed interactions utilize the existing sensors in a typical tablet computer or smart phone in order to interactively relight a digitized inscription and manipulate the user's perspective, using a set of intuitive gestures that imitate the natural interaction with a physical object. In the proposed design, the epigraphist can "hold" a digital inscription, relight it by reorienting it as a tangible object, observe it from different perspectives, and finally interact with other modalities by following a set of 2D touch gestures. The prototype system was developed as part of the Digital Epigraphy and Archaeology (DEA) project, using the open-source library VisiNeat for 3D visualization and interaction, and is compatible with iOS, Android, and Microsoft RT tablet and smart phone devices. The interface is available through the web-site of the project: <http://www.digitalepigraphy.org>.

In the future, we plan to quantitatively evaluate the designed interface by tracking the user activities and analyze their motion patterns in the 3D space while they are interacting with their handheld device.

Acknowledgement

We would like to acknowledge our collaborators for their invaluable interaction and feedback during the past three years. We would also like to thank the UF College of the Arts and the UF Center for Greek Studies for providing continuous funding support for this project.

References

- ABATE, DANTE and FEDERICO FANTI. 2014. "La valorizzazione digitale del Museo Geologico Giovanni Capellini di Bologna." *Archeomatica* 5(1). <http://mediageo.it/ojs/index.php/archeomatica/article/view/319>.
- ASWATHA, SHASHAANK M., ANANTH NATH TALLA, JAYANTA MUKHOPADHYAY, and PARTHA BHOWMICK. 2014. "A Method for Extracting Text from Stone Inscriptions Using Character Spotting." In *Computer Vision-ACCV 2014 Workshops*, 598–611. Berlin: Springer. http://vigir.missouri.edu/~gdesouza/Research/Conference_CDs/ACCV_2014/pages/workshop13/pdffiles/w13-p5.pdf.

- BABEU, ALISON. 2011. "Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics. Council on Library and Information Resources. <https://www.clir.org/pubs/reports/reports/pub150/pub150.pdf>.
- BARMPOUTIS, ANGELOS, ELENI BOZIA, and DANIELE FORTUNA. 2015. "Interactive 3D Digitization, Retrieval, and Analysis of Ancient Sculptures, Using Infra-red Depth Sensors for Mobile Devices." In *Universal Access in Human-Computer Interaction. Access to the Human Environment and Culture*, 3–11. Berlin: Springer. https://link.springer.com/chapter/10.1007/978-3-319-20687-5_1.
- BÖHLER, WOLFGANG and ANDREAS MARBS. 2004. "3D scanning and photogrammetry for heritage recording: a comparison." In *Proceedings of the 12th International Conference on Geoinformatics*, 291–298. Citeaser. <http://docplayer.net/29841597-3d-scanning-and-photogrammetry-for-heritage-recording-a-comparison.html>.
- BOZIA, ELENI, ANGELOS BARMPOUTIS, and ROBERT WAGMAN. 2014. "Open-Access Epigraphy: Electronic Dissemination of 3D-digitized Archaeological Material." In *Proceedings of the International Conference on Information Technologies for Epigraphy and Cultural Heritage*, edited by Silvia Orlandi, Raffaella Santucci, Vittore Casarosa, and Pietro Maria Liuzzo, 421–435. Rome: La Sapienza University Press.
- COOPER, M., A. LA PENSÉE, and P. BRYAN. 2007. "Chiswick House, London: Laser scanning tests on a gilded 18th century table." *English Heritage Research Department Series Report* 19.
- ESTEBAN, CARLOS HERNÁNDEZ and FRANCIS SCHMITT. 2004. "Silhouette and stereo fusion for 3D object modeling." *Computer Vision and Image Understanding* 96(3): 367–392. <http://www.sciencedirect.com/science/article/pii/S1077314204000542>.
- GONIZZI BARSANTI, S. and G. GUIDI. 2013. "3D digitization of museum content within the 3D-ICONS project." *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-5 W 1: 151–156. <http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/II-5-W1/151/2013/isprsanals-II-5-W1-151-2013.pdf>.
- IKEI, YASUSHI, SEIYA SHIMABUKURO, SHUNKI KATO, KOHEI KOMASE, KOICHI HIROTA, TOMOHIRO AMEMIYA, and MICHITERU KITAZAKI. 2015. "Experience Simulator for the Digital Museum." In *Human Interface and the Management of Information. Information and Knowledge in Context*, 436–446. Berlin: Springer. URL [http://link.springer.com/chapter/10.1007/978-3-319-20618-9/s\do5\(4\)4](http://link.springer.com/chapter/10.1007/978-3-319-20618-9/s\do5(4)4).
- LANDON, GEORGE V. AND W. BRENT SEALES. 2006. "Petroglyph digitization: enabling cultural heritage scholarship." *Machine Vision and Applications* 17(6): 361–371. URL <http://link.springer.com/article/10.1007/s00138-006-0044-0>.
- LEVOY, MARC, KARI PULLI, BRIAN CURLESS, SZYMON RUSINKIEWICZ, DAVID KOLLER, LUCAS PEREIRA, MATT GINTON, SEAN ANDERSON, JAMES DAVIS, JEREMY GINSBERG, and OTHERS. 2000. "The digital Michelangelo project: 3D scanning of large statues." In *Proceedings of the 27th annual conference on*

- Computer graphics and interactive techniques*, 131-144. ACM Press/Addison-Wesley Publishing. <http://dl.acm.org/citation.cfm?id=344849>.
- MALZBENDER, TOM, DAN GELB, and HANS WOLTERS. 2001. "Polynomial texture maps." In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 519-528. ACM. <http://dl.acm.org/citation.cfm?id=383320>.
- OSMAN, MOHD AZAM, ABDULLAH ZAWAWI TALIB, ZAINAL ABIDIN SANUSI, TAN SHIENG YEN, and ABDULLAH SANI ALWI. 2011. "An exploratory study on the trend of smartphone usage in a developing country." In *Digital enterprise and information systems*, 387-396. Berlin: Springer. https://link.springer.com/chapter/10.1007/978-3-642-22603-8_35.
- PAPADAKI, ALEXANDRA, PANAGIOTIS Agraftotis, ANDREAS GEORGOPOULOS, and SEBASTIAN PRIGNITZ. 2015. "Accurate 3D Scanning of Damaged Ancient Greek Inscriptions for Revealing Weathered Letters." *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 1: 237-243. <http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-5-W4/237/2015/isprsarchives-XL-5-W4-237-2015.pdf>.
- PAVLIDIS, GEORGE, ANESTIS KOUTSOUDIS, FOTIS ARNAOUTOGLOU, VASSILIOS TSIUKAS, and CHRISTODOULOS CHAMZAS. 2007. "Methods for 3D digitization of cultural heritage." *Journal of cultural heritage* 8(1): 93-98. <http://www.sciencedirect.com/science/article/pii/S129620740600121X>.
- POLLEFEYS, MARC, LUC VAN GOOL, MAARTEN VERGAUWEN, KURT CORNELIS, FRANK VERBIEST, and JAN TOPS. 2001. "Image-based 3D acquisition of archaeological heritage and applications." In *Proceedings of the 2001 conference on Virtual reality, archeology, and cultural heritage*, 255-262. ACM. URL <http://dl.acm.org/citation.cfm?id=585033>.
- PREECE, JENNY, HELEN SHARP, and YVONNE ROGERS. 2015. *Interaction Design-beyond human-computer interaction*. New York: John Wiley & Sons.
- RAMÍREZ-SÁNCHEZ, MANUEL, JOSÉ-PABLO SUÁREZ-RIVERO, and MARÍA-ANGELES CASTELLANO-HERNÁNDEZ. 2014. "Epigrafía digital: tecnología 3D de bajo coste para la digitalización de inscripciones y su acceso desde ordenadores y dispositivos móviles." *El profesional de la información* 23(5): 467-474. <http://eprints.rclis.org/23920/>.
- SOILE, SOPHIA, KATERINA ADAM, CHARALABOS IOANNIDIS, and ANDREAS GEORGOPOULOS. 2013. "Accurate 3d Textured Models of Vessels for the Improvement of the Educational Tools of a Museum." *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 1(1): 211-217.
- SULLIVAN, STEPHANIE MARIE. 2011. *Analytical methods for the extraction of content from high resolution 3D data sets: Epigraphical applications to the Drakon Stele*. Ph.D. thesis, University of Arkansas. https://www.researchgate.net/publication/235618425_Analytical_Methods_for_the_Extraction_of_Content_from_High_Resolution_3D_Data_Sets_Epigraphical_Applications_to_the_Drakon_Stele.

- UJITOKO, YUSUKE and KOICHI HIROTA. 2015. "Application of the Locomotion Interface using Anthropomorphic Finger Motion." In *Human Interface and the Management of Information. Information and Knowledge in Context*, 666–674. Berlin: Springer. https://link.springer.com/chapter/10.1007/978-3-319-20618-9_65.
- WACHOWIAK, MELVIN J. and BASILIKI VICKY KARAS. 2009. "3D scanning and replication for museum and cultural heritage applications." *Journal of the American Institute for Conservation* 48(2): 141–158.

13. TIGLIO. Translations and Images of Greek and Latin Inscriptions Online

*Almas Bridget, Baumann Ryan, Beaulieu Marie-Claire, Cayless Hugh
Cowey James, Liuzzo Pietro, McCourt Finlay, Sosin Joshua**

Abstract

This paper describes the aims of a project funded by The Andrew W. Mellon Foundation to speculate on the best ways to deal with two forgotten types of content in the realm of ancient epigraphy: translations and images. Translations available online are less than 3% of the total available transcriptions online; images are often subject to policies which make extremely difficult their use for research, publication, even simple viewing. There has been no thought given to these before in an articulated manner although these are the types of content which can bring a much larger group of users to epigraphy.

