The EAGLE Europeana network of Ancient Greek and Latin Epigraphy: a technical perspective

Andrea Mannocci, Vittore Casarosa, Paolo Manghi, Franco Zoppi

Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Via Moruzzi 1, 56124 Pisa, Italy

name.surname@isti.cnr.it

Abstract. The project EAGLE (Europeana network of Ancient Greek and Latin Epigraphy, a Best Practice Network partially funded by the European Commission) aims at aggregating epigraphic material provided by some 15 different epigraphic archives (about 80% of the classified epigraphic material from the Mediterranean area) for ingestion to Europeana. The collected material will be made available also to the scholarly community and to the general public, for research and cultural dissemination. This paper briefly presents the main services provided by EAGLE and the challenges encountered for the aggregation of material coming from heterogeneous archives (different data models and metadata schemas, and exchange formats). EAGLE has defined a common data model for epigraphic information, into which data models from different archives can be optimally mapped. The data infrastructure is based on the D-NET software toolkit, capable of dealing with data collection, mapping, cleaning, indexing, and access provisioning through web portals or standard access protocols.

Keywords: Data Infrastructure, Aggregation System, Metadata Formats, Data Interoperability, Data Harmonization, Epigraphy, D-NET.

1 The EAGLE project

Ancient inscriptions are a valuable source of information about otherwise undocumented historical events and past laws and customs. However, centuries of unregulated collection by individuals and by different institutions has led to an extremely fractioned situation, where items of the same period or from the same geographical area are presently scattered across several different collections, very often in different cities or countries.

One of the main motivations of the project EAGLE (Europeana network of Ancient Greek and Latin Epigraphy [1], a Best Practice Network partially funded by the European Commission) is to restore some unity of our past by collecting in a single repository information about the thousands of inscriptions now scattered across all Europe.

The collected information (about 1,5 million digital objects at project's end, representing approximately 80% of the total amount of classified inscriptions in the Mediterranean area) are to be ingested into Europeana, as they represent the origins of the European culture. That information is also made available to the scholarly community and to the general public, for research and cultural dissemination, through a user-friendly portal supporting advanced query and search capabilities.

The EAGLE project comprises 19 partners from 13 European countries. Fifteen partners (called Content Providers) are cultural institutions owning collections of Latin and Greek inscriptions, and are the ones actually providing the data to EAGLE and to Europeana. Four partners (called the Technology Providers) are the ones in charge of setting up the aggregation infrastructure, connecting with Europeana, de-duplicating the received data (the same inscription can appear in the collections of more than one Content Provider) and developing the services available at the EAGLE portal.

In addition to the "normal" query capabilities (full text search a la Google, fielded search, faceted search and filtering), the EAGLE portal supports two applications intended to make the fruition of the epigraphic material easier and more useful. The Mobile Application will enable a (mobile) user to get information about one visible epigraph by taking a picture with a mobile device, and sending it to the EAGLE portal specifying the recognition mode. In Exact Match mode, the EAGLE system will return all the information associated with the recognized image, or a message indicating that the image was not recognized. In Similarity Search mode, the EAGLE system will return a list of the most visually similar epigraphs, ranked in order of similarity.

The Story Telling application will provide tools for an expert user (say a teacher) to assemble epigraphy-based narratives providing an introduction to themes and stories linking various inscriptions together (e.g. all the public works done by an emperor). The stories are then made available at the EAGLE portal, and are (usually) intended for the fruition of the epigraphic material by less knowledgeable users or young students.

Along the same lines, in order to make the epigraphic material more interesting and usable also by non-epigraphists, EAGLE, in collaboration with the Italian chapter of the Wikimedia Foundation, is leading an effort for the enrichment of the epigraphic images and text with additional information and translations into modern languages. This additional material, residing on Wikimedia, is periodically harvested and included in the information associated with each epigraph.

2 The EAGLE Aggregation Infrastructure

As mentioned before, EAGLE aggregates content