Keywords: Translations, Images, Greek and Latin, Epigraphy, Online

13.1. Introduction

Nearly all Greek and Latin epigraphic texts are available, sometimes in multiple versions, after three decades of continuous digitization and online publication. Without counting repeated inscriptions and *instrumenta*, there are about 300.000 Latin inscriptions¹ online and about

* Almas Bridget, Tufts University (USA); Baumann Ryan, Duke Collaboratory for Classics Computing (USA); Beaulieu Marie-Claire, Tufts University (USA); Cayless Hugh, Duke Collaboratory for Classics Computing (USA); Cowey James, Ruprecht-Karls-Universität Heidelberg; Liuzzo Pietro, Ruprecht-Karls-Universität Heidelberg, corresponding author, email: pietro.liuzzo@zaw.uni-heidelberg.de; McCourt Finlay, Glasgow University - Attic Inscriptions Online (UK); Sosin Joshua Duke Collaboratory for Classics Computing (USA).

¹ EAGLE disambiguated total: 235.626; EDCS not disambiguated total without *instrumenta* and *inscriptiones christianae*: 308.581. On the definition of Roman Epigraphy, see Panciera (2012).

200.000 Greek Inscriptions². In some databases there is also abundant metadata and a structured bibliography. The situation for translations and images of these inscriptions (text and support) available in the digital space is nevertheless quite different. The ratio between images of inscriptions and text is of one image every two inscriptions,³ but those inscriptions which have photographic documentation usually have many photos.⁴ Translations are present only in smaller corpora edited online in most cases and the Attic Inscriptions Online project⁵ is an *uniquum* in its intent to provide mainly translations of inscriptions.⁶ There are many publications offering translations in print, but these are not published online. An estimate calculation, based on the 11.000 translations present in the EAGLE Media Wiki, and known collections of printed translations of inscriptions, compared to a total of texts usefully translatable of around 300,000 texts, brings to an alarming 10% of translated texts, of which only a third (slightly more than 3%) is online. Translations are perhaps not a priority for researchers who know Greek and Latin, but are a way to clarify the interpretation of a text and an invaluable tool for didactical purposes and teaching: they are the only way in which an inscription can reach a wider public in a significant way as part of cultural heritage. The same could be said for images, even more obviously, since researchers also need them because: 1) they cannot always reach the place where an inscription is stored to study it (given that the inscription is still there); 2) there are cases in which a photo might be all we are left with and these are quickly increasing as monuments get lost or are destroyed. The imbalance in the documentation is thus pressing, since translations and images are our two best controls on the constitution

² Data from the latest Integrating Digital Epigraphies's (https://youtu.be/OPfDj_hjeok) harvest from the Packard Humanities Institute, Searchable Greek Inscriptions (<http://epigraphy.packhum.org/>), with some duplicates, 207.964.

³ At the time in which this paper is being written () the Epigraphic Database Heidelberg has ca. 35.000 photos and ca. 71.000 texts (0,5); The Epigraphic Database Roma has a slightly better ratio with 45.000 photos for 71.000 texts (0,6); The Epigraphic Database Bari has 34538 texts and 10341 images (0,3). In the EAGLE aggregator, the total ratio (excluding the related content of Arachne), is of 0,79 images for each text (235626 documental entities per 185999 visual entities) because smaller corpora tend to have a better photographic documentation.

⁴ The Epigraphic Database Heidelberg has to date ca. 14.000 records with a photo or a drawing attached, bringing the average number of images per inscription to 2,5.

⁵ <https://www.atticinscriptions.com>.

⁶ Lambert and McCourt 2014.

and interpretation of ancient documentary texts. To an extent, digital epigraphy today is the direct descendant of epigraphy's 19th century analog self: many texts, few translations, few images. This project aims to take initial steps to redress that imbalance, building resources that allow epigraphists and ancient historians to bring translations and images more closely into the suite of existing digital epigraphy resources.

13.2. Problems

Let's look at the data we have. The EAGLE project has gathered some insights on small collections of images of inscriptions openly published online, on Wikimedia Commons,⁷ and on a set of translations of inscriptions, collected in the EAGLE Media Wiki.⁸ We shall compare their impact and reach on the wider public to that of the texts of inscriptions, to underline the urgency for these materials to be produced also in order to bring epigraphy outside its restricted circles. Let's look at the visits to the Epigraphic Database Clausss-Slaby, the largest collections of texts (with minimal metadata and no directly stored image): EDCS has an average 3000 requests per day.⁹ The result page is always one, containing all the results from the database, which has a total of 491353 texts. We have no means to provide better data unfortunately but for the comparison these will be enough. The images collected under the category "Media Contributed by EAGLE" on Wikimedia Commons contains instead around 8000 photos of inscriptions and we have some good insights on this data¹⁰. These photos have been viewed in 19 months 22,236,085 times. Another interesting information is the number of people who have worked on them, by no means only members of the EAGLE project: 7 users have made more than 1000 edits, which could be anything above the figure; 20 have made between 100 and 1000 edits; 71 have made between 10 and 100, and even more interestingly 581 have made between 1 and 10 edits. This is a critical mass of active users, uploading, editing, curating, using

⁷ https://commons.wikimedia.org/wiki/Category:Media_contributed_by_EAGLE

⁸ Liuzzo et al. 2014 http://www.eagle-network.eu/wiki/index.php/Main_Page

⁹ This can be easily monitored looking at the counter on the website at the end of each day. No better statistics are available.

¹⁰ Thanks to user:Fae and the authors of the BaGLAMa2 tool. See https://commons.wikimedia.org/wiki/Category_talk:Media_contributed_by_EAGLE/reports and <https://tools.wmflabs.org/glamtools/baglamma2/#gid=148&month=201508>.

data they are interested in the same situation can be noted for the 11.000 translations in the EAGLE Media Wiki, which were viewed in 18 months 1.380.000 times and have seen 280 active users, who have at least made 1 edit. The tool is not well known outside the EAGLE consortium and is a very small prototype, but the fact that it has already attracted such a mass of views is significant. What would happen if we gave 100.000 translations in the way we have given them to the people in the Mediawiki, completely openly? What would happen if those people contributing to images and translations were empowered to operate easily and intuitively to enter more and more data? Inscriptions will never get as many fans as we would like to, but perhaps their content and related resources would be a bit more accessible to non-initiated. This comparison confirms also that the usability of resources is measured at a different level when they are made open, and that images and translations have an undeniable higher relevance as an online resource, thus attracting interest also to the transcriptions, while this does not happen the other way around and only those who know what they are looking for will stumble upon an ancient inscription published online. Nothing new in these observations, but this obvious observation is in contrast with the actual situation in which photos are few and translation even fewer.¹¹ Why so, then? The possible reasons are:

1. the lack of an entry point which is easy to access and use;
2. people get easily worried by copyright due diligence and find difficult sometime to track back who is the author or the copyright owner of a photo;
3. a lack of coordinated effort, planning and management of the storage of both these types of content;
4. researchers working on inscriptions identify their intended audience in a very specific academic community which does not need translations and instead needs edited texts (transcriptions and metadata);
5. publication of content with (sometimes unnecessary) restrictions;
6. lack of time for this effort, unrecognized in academic settings as a contribution to the progress of knowledge, as, sadly, most other digital efforts.

¹¹ Photographs are very useful for simple stones inscribed monuments in large lettering (as e.g. Roman monumental texts), but for most longer documentary inscriptions, including most Attic material, after autopsy, squeezes are the most important editorial tool, because photos are mostly 2 D images of 3 D objects.

We shall point out what has been done to solve these problems and cater for an improvement in this part of documentation and production of online content in the future.

The international group of partners, which includes University of Heidelberg, University of Cardiff, Duke University and Tufts University is holding regular workshop meetings to design and develop a suite of resources that support generation of epigraphic translations, with peer-review and publication workflows supported by Perseids' extension of the Son of the Suda On Line code (SoSOL), with publication supported by the EAGLE Mediawiki, and image management, reference ontology, geo- and other services, supported by Integrating Digital Epigraphies, and with Attic Inscription Online translations as the key content stream for development and testing.

13.3. Translations

To face these challenges with regard to translations, it is the opinion of the project team that tying together existing resources is a better way to tackle the issues rather than trying to superimpose a new tool or system. The available building blocks for such systematization of existing resources are the following:

- the existing local data entry point of Attic Inscriptions Online, in the process of moving to TEI-EpiDoc markup for the underlying data;
- the EAGLE Mediawiki, with the Wikibase Extension, which collects translations from several users as a part of the work of the EAGLE consortium to bring epigraphy to a wider public;
- the Perseids peer review system,¹² which uses the Son of the Suda Online¹³;
- Leiden+,¹⁴ a simplified markup which allows the use of normal diacritics instead of tags to enter XML markup;
- identification and disambiguation done by content providers (the epigraphic databases) and by projects such as Trismegistos¹⁵ for the members of the EAGLE consortium and IDEs¹⁶ for Greek Epigraphy projects;

¹² <http://sites.tufts.edu/perseids/>.

¹³ Baumann 2013.

¹⁴ <http://papyri.info/docs/about> and J. Sosin presentation at <http://www.stoa.org/archives/1263>.

¹⁵ <http://www.trismegistos.org/>.

¹⁶ <http://blogs.library.duke.edu/dctthree/projects/>.

- referencing and resolution services provided by IDEs which do not just align content relating to one resource but describe the relation among them

These consider two kinds of users:

- users involved in a project with access to a data entry point in a database (using XML);
- independent contributors.

With the available building blocks what can be done currently for translations is:

- a standalone javascript library to enter translations using Leiden+ (to be implemented and tested in AIO, as the best candidate for its focus on translations) which implies;
- an enlargement of the encoding guidance and conversion capabilities of Leiden+ to EpiDoc for translations;¹⁷
- recommendation on how to mark up translations for the EpiDoc guidelines.

These developments will hopefully be beneficial to the EpiDoc users community as well, which has in the past asked for more guidance on how to encode translations.

- An export of AIO in EpiDoc to the Perseids platform in which translations will be peer-reviewed for ingestion in the EAGLE wiki, bypassing the harvest process for EAGLE. This might be useful as a use case for future project willing to publish their translations with the others collected into the EAGLE Media Wiki;
- facilitate flow between existing tools and services;
- EAGLE and Perseids worked together in the past years in order to integrate the two services offered, but this had a number of limitations, e.g. the requirement for a translation to be already present in the wiki in order to be able to publish another one via Perseids. This reduced the number of items for which the integration could be used to a minimum, forcing the use of workarounds as mock text and placeholders. The integration of the EAGLE wiki with Perseids will now enable users to enter translations for any inscriptions and even a new translation from scratch with a specific new language, thus covering all possibilities for the EAGLE Mediawiki user.

¹⁷ <https://github.com/TIGLIOPROJECT/documentation/wiki>.

This requires nevertheless:

- unique identifiers for Latin and Greek texts, which are currently provided for the first by Trismegistos, and for the second group of texts by IDEs;
- a citation URN structure which is agreed upon and otherwise usable. This will be based on the scheme already in use for EAGLE built on CTS URN syntax as in the example:

urn:cts:pdlepi:eagle.tm12345.perseids-translation-1

where the structure is

urn:cts:namespace:textgroup.work.version

- The complete workflow from data entry to publication in a website and to the common EAGLE resource via Perseids will then be a complete and replicable workflow, scalable for use from larger projects and documented to guarantee easy and sensible connection of the resources online.

The workflow for the connection of new translations and images to existing online content will be then facilitated in this way. A project with its own data entry interface should be able to use the javascript library to enter translations using Leiden+ and following the conventions already extended and public in papyri.info. They should then be able to identify with a TM or IDEs id these translations and infer a URN to push these directly to the Perseids system. Here the translations will be peer-reviewed and then sent both back to the source with an approved status and to the EAGLE Media Wiki. If this database is partner of the EAGLE project the texts will be harvested separately and the translations linked back from the Media Wiki. From the Media Wiki API they will also be available as such to external users. An independent contributor instead will be able to look up the TM text id or IDeSt of the inscriptions he wants to translate and enter it to the EAGLE Media Wiki. From here this will be sent to the Perseids system and returned as described above.

The short term goal is therefore to stitch together existing resources already in development, and especially to provide ids and a clear citation syntax for all available inscriptions, which will have counter benefits also for any other digital project with these requirements.

13.4. Images

Most of the existing images of inscriptions are currently safely stored in private computers. Large collections of images of inscriptions are available at the major databases, and can count up to thousands of photos of inscriptions or drawings. Large collections of images are also on Flickr (e.g. Visible Words) and Wikimedia Commons (CIL and AE categories provide a good overview of what is available).

The major problem preventing publication is that people do not know if they can share the images which they have. The copyright regulations are too complicated and possible contributors opt for doing nothing instead of taking any unknown and unwanted risk. Storage of images of inscriptions in the major databases happens under very strict conditions of reuse and publication and while it is the best possible way to operate for these projects, there can be no mirroring or distribution of the resources available so that the life expectancies of content curated for decades is tied to the lifetime of these databases. A test has been done with the images of non identified items in the Epigraphic Database Heidelberg, uploaded to Wikimedia Commons and users have contributed to identify a number of those. Uploading photos on Wikimedia Commons requires an open license, which cannot always be guaranteed, whereas on Flickr it is possible to retain rights whilst publishing the photos, so that it makes a nicer tool for this kind of content, although it does not allow community editing and batch upload, which is instead possible via tools developed by Europeana and the Wikimedia Foundation for Commons.¹⁸ A unique repository or a connection hub for all these photos would be a solution to keep the content under sight but this would still require control over resources daily published in Flickr and Commons, an involvement into the communities of users online, together with the extension of citation structures to these groups. There is no easy solution for the copyright side of the problem, but a continued encouragement to share will build towards

¹⁸ The GLAM Wiki toolset, https://commons.wikimedia.org/wiki/Commons:GLAMwiki_Toolset_Project.

the critical mass needed for a change in perspective on this issues in the coming few years. During this project we will try to list requirements for a tool to ease out the upload of images online, which will match images and metadata provided in various formats, suggesting ids and adding them in a format compliant to the citation scheme agreed. It is in fact true that the amount of time required to use tools which ask for a one to one upload of images is another important factor which reduces the amount of content shared even from those willing to do so.

13.5. User involvement and expert sourcing

The problems of all digital projects looking at putting together materials from various sources and contributors are on one side to get people involved, on the other to overlook the work done and take care of the administration. The possible user scenarios are eventually many more. Two especially deserve to be mentioned here. Most input has been provided in term of new translations in the EAGLE Media Wiki during Workshops and Secondary Schools class work. Some teachers with a background in epigraphy have been contacted by the Epigraphic Database Rome to start and experiment with a didactic model which would include translating inscriptions. The students worked on a specific corpus of inscriptions, studied the text and the support and produced a translation which they entered in the EAGLE Media Wiki with the supervision of their teacher. This experience has proved successful for the students which have seen their contribution directly where it should be, together with other scholarly content. On the other side, every new translation matters in such a state as the one described above. The second example is the work done in two consecutive workshops, held in Ercolano and at the Centre for Hellenic Studies in Washington DC of the Ancient Graffiti Project,¹⁹ which has published multiple translations for all the graffiti of Herculaneum. In this context the usability of the Wikibase software has been tested and it proved to be an extremely intuitive and powerful tool. It takes very little explanation, but there are caveats for this simplicity and namely that it is extremely easy to do things in slightly creative ways, as entering statements as source information or typing an id in a slightly different way, which then need to be monitored and fixed.

¹⁹ <http://ancientgraffiti.wlu.edu/>.

13.6. Conclusions

Some of the tasks above have been already carried out, some are under way, but there are some general conclusions which can be summarized. There is a need to unify the resources, agree on standards for reference and citation and provide stable identifiers and citation structures, providing a comprehensive list of epigraphic publications with the relevant abbreviation in use. Although some work has been done, the amount of data makes this task continuously needed together with that of disambiguation. Other efforts need to go in the direction of flexible but harmonized standards for encoding and working on data entry giving priorities probably in a slightly different way as before, updating the tools to be able to cope. More generally, while tools are abundant and so are guidelines and cookbooks, an agreed venue for coordination of the efforts is still a desideratum, and should include not only scholarly project but also community based efforts such as those of the Wikimedia Commons users and of the Flickr user's groups. These people could be also part of the peer review process, thus contrasting the side effects of an inactive board.²⁰

Aknowledgements

The Andrew W. Mellon Foundation has funded the meetings of the project TIGLIO and the work which will be carried out for Attic Inscriptions Online. We would like to thank also all the Wikimedia users who have contributed to parts of this work in various ways and especially Aubrey, CristianCantoro, Fæ, Laurentius, Magnus Manske, Sannita and Wittylama. Another special aknowledgement goes to all users of the EAGLE Mediawiki and to all the volunteers working on photos of inscriptions online.

²⁰ The last meeting of the project, dealing with images and CTS URN structure still had to take place at the date of submission of the present contribution.

References

- BAUMANN, RYAN. 2013. "The Son of Suda On-Line." In *The Digital Classicist 2013*, edited by Stuart Dunn and Simon Mahony. London: The Institute of Classical Studies University of London. <http://ryanfb.github.io/papers-BICS/sosol-bics-draft.pdf>.
- LAMBERT, STEPHEN and FINLAY McCOURT. 2014. "Attic Inscriptions Online." In *Information Technologies for Epigraphy and Cultural Heritage*, edited by Silvia Orlandi, Vittore Casarosa, Raffaella Santucci, and Pietro Maria Liuzzo, 155–166. Rome: Sapienza Università Editrice. <http://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf>.
- LIUZZO, PIETRO MARIA, ANDREA ZANNI, LUCA MARTINELLI, LORENZO LOSA, and PIETRO DE NICOLAO. 2014. "The EAGLE Mediawiki. A fully collaborative database for academics, data engineers and the general public." In *Information Technologies for Epigraphy and Cultural Heritage*, edited by Silvia Orlandi, Vittore Casarosa, Raffaella Santucci, and Pietro Maria Liuzzo, 187–200. Roma: Sapienza Università Editrice. <http://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf>.
- PANCIERA, SILVIO. 2012. "What Is an Inscription? Problems of Definition and Identity of an Historical Source." *Zeitschrift für Papyrologie und Epigraphik* 183: 1–10. <http://www.digitalmeetsculture.net/wp-content/uploads/2013/10/Panciera-Inscription-ZPE-2012.pdf>.

14. A Virtual Research Environment to Document and Analyze Non-alphabetic Writing Systems. A Case Study for Maya Writing

*Katja Diederichs, Sven Gronemeyer, Christian Prager, Elisabeth Wagner, Franziska Diehr, Maximilian Brodhun, Nikolai Grube**

Abstract

No existing digital work environment can sufficiently represent the traditional epigraphic workflow ‘documentation, analysis, interpretation, publication’ for a non-alphabetic writing system. Using the Maya hieroglyphic script, this workflow will be transferred to a digital epigraphy. Digital methods and tools will be developed and reused in a Virtual Research Environment to create a freely accessible, annotated textual corpus, including metadata on cultural and object history and references.

Keywords: Classic Mayan, Digital Epigraphy, TEI, EpiDoc, CIDOC-CRM, Metadata Schema, Open Science

14.1. Digitalizing the epigraphic workflow

In order to develop a digital work environment that represents the traditional epigraphic workflow for documenting and studying inscriptions in a non-alphabetic script, the individual steps of this workflow need to be technically implemented using methods and tools from the Digital Humanities. The workflow begins with the documentation of the text carriers and compilation of descriptive data; proceeds to epigraphic analysis, including sign classification and transliteration

^a Katja Diederichs, Rheinische Friedrich-Wilhelms-Universität Bonn. Corresponding author. Email: diederichs@uni-bonn.de; Sven Gronemeyer, Rheinische Friedrich-Wilhelms-Universität Bonn. La Trobe University Melbourne; Christian Prager, Rheinische Friedrich-Wilhelms-Universität Bonn; Elisabeth Wagner, Rheinische Friedrich-Wilhelms-Universität Bonn; Franziska Diehr, Niedersächsische Staats- und Universitätsbibliothek Göttingen; Maximilian Brodhun, Niedersächsische Staats- und Universitätsbibliothek Göttingen; Nikolai Grube, Rheinische Friedrich-Wilhelms-Universität Bonn.

and transcription of the texts; continues with morphological segmentation and linguistic interpretation; and concludes with translation and digital publication of the inscription (Diederichs 2015).

The Virtual Research Environment (VRE) TextGrid initially provides the necessary framework for executing the workflow that technically facilitates documentation and digital registration of the text carriers in the form of images and drawings, as well as annotation of the objects with metadata. TextGrid offers input forms, annotation tools, and a Text-Image-Link-Editor, and provides long-term storage, and free access to the data in an open web repository (Neuroth *et al.* 2011).

Although more detailed steps in an epigraphic workflow, such as detailed linguistic and epigraphic analysis of primary texts, can be technically carried out within a VRE such as TextGrid, they still necessitate creation and adaptation of their own XML-based metadata model, as well as annotation schemas for object and textual contexts (Prager 2015). Using this approach, the Maya hieroglyphic script will for the first time be able to fulfill a central requirement of corpus linguistics, namely machine readability (McEnery and Wilson 2001).

Fig. 1 provides a schematic illustration of the following points that represents in great detail the steps in epigraphically analyzing Maya inscriptions, a process which can also be used in modified form to study other non-alphabetic writing systems. The modulated VRE is thus applicable not only to the study of Maya texts, but also to researching texts composed in other hieroglyphic, cuneiform, or linear writing systems.

1. Identification of hieroglyph blocks by alphanumeric classification
2. Original spelling
3. Reading order of individual signs
4. Number of signs within one block
5. Identification and isolation of signs in one block
6. Classification of signs based on the sign catalog by Eric Thompson
7. Classification of signs in one block
8. Transliteration of individual signs
9. Description of sign function (syllables, logographs)
10. Transliteration
11. Broad transcription
12. Morphological segmentation
13. Morphological analysis
14. Determining congruence between block- and word-boundaries
15. Determining syntactic function















1	D13			E13					D14		
2											
3	type A			type B					type A		
4	3			5					3		
5											
6 ⁶	T264	T21	T575	T1	T257	T1	T624	T178	T74	T712°504	T178
7	T264:21:575			T1:257.1:624:178					T74:712°504:178		
8	ju	bu	yi	U	TOK'	U	PAK	la	ma	CH'AB° AK'	la
9	syllable	syllable	syllable	logograph	logograph	logograph	logograph	syllable	syllable	logograph	syllable
10	ju-bu-yi			U-TOK' U-PAK-la					MA' CH'AB'AK'-la		
11	jubuy			utok' upakal					ma' ch'ab [ma'] ak'al		
12	jub-uy-ø			u-tok' u-pak-al					ma' ch'ab ma' ak'-al		
13	to come down-THM-3SA			3SE-flint 3SE-shield					NEG-creation-[NEG]-night		
14	1:1			1:2					1:4		
15	verb			subject							
16	verb			patient							
17	verbal phrase			nominal phrase							
18	"came down"			"his flint, his shield"					"no creation, no night"		
19	"it came down his flint, his shield, he who is without creation, without night"										
20	"defeated was the army of the captive"										

Fig. 14.1. Visualization of the important steps in the workflow of epigraphic analysis, drawings by Christian Prager (2015).

14.1.1. Conceptual questions

When developing and technically implementing the digital epigraphic research environment, the major focus is on pursuing the following objectives:

1. Creation of a markup schema to represent these steps (which in practice are interconnected and may also occur parallel to each other in the form of alternative interpretations);
2. Defining of the precise requirements that a data structure must fulfill in order to sufficiently represent complex i.e. non-alphabetic textual data for epigraphic analysis and research.

14.2. Prerequisites for making a non-alphabetic writing system digitally accessible

In order to make syllabic or logo-syllabic writing systems accessible for corpus linguistics, for instance to carry out a corpus analysis as part of the basic lexicography required to compile a dictionary of Classic Mayan, the phonology, morphology, syntax, semantics, and pragmatics of the language in question must be captured, marked up, and saved in the corpus. As such, the basic characteristics that distinguish the graphematic

lexicon of syllabic and logo-syllabic scripts from those of alphabetic writing systems must be taken into account. Thus, methodological and technical prerequisites must be fulfilled in order to be able to digitally represent and analyze a non-alphabetic script, such as Classic Mayan, for the purposes of epigraphic analysis.

14.2.1. Registering object data

When designing the digital epigraphic work environment, one must also account for the current state of decipherment of the script and language that are to be documented. Thus, Classic Mayan presents great challenges, as at least one third of the known graphemes have not yet been completely deciphered. Even when a lexeme can be read, its etymology or semantic domain is often unknown, and in some cases not even its lexical class can be determined, which causes difficulties when attempting to read and understand a text. In order to decode an unknown sign, word, or sign sequence in its respective context of use, for instance, semantic fields are generally investigated by means of applied substitution. From a structuralist perspective, this method investigates the paradigmatic relation between two linguistic entities as indicated by whether or not they appear in the same context. This step yields results when sufficient quantities of data can be digitally compiled, marked up and investigated using corpus linguistics.

The VRE currently being planned is the laboratory for this process, where new decipherments will be obtained and existing readings will be tested against primary source material while also considering all of the contextual data. The context of the contents that is referenced during decipherment may refer to the text carrier itself. For this reason, registering data on the object on which the text is written is just as important as documenting the text itself.

The decoding of Ugaritic cuneiform at the beginning of the twentieth century provides an example of a successful decipherment that resulted from studying the relationship between object and textual data. Scholars suspected that the word “axe” occurred in repeating, brief sign sequences on inscribed bronze axes. Building on this hypothesis, they determined the phonetic values of individual signs using data on the language and completely deciphered the script (Day 2002). Such investigations enable initial association of signs with both content and function in the relevant language so that the signs may later be gram-

matically, morphologically, semantically, etc., defined and eventually deciphered. This work also facilitates more precise typological classification of a writing system (Gronemeyer 2015).

Like the relationship between writing and the object, the relationship between text and image must also be coded. Associated depictions can potentially illuminate the contents of an inscription. Working from this analytical basis, a writing system can be registered in its respective contexts of use in order to derive semantic meanings and, ideally, linguistic decipherments. As such, contextual information concerning the text carrier as an object must be recorded, because its physical features may influence the text's contents and arrangement. In this respect, the size of the object, for example, may be just as influential as the material itself.

For example, the text carrier may influence scribal economy. The recent discovery of an inscription on very hard jade in Nim Li Punit (Belize) illustrates how the use of this particular writing surface can lead to a reduced spelling, meaning that the scribe omitted final syllables or used very simple sign variants that were almost geometric.

The form of a text carrier and its function also influence the text. Different lexemes may thus be determined for describing round or quadrilateral altars, with the result that conclusions may be drawn concerning the semantics and possible linguistic decipherments. Physical description of the text carrier as an object and its linguistic, historical, and cultural contextualization are indispensable for deciphering a previously unknown text or analyzing its arrangement. These metadata must be consulted in order to grammatologically and linguistically analyze the text's contents. Modelling an object biography and culturally contextualizing the texts must be initially addressed with information technology in order to facilitate epigraphic analysis of the text's contents in its extra-linguistic, cultural context. Thus, it is necessary to create an object metadata schema that captures such information concerning the language's cultural area and makes it available for use.

For these purposes, an epigraphic object metadata schema that compiles and represents extra-linguistic information about the inscriptions in an ontological structure must be created, in addition to an XML-based analytical metadata schema for linguistics.

The CIDOC-CRM standard for documenting cultural heritage offers a broad foundation that can be supplemented with other

standards, such as Dublin Core or the SKOS vocabulary. In addition to taking advantage of the comprehensive understandability of CIDOC-CRM, the scheme also requires to reuse data and to connect them with data from other research projects in the spirit of Linked Open Data (Diederichs 2015; Prager 2015).

14.2.2. Making linguistic data accessible

An interdisciplinary and widely accepted standard that takes into account both generalized and special characteristics of writing systems beyond those of a single script must be developed in order to be able to satisfactorily annotate records from various writing traditions that use syllabic signs and logograms to represent language.

In most cases, the scripts annotated and encoded in TEI and also those prepared for epigraphic analysis in EpiDoc are alphabetic, and most of them are linear and arranged in rows. Non-alphabetic writing systems, such as Egyptian, various cuneiform systems, or Hieroglyphic Luwian, arrange their graphemes principally in groups or blocks, and only secondarily in rows or columns, with a high rate of metathesis (Lacau 1903) in order to optimize the use of space. Furthermore, many of these writing systems demonstrate a high degree of allography, which is no longer apparent in a pure transliteration. Various desiderata become apparent when an epigraphic project attempts to use the standards provided by TEI and EpiDoc to edit texts composed in these scripts.

Annotating texts composed in such writing systems that use only alphabetic transcription is insufficient; instead, the original spelling of a lemma should be represented. This approach facilitates studies of paleography or sign usage across time and space. In addition, it can reveal preferred sign arrangements within a block or preferences for particular signs that correspond to the material used or to the subject of the text, for instance. Thus, when creating a standard for non-alphabetic scripts, object metadata and annotations for orthographic and linguistic analysis must be taken into account, and the creative process should also promote discussion of terminological and typographic conventions for these annotations (Sachse and Dürr 2016). The goal is to create a generic markup model that can be implemented regardless of the particular script or language in question.

The inscriptions of the Maya hieroglyphic writing system alone feature attestations of various vernacular languages (Lacadena and Wichmann 2002), a phenomenon which raises several preliminary questions: what exactly are the demands, and what are the goals of an epigraphic and linguistic annotation? Where do possibilities and limitations exist for annotating these writing systems using established standards, such as TEI or EpiDoc? Existing standards have to be modified to some extent and possibly new standards have to be created, to overcome these limitations.

14.2.2.1. Representing the primary text source

When defining a text, acknowledging an interpretation of the source text itself is unavoidable. The meaning of the “original text” that resides in the “source text” is an important point of discussion in epigraphy. Thus, the annotation does not contain a guiding text in the conventional sense, because it is merely reconstructed using all given text information and one’s own conclusions and interpretations, and then represented with the aid of alphabetic transliteration and transcription. In this respect, one should always leave open the possibility of separating primary data (in the case of hieroglyphic texts, photographs of the text carriers, for example) and secondary data (e.g. drawings or interpretive annotations) (Stührenberg 2012). This strategy can be implemented using stand-off annotations for data. Similarly, such an annotation should permit multiple descriptions, i.e. alternative statements concerning the data (Stührenberg 2012). For example, sign sequences in Classic Mayan can be variously analyzed depending in a particular vernacular context (Gronemeyer 2014). Only in this manner can various ways of thinking be appropriately recorded and relayed to the scientific community for discussion.

A key objective of the XML-based markup of hieroglyphic, cuneiform, or linear texts should thus be to represent the original spelling and arrangement of the signs in their respective contexts. A linear transcription alone cannot represent the original text or primary source in its entirety, as many details remain undocumented. For example, signs in Maya hieroglyphic texts are not arranged according to their literal reading order, but instead in spatially distinct, square or rectangular units (so-called “blocks”), each of which in most cases corresponds to a word or morpheme sequence. A detailed markup of the original text is therefore of great importance, particularly for partially deciphered

and undeciphered graphemes and writing systems in general. In such cases, an alphanumerical or numerical nomenclature is often used to refer to the signs in order to carry out a corpus linguistic analysis of the texts. The arrangement of each block, as well as the text and its position on a text carrier, should be documented as well. To fully understand a writing system – i.e. the language and information expressed in it – a detailed representation of the primary source and its context must be a key objective in the digital markup of documents. When studying complex writing systems (respectively non-alphabetic writing systems), digital documentation of the original spelling using annotation standards like TEI is a basic prerequisite for conducting a detailed graphemic and graphetic analysis of the relevant script and for providing a basis for a linguistic and corpus linguistic investigation. This need represents a significant desideratum in epigraphic research, and it also constitutes a core pillar of computational linguistics (McEnery and Wilson 2001).

In an interdisciplinary effort, epigraphers and experts of digital markup-languages alike need to discuss methods for investigating syllabic and logo-syllabic writing systems using the XML-based standards like TEI or EpiDoc. When doing so, they should discuss the following points of emphasis, among others:

14.2.2.2. Graphetics

Graphetics concerns the formal structure of linguistic units and the structure of texts (Crystal 1997). Classic Maya hieroglyphic inscriptions, for instance, may display a very high degree of variation in writing styles for arranging texts, a phenomenon which is deserving of investigation. In this manner, various systems of notation are studied in their individual, social, and typographic aspects, for instance. Similarly, in paleography, script decipherment is examined from graphetic perspective. The arrangement of texts, for instance, different grapheme or block sizes of text fields, or different styles of hieroglyphic writing (equivalent to alphabetic font style and size) are analyzed.

Research addresses the degree to which attested variation in the composition of hieroglyphic texts may be ascribed to specific scribal schools or to regional forms of expression, rather than to differences in the meaning of the signs' contents.

14.2.2.3. Graphemics

When analyzing the contents of these texts, it is important to investigate meaningful variation underlying the manner in which the writing is arranged and to thereby identify distinct units. Graphemics is devoted to this area of research, and consequently to exploring the meaningfully distinct features of signs.

Allography of graphemes

A very important example of a subject of graphemic research is provided by allography, a phenomenon which describes a 1:n relationship between a grapheme and its various graph representations (Crystal 1997).

The process of establishing a suitable annotating-tool thus must be oriented toward a series of questions: What is the significance of allography for the respective grapheme inventories? Are allographs annotated in the transliteration or transcription of the respective writing systems?

For example, in Maya writing, there are over twenty different allographs for the vowel sign <u>, which is used e.g. as the 3rd person ergative pronoun u- "he, she, it". Additional research questions may comprise 1) the reading order of the signs in context; 2) the existence of word separators or other graphic aids used to differentiate between meaningful units or words; 3) multiple possibilities for establishing a reading order of signs within a block that are equally meaningful (with alternatives documented or marked by epigraphers); 4) the reading order of meaningful units within the texts; 5) variations or violations of these orders (and if so, with a markup as well); 6) cases where images appear with the text and their possible relation; 7) integration of texts into the images or vice versa.

Graphotactic strategies of a writing system

For analysis, it is important to convey the graphemics of a writing system, which are lost in an alphabetic transliteration. Graphemics concerns sign function, or graphotactic strategies for constituting a meaningful unit or word. Representing the reading order of signs in the text and in each block is similarly imperative. To graphotactically link individual signs in a hieroglyphic block in Maya writing, various representational rules may be used (see Fig. 2), e.g. affixation, infixation, conflation, or superimposition (Zender 1999).

Additional typical, functional traits of logo-syllabic writing systems include underspellings, diacritics, phonemic complements and indicators, semantic classifiers and determinatives, and sign class convergences. However, these characteristics may have very diverse grapho-tactical manifestations in different writing systems, which nonetheless need to be standardized and communicated in a digital structure in order that researchers may thus take note of various interpretations under discussion.

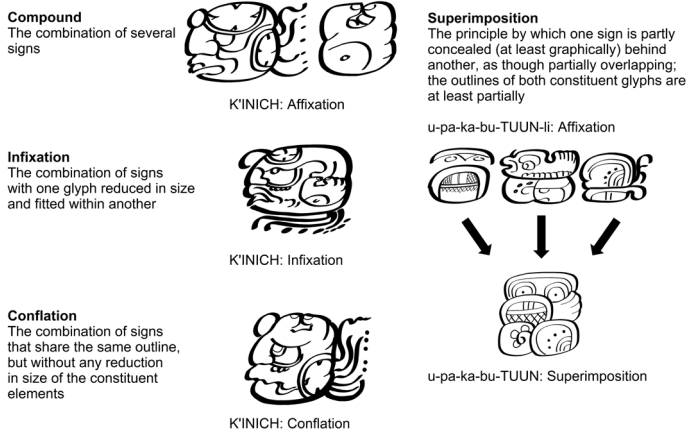


Fig. 14.2. Some possible sign combinations in Classical Mayan hieroglyphs after Zender (1999), drawings by Christian Prager (2015).

14.2.2.4. Signary: sign classification and sign catalogs

The graphemic lexicon provides a central reference for marking up texts composed in non-alphabetic writing systems. In the case of the Maya script, signs whose phonemic reading is unknown or unconfirmed are denoted using reference systems from various classification catalogues. Use of question marks for indicating unclear readings, for example, should be avoided, because such signs impair machine readability and sometimes represent control characters or part of an escape sequence. In this context, it should be determined what significance sign classifications and nomenclatures have in their respective branches of epigraphic research. Also their role in transliterating and transcribing texts in the respective writing system, as well as in sign classification is to be examined.

Sign inventories constitute a central authority in an epigraphic VRE, and they need to be modified or supplemented in accordance with the ever-changing state of research as soon as newly discovered material is entered and signs are identified that had not previously been isolated or identified as discrete graphemes.

14.2.2.5. The state of decipherment and the readability of texts

In the case of unreadable or only partially deciphered writing systems, the question in the forefront is the state of decipherment of the relevant writing system and the handling of undeciphered signs or text passages. As archaeological artifacts, texts are also subject to processes of decay that inhibit visible legibility. It has several implications: 1) the treatment of undeciphered signs or text passages; 2) the indication of physical gaps (as texts are also subject to processes of decay that inhibit visible legibility) and their marking in the transliteration and transcription of the text; 3) the treatment of hypothetical, yet plausible, readings of singular signs or passages; 4) the handling of alternative readings of the same passage; and 5) a review to what extent the EpiDoc Metadata standard, into which the Leiden Conventions have also been incorporated, is able to capture these features of readability and alternative interpretations of inscriptions.

These and other characteristics of non-alphabetic scripts, which to a certain extent reach beyond the epigraphic problematic of alphabetic writing systems, illuminate a desideratum here and the need to formulate ideas that contribute to the design and modelling of appropriate XML-based metadata schemas.

14.3. Open science strategy

A Digital Humanities project working in the spirit of Open Science fundamentally intends the work environment that it develops to be reused. As such, a VRE for non-alphabetic writing systems such as that of the Classic Maya should be developed as generically as possible, in order that structural interoperability may ensure that they can be reused by similar epigraphic projects.

In this respect, any metadata schema that is created should be widely intelligible and ensure the data's syntactic and semantic interoperability by using common, given, XML-based annotation and metadata

standards, like TEI and EpiDoc, and controlled vocabularies, such as SKOS, that are already established in the Digital Humanities (Diederichs 2015).

Another objective of database projects in the Digital Humanities should be to ensure open access to their data and metadata, and thus to maximize usage of their research by maximizing user access to it. As all current and future research and innovation stands on the shoulder of giants, an efficient system for broadly disseminating and allowing uninhibited access to the project's research (raw data and metadata), as well as for guaranteeing the contents' productive reuse, must be ensured through the use of free licenses (DFG 2013).

As such, all the contents of a database and data infrastructure that are being planned will be made available under so-called "Open" licenses (Open Access, Open Source, Open Methodology, Open Data, etc.), with the goal of pursuing a comprehensive Open Science strategy (Diederichs 2016).

All methods and the process of developing tools must thus be carefully documented to maximize reuse of all schemas and software. In particular, all research information and results should be made available not only to the scientific community, but also to the general public, to contribute to the digital safeguard and dissemination of humanity's cultural heritage.

14.4. Summary

The creation of digital infrastructures and metadata schemas that facilitate recording and further analysis of the appropriate annotation for non-alphabetic writing systems, including their respective cultural contexts, is a desideratum in the Digital Humanities that needs to be addressed.

The recommendations articulated here for an object database, as well as for a structure for recording epigraphic and linguistic data, can make a significant contribution to understanding how traditional epigraphy can be transferred to digital form. In working towards this goal, methods and tools from the field of Digital Humanities will be suitably adapted to the needs of epigraphy, and new methods and structures from the digital world will find their way into digital epigraphy.

References

- DFG. *Deutsche Forschungsgemeinschaft (DFG), Praxisregeln 'Digitalisierung'*. Bonn: Deutsche Forschungsgemeinschaft, 2016. http://www.dfg.de/formulare/12_151/12_151_de.pdf.
- CRYSTAL, David. 1997. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.
- DAY, Peggy L. 2002. "Dies Diem Docet: The Decipherment of Ugaritic." *Studi Epigrafici e Linguistici*: 19–37. http://www.proyectos.cchs.csic.es/SEL/sites/default/files/06day_2a4aeb99.pdf.
- DIEDERICH, Katja. 2015. "The 'Open Science' Strategy of the Project 'Text Database and Dictionary of Classic Mayan'". *Textdatenbank und Wörterbuch des Klassischen Maya, Working Paper 1*. http://mayawoerterbuch.de/wp-content/uploads/2015/06/twkm_paper_001en.pdf.
- GRONEMEYER, Sven. 2014. "E pluribus unum: Embracing Vernacular Influences in a Classic Mayan Scribal Tradition." In *A Celebration of the Life and Work of Pierre Robert Colas*, edited by Christophe Helmke and Frauke Sachse, 147–162. Markt Schwaben: Anton Saurwein.
- . 2012. "Class Struggle: Towards a Better Understanding of Maya Writing Using Comparative Graphematics." In *Proceedings of the 17th European Maya Conference 2012*, edited by Harri Kettunen and Christophe Helmke. Markt Schwaben: Anton Saurwein.
- LACADENA, ALFONSO AND SØREN WICHMANN. 2002. "The Distribution of Lowland Maya Languages in the Classic Period." In *La Organización Social entre los Mayas: Memoria de la Tercera Mesa Redonda de Palenque*, edited by Vera Tiesler Blos, Merle Greene Robertson, and Rafael Cobos, 275–314. México D.F.: Instituto Nacional de Antropología e Historia/CONACULTA and Universidad Autónoma de Yucatán.
- LACAU, Pierre. 1903. "Métathèses Apparentes en Égyptien." *Recueil de Travaux Relatifs à la Philologie et à l'Archéologie Égyptiennes et Assyriennes* 25: 139–161.
- MCENERY, TONY AND ANDREW WILSON. 2001. *Corpus Linguistics: An Introduction*. 2nd ed. Edinburgh: Edinburgh University Press.
- NEUROTH, HEIKE, FELIX LOHMEIER, AND KATHLEEN MARIE SMITH. 2001. "TextGrid - Virtual Research Environment for the Humanities." *International Journal of Digital Curation* 6: 222–231.
- PRAGER, CHRISTIAN. 2015. "Das Textdatenbank- und Wörterbuchprojekt des Klassischen Maya: Möglichkeiten und Herausforderungen digitaler Epigraphik." In *TextGrid: Von der Community - für die Community*, edited by Heike Neuroth, Andrea Rapp, and Sybille Söring, 105–124. Göttingen: Werner Hülsbusch.
- SACHSE, FRAUKE AND MICHAEL DÜRR. 2016. "Morphological Glossing of Mayan Languages under XML: Preliminary Results." In *Textdatenbank und Wörterbuch des Klassischen Maya, Working Paper 4*. http://mayawoerterbuch.de/wp-content/uploads/2016/01/twkm_paper_004.pdf.

- STÜHRENBERG, MAIK. 2012. "Auszeichnungssprachen für linguistische Korpora: Theoretische Grundlagen, De-facto-Standards, Normen". PhD diss., Universität Bielefeld. <http://pub.uni-bielefeld.de/download/2492772/2492773>.
- ZENDER, MARC UWE. 1999. "Diacritical Marks and Underspelling in the Classic Maya Script: Implications for Decipherment." MA thesis. University of Calgary, 1999. http://www.collectionscanada.gc.ca/obj/s4/f2/dsk1/tape9/PQDD_0020/MQ47975.pdf.

15. Integration of Multimedia Collections and Tools for Interaction with Digital Content. The case study of the Archaia Kypriaki Grammateia Digital Corpus

Uros Damnjanovic, Valentina Vassallo**, Sorin Hermon****

Abstract

Supporting a discovery, use, and navigation of digital collections is a fundamental part of providing access and encouraging inquiry, interpretation, and knowledge. In this paper we present our efforts to store and explore multimedia collections of archaeological data. Particularly, the case study of the Archaia Kypriaki Grammateia epigraphic collection is presented. Our work can be seen twofold. One aspect of our work is to provide a place where the data coming from various sources can be stored and accessed. Another aspect is to provide users with means to explore this data. We argue that currently digital libraries are constrained by their webpage-based paradigm, thus not providing the means for utilizing the full potential of the heritage data.

Keywords: Digital libraries, Data repository, Epigraphy, User interfaces, Data visualization, Interaction

15.1. The world going digital

Nowadays digital information format found its ways to all the spheres of our lives. Cultural heritage is no exception to this. Gigabytes of data are being created on an every-day basis. For some time now, digital libraries are used to store and provide access to digital heritage data. However, existing digital libraries often fail to provide the ability to iteratively explore items, compare data trends, and engender the wisdom that comes from exploring data in new ways. Current digital libraries are limited by their inadequate webpage-based paradigm,

* The Cyprus Institute - STARC.

** The Cyprus Institute - STARC. Corresponding author. Email: v.vassallo@cyi.ac.cy.

*** The Cyprus Institute - STARC.

and it is easy for even the most experienced scholar to get lost (Bergstrom and Atkinson 2009). To overcome this issue, a new generation of digital libraries should emerge that will serve as an interactive knowledge environment in which knowledge is created by interacting with the data. Digital collections often lack features for deeper quantitative and qualitative analysis, and even very useful functions, such as the ability to annotate or bookmark content, are often not supported. After digitisation, these collections are typically monolithic, difficult to navigate, and can contain text which is of variable quality in terms of language, spelling, punctuation and consistency of terminology.

According to Fast and Sedig 2010 digital libraries are seen as a store of epistemic potential with interaction having the role to reveal much of the hidden complexity. In Buchel and Sedig 2014 the authors demonstrated the role that a set of interactions can play in supporting users' understanding of spaces and their formation of cognitive maps when working with map-based visualisations. According to Algee, Bailey, and Trevor 2012 interactive visualizations empower users to discover meaning and patterns within digital collections using dynamic, interactive displays. One example of utilizing the interactive graphics in a real world digital library is presented in Hienert et al. 2012. Authors developed interactive visualization tools that support exploration of search queries and search results and help users in formulating new queries. Another example, the application called DaisyViz shown in Ren et al. 2010, tries to help users acquire better insights of the data by enabling users to rapidly develop domain-specific information visualizations without traditional programming. In Nasar, Mohd and Mohamad Ali 2011 authors proposed a conceptual framework that uses interactive visualizations for managing personal collections of images and videos with focus on data re-finding and improved filtering. INVISQUE (Wong et al. 2011) is a novel system designed for interactive information exploration. Instead of a conventional list-style arrangement, in INVISQUE information is represented by a two-dimensional spatial canvas, with each dimension representing user-defined semantics. Search results are presented as index cards, ordered in both dimensions. Intuitive interactions are used to perform tasks such as keyword searching, results browsing, categorizing, and linking to online resources such as Google and Twitter. An approach for overview-first exploration of data collections based on user-selected metadata properties is presented in Bernard et al. 2012.

In a 2D layout representing entities of the selected property are laid out based on their similarity with respect to the underlying data content. The display is enhanced by compact summarizations of underlying data elements, and forms the basis for exploratory navigation of users in the data space. In Nan et al. 2012 authors propose a novel visual design termed “Whisper” to fulfil the need for tracing information diffusion processes in social media, in a real time manner. BirdVis (Ferreira et al. 2011) leverages visualization techniques and uses them in a novel way to better assist users in the exploration of interdependencies among model parameters. Furthermore, the system allows for comparative visualization through coordinated views, providing an intuitive interface to identify relevant correlations and patterns.

15.2. “A place where the past meets the future”: the STARC repository

STARC repository was initially developed to store the data coming from the daily work of the research group and as content aggregator for various digital libraries projects: data from archaeological excavations, museums artefacts, epigraphic corpora, etc. Data stored in the repository ranges from high resolution images, 3D models, texts, maps, videos and audio resulting from various data acquisition procedures such as photogrammetry, laser scanning, 3D modeling, photography, sketching, drawing and so on. Every item in the repository is described using a metadata schema that depends on the type of data and according to the needs of the project. For most of the items, we are using our own metadata schema (STARC metadata schemas) (Ronzino, Hermon and Niccolucci 2012; Vassallo et al. 2013), that was developed to thoroughly describe various aspects of data creation process. Once data is uploaded and available, it can also be used in different systems. For example, the repository serves as source of data for aggregation procedures by which our data feed the Europeana portal (<http://www.europeana.eu/portal/>) through different digital libraries projects (e.g. Athena, CARARE, Linked Heritage, AthenaPlus, EAGLE). In order to facilitate the migration of data from heterogeneous sources, the repository provides ingestion capabilities, that enable easy transfer of data from any available data end-point. Another aspect of the back-end is the user management. It deals with providing different access levels for different users’ groups.

It also manages personal user space, where users can add their own information to the repository and have access to it when exploring the data. Data annotations and personal collections are some of the examples of how users can add information to the repository and use it for data exploration.

15.2.1. Tools for accessing data

We wanted to provide users with a number of tools that they can use while exploring the repository. In traditional web based system, navigating from one tool to another usually means that we click on a link and then move to another page where the tool is available. Once the user is on a page s/he can start interacting with the tool by adding or modifying data or by performing any action relevant to the selected functionality. The process of exploring the repository starts by running the initial query (Fig. 1).

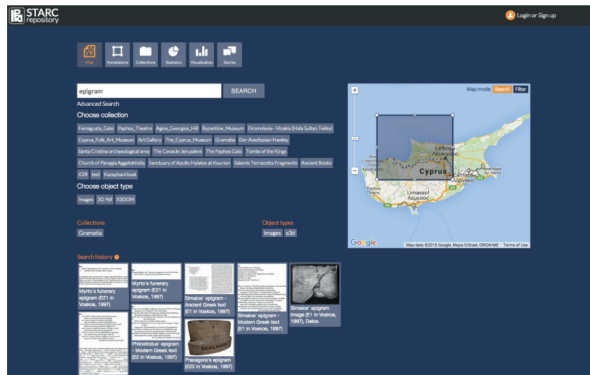


Fig. 15.1. The exploration process starts by running the search. Once the results are available the user can add or remove exploration tools. By default map tool is present in the view. The map tool is used to mark the geo-referenced data on the map, and also to perform geo-search by defining the region of interest.

Once the results are returned and displayed, the exploration process can start. By default the map functionality is available in the view. When a new tool is added to the screen, it automatically gets an access to all the data available in the current search. It also gets notified on any other tools that might be active in the view, so the necessary communication channels are established. Once the results are ready, there are a number of filters available to filter out the data, either by data type, collection. Search form can be used to filter out the data by typing

the text used to filter the data. As the user types, the data is automatically updated to show only those items that are related to the text in the search field.

The user can then run a new search repeatedly, and the search history functionality will store all those searches for later reference. The map tool (Fig. 1) is used to show information about the geo-referenced data. It provides two main functions. One is that it accompanies the search tool. Whenever the new search is ran, and the results become available, the map tool updates itself with new data, showing the geographical distribution of the search results. The user can then use the map to filter out the data by clicking on the markers on the map. Also the map can also be used as a search tool. By selecting the search map mode, the rectangle mask appears on the map, that is used to define the region of interest on the map. This region is then used as an additional parameter for the search.

Annotation tool is used to attach the information to the objects in the repository and share it with other users, or use it to facilitate more efficient search. The tool is used by selecting object from search results, then either by selecting a region in the image files (Fig. 2), or by selecting a point for 3D models, and by adding textual description of the selected part.

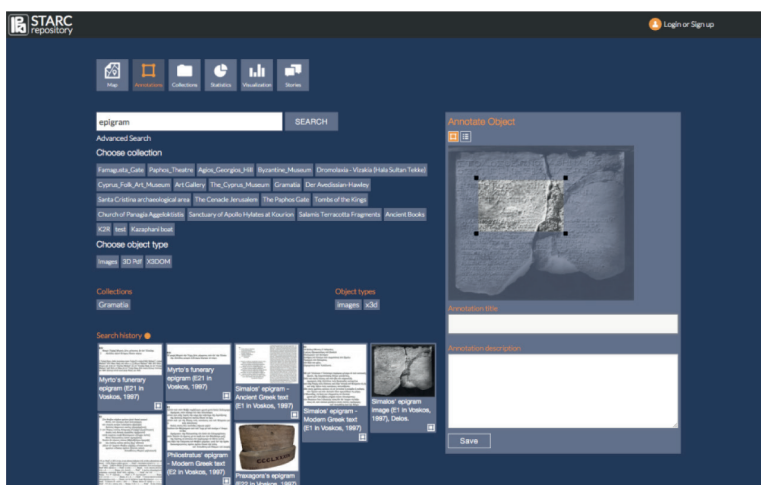


Fig. 15.2. The annotation tool. The user first selects the object in the search results, then selects the region of the object and then adds the annotation. The annotation can be accessed when the annotated object is accessed, and by specifying in the search that the search should also include annotations.

Once the search results are available, and the user selects the collection tool, the user can start adding and removing the objects from the collection. Finally the collection is described and saved. Same as for the annotations, the collection information is saved in the user's personal space, and can be also used in search, by specifying to the search that it should include the personal collections in the search. This means that the search will look not only for objects descriptions and metadata, but will also look inside collection descriptions to try and match the query.

In order to help in exploring the datasets, there are statistics and visualization tools. Statistics tool shows the basic properties of the repository. A visualisation tool is used to accompany the search tool (Fig. 3). Story creation tool is another tool that supports collaborative user generated content. It provides users with an online document editor, where users can write their own ideas, notes, insights about the data. The document created in this way is stored in the repository and can be accessed in the personal workspace, and shared with the others.

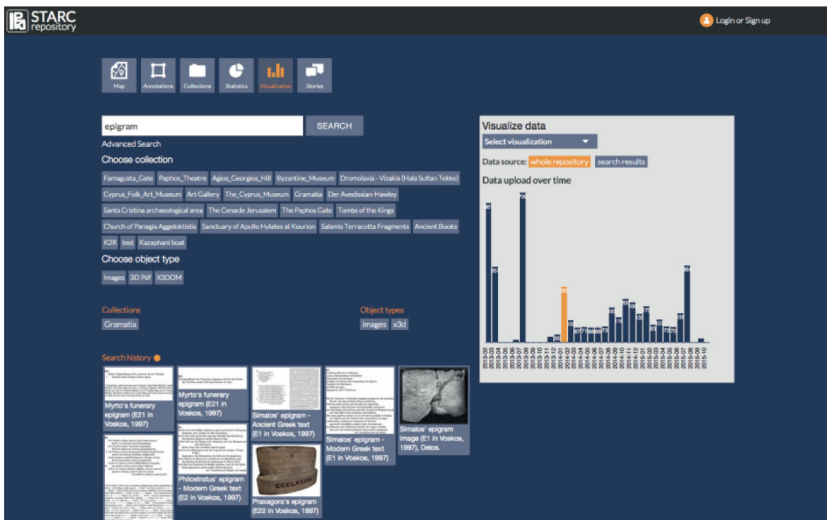


Fig. 15.3. Visualization tool. Visualization tool is used to accompany the search tool by showing live visualizations as the search results change. When the search is ran and the search results are available the visualization tools automatically update itself and shows the visualisation related to the current search results.

15.3. Archaia Kypriaki Grammateia, the STARC Repository and the EAGLE infrastructure

One of the most interesting collection stored in the repository is the Archaia Kypriaki Grammateia Digital Corpus (Pitzalis et al. 2012). The collection is composed of Ancient Greek and Latin epigraphic texts produced within a time span of circa 13 centuries (from the 7th century BC to the 6th century AD). The texts are attributed to Cypriot authors or were produced in Cyprus. The corpus consists mainly of funerary or dedicatory epigrams published with their translation in Modern Greek, critical apparatus and philological comments (Voskos 1997).

The peculiarity of the digital collection consists in the fact that, beyond the epigraphic texts, it is composed of a series of multimedia digital resources that describe the content in a multidisciplinary way: digital texts, images related to the epigraphic supports, 3D representation of the inscriptions, video, audio files, and so forth (Fig. 4).


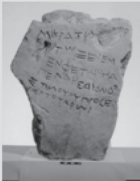


Object				
Format	Text	Image	3D	Audio

Fig. 15.4. Multimedia data constituting the AKG digital collection stored in the STARC repository.

Such a material needs to be described in the right way, in order to provide detailed information about all the elements that constitute the collection. For this reason a specific metadata schema has been developed: it covers the multidisciplinary research carried out on this material in a manner that users comprehend the complexity of such an approach through access to heterogeneous types of information (Vassallo et al. 2013).

The metadata schema for Ancient Cypriot inscriptions integrates multidisciplinary information regarding the objects and their multiple digital resources. It is the result of a research line developed within the group (Liuzzo, Rivero Ruiz and Vassallo 2014) and of the assessment and comparisons of schemas, models and ontologies already in use in the digital epigraphy field (e.g. TEI Epidoc, Dublin Core, CIDOC-CRM).

The metadata schema is organized in groups corresponding to different research areas clustered in wrappers and sub-wrappers, in order to fully describe all information that the 'asset inscription' contains. This schema, distinguished by its multidisciplinary structure, include all disciplines relevant for description and representation of epigraphies: archaeology (investigating for example the context of the finds), philology (analyzing the text, the writing style, the scripture), chemistry and geology (providing details on the material upon which inscriptions were carved), conservation (giving information about the state of the artefact), visualization and museology (about the museums and places of conservation).

Besides the inscriptions, since the Cypriot collection consists of their digital representations (pictures, 3D models of inscriptions, videos, etc.), the metadata schema takes into consideration their descriptive features, their digital provenance and other related information. For example, through the metadata it is possible to describe the digital provenance of the 3D model of an inscription and to give information about the acquisition phase (the technique, the tool used, the specification of the tool, the specification of the output, etc.) the post-processing (the operative info, the file specifications, etc.) and the digital output (data format, software), according to the digital resource obtained (Fig. 5).

The metadata schema for ancient Cypriot inscriptions aims at the following goals:

- to describe in detail the digital resources and its digital provenance,
- to provide related information about the context of the inscription,
- to enable harvesting to larger initiatives (e.g. Europeana, Ariadne).

15.3.1. From STARC Repository to EAGLE infrastructure

Beyond the functionalities developed for an enhanced users' access to the content, the STARC repository plays as aggregator for different digital libraries projects that in turn provide content to Europeana portal. The epigraphic content of the Archaia Kypriki Grammateia digital corpus is ingested through the EAGLE Project.

The “Europeana network of Ancient Greek and Latin Epigraphy”, is an EU funded project under the umbrella of the CIP-Best Practice Network and whose aim is to bring together some of the most prominent European institutions and cultural archives in the field of Classical Latin and Greek epigraphy. One of the aim is to provide Europeana with a comprehensive collection of unique and curated online edition of historical and archaeological sources.

To encode inscriptions from different epigraphic databases, EAGLE developed a metadata format that assessed the provider’s metadata structures and considered two sets of standards: TEI EpiDoc and CIDOC CRM. EpiDoc allows a full description of the text of inscriptions while CIDOC CRM enables a further full description object-oriented, reflecting at the same time the different souls of Epigraphy, the philological and the archaeological one. Content providers use different models according to the fact their databases are more oriented towards the text of an inscription or towards the archaeological object that bears the text. According to the epigraphers’ community needs, EAGLE consortium developed the data model on the base of the two sets of standards (Liuzzo, Rivero Ruiz and Vassallo 2014). Finally, the EAGLE data model is mapped to the Europeana Data Model (EDM), in order to be ingested into the Europeana portal.

The content provided by the EAGLE consortium, due to the transformation into the EAGLE metadata model, produce three different groups of metadata according to the digital objects they are connected to: artefacts, documental manifestation, visual representation. The data that are sent to Europeana belong to the category of the artefacts (the inscriptions represented as texts, called by Europeana “Cultural Heritage Object” - CHO) and of the visual representation (the images related to the inscriptions, called by Europeana “WebResource”). The distinction between CHO and WebResource has been introduced in Europeana as a result of the introduction of the new EDM schema to provide the users with a better navigation and data retrieval and to aggregate under the same umbrella all the resources available for a Cultural Heritage Object, avoiding data duplication in the portal.

Most of the consortium data, as well as the majority of the existent epigraphic databases worldwide, are compliant to Epidoc. This allows to perform a smooth mapping to the EAGLE data model, to have results that are aligned to the project aims (e.g. the possibility

to create dedicated groups of metadata associated with the category of the object) and to simplify the involvement of other databases in EAGLE through the enlargement of the consortium.

As previously mentioned, the *Archaia Kypriaki Grammateia* digital corpus data integrated in the STARC Repository are different from other epigraphic databases data. The study of the epigraphy is not text-centric. The text is one of the representations of the artefact: the archaeological object, the support, the 2D image or the 3D of the object are all resources that, even if connected, have their own identity and their own set of information. Moreover, the AKG is a collection of texts accompanied by their translations, notes and commentaries as published in its last edition by Voskos (Voskos 1997), therefore it is a digital form of this important editorial printed work, to which are connected all a series of digital resources that help to enrich the corpus.

Such a material and the metadata schema used for the description of this content converges in a difficulty to map our metadata schema to EAGLE data model. This implies a further effort to integrate the content into the EAGLE infrastructure for a compliant visualization with the other content and retrieval in the EAGLE and Europeana portals. Concerning the issues connected to the mapping and integration of the AKG data into EAGLE portal, possible solutions have been investigated and are currently under tests:

- the creation and use of relations that are able to aggregate all the resources under a specific resource identified as reference item.
- the creation of rules for the development of a script that will be able to map to the EAGLE data model every time according to the type of object we are dealing with. This implies also an editing of the metadata, creating single items that have as head the information about the inscription (e.g. the Ancient Greek text) and to which are attached all the other metadata sets concerning the related objects (e.g. the Modern Greek text, the image support of the inscription, the commentary, the 3D of the support-inscription, etc.).

Even if the first solution would be much faster from a technical point of view, at the moment is under preparation a test for the second solution. In fact the latter, even if it is time-consuming in terms of elaboration, seems to be the best solution for avoiding data duplication and for having more cohesive data.

15.4. Conclusions

We presented in this paper a digital data repository and showed numbers of innovative tools used to access and explore its collections. We argued that by interacting with data in an innovative ways can help users better understand data and stimulate knowledge creation. One of the most interesting collections stored in the repository is the Archaia Kypriaki Grammateia Digital Corpus composed of Ancient Greek and Latin epigraphic texts. With the set of proposed functionalities we also enabled exploration of the epigraphic text in a way that supports sense making, understanding and collaboration. We also showed how the collection of epigraphic texts can be used within other projects by mapping our data to the appropriate data formats. Next step is to set up an evaluation framework that will try to measure and evaluate the contribution of each tool to the users' daily tasks. We want to measure the benefits of using such a tools by performing various evaluation tasks and measuring the user's performance.

References

- ALGEE, LAUREN, JEFFERSON BAILEY, and TREVOR OWENS. 2012. "Viewshare and the Kress Collection: Creating, sharing, and rapidly prototyping visual interfaces to cultural heritage collection data." *D-Lib Magazine* 18(11): 3. <http://www.dlib.org/dlib/november12/algee/11algee.html>.
- BERGSTROM, PETER and DARREN C. ATKINSON. 2009. "Augmenting the exploration of digital libraries with web-based visualizations." In *Fourth International Conference on Digital Information Management*, volume 1, 1–7. IEEE. doi:10.1109/ICDIM.2009.5356798. <http://ieeexplore.ieee.org/document/5356798/>.
- BERNARD, JÜRGEN, TOBIAS RUPPERT, MAXIMILIAN SCHERER, JÖRN KOHLHAMMER, and TOBIAS SCHRECK. 2012. "Content-based layouts for exploratory meta-data search in scientific research data." In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 139–148. ACM. <http://dl.acm.org/citation.cfm?id=2232844>.
- BUCHER, OLHA and KAMRAN SEDIG. 2014. "Making sense of document collections with map-based visualisations: the role of interaction with representations." *Information Research* 19(3). <http://www.informationr.net/ir/19-3/paper631.html#.WOJPr2clHIU>.
- FAST, KARL V. and KAMRAN SEDIG. 2010. "Interaction and the epistemic potential of digital libraries." *International Journal on Digital Libraries* 11(3): 169–207. <http://link.springer.com/article/10.1007/s00799-011-0066-8>.

- FERREIRA, NIVAN, LAURO LINS, DANIEL FINK, STEVE KELLING, CHRIS WOOD, JULIANA FREIRE, and CLAUDIO SILVA. 2011. "Birdvis: Visualizing and understanding bird populations". *Visualization and Computer Graphics, IEEE Transactions on* 17(12): 2374–2383. <http://dl.acm.org/citation.cfm?id=2068640>.
- HIENERT, DANIEL, FRANK SAWITZKI, PHILIPP SCHAEER, and PHILIPP MAYR. 2012. "Integrating interactive visualizations in the search process of digital libraries and IR systems." In *Advances in Information Retrieval Lecture Notes in Computer Sciences*, 447–450. Berlin: Springer. http://link.springer.com/chapter/10.1007/978-3-642-28997-2_38.
- LIUZZO, PIETRO MARIA, EYDEL RIVERO RUIZ, and VALENTINA VASSALLO. 2014. "Networking EAGLE with CIDOC and TEI." In *CIDOC 2014: Access and Understanding – Networking in the Digital Era*, Dresden. http://www.cidoc2014.de/images/sampleddata/cidoc/papers/J-2_Vassallo_Ruiz_Liuzzo_paper.pdf.
- NAN, CAO, LIN YU-RU, SUN XIAOHUA, DAVID LAZER, LIU SHIXIA, and QU HUAMIN. 2012/ "Whisper: Tracing the spatiotemporal process of information diffusion in real time." *Visualization and Computer Graphics, IEEE Transactions on* 18 (12): 2649–2658. <http://ieeexplore.ieee.org/document/6327271/>.
- NASAR, MOHAMMAS AL, MASNIZAH MOHD, and NAZLENA MOHAMAD ALI.. 2011. "A conceptual framework for an interactive personal information management system." In *User Science and Engineering (i-USER), 2011 International Conference on*, 100-105. IEEE <http://ieeexplore.ieee.org/document/6150545/>.
- PITZALIS, DENIS, ELINA CHRISTOPHOROU, NIKI KYRIAKOU, ARISTOULA GEORGIOUDOU, and FRANCO NICCOLUCCI. 2012. "Building scholar e-communities using a semantically aware framework: Archaia Kypriaki Grammateia Digital Corpus". In *VAST12: The 13th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, 89–95. <http://diglib.eg.org/handle/10.2312/VAST.VAST12.089-095>.
- REN, LEI, FENG TIAN, LIN ZHANG, and GUOZHONG DAI. 2010. "DaisyViz: A Model-based User Interfaces Toolkit for Development of Interactive Information Visualization." In *Visual Information Communication*, 209–229. Berlin: Springer. http://link.springer.com/chapter/10.1007/978-1-4419-0312-9_14.
- RONZINO, PAOLA, SORION HERMON, AND FRANCO NICCOLUCCI. 2012. "A metadata schema for cultural heritage documentation." *Electronic Imaging & the Visual Arts*: 36–41.
- VASSALLO, VALENTINA, ELINA CHRISTOPHOROU, SORIN HERMON, AND FRANCO NICCOLUCCI. 2013. "Revealing cross-disciplinary information through formal knowledge representation—A proposed Metadata for ancient Cypriot inscriptions." In *Conference Proceeding of Digital Heritage Heritage International Congress (Digital Heritage)*, edited by A. Addison, L. De Luca, G. Guidi, and S. Pescarin. IEEE. <http://ieeexplore.ieee.org/document/6744732/>.
- VOSKOS, ANDREAS 1997. *Αρχαία Κυπριακή Γραμματεία*. Nicosia: A.G. Leventis Foundation.

- WONG, WILLIAM, RAYMOND CHEN, NEESHA KODAGODA, CHRIS ROONEY, AND KAI XU.
2011. "INVISQUE: intuitive information exploration through interactive
visualization." In *CHI'11 Extended Abstracts on Human Factors in Computing
Systems*, 311-316. ACM. <http://dl.acm.org/citation.cfm?id=1979720>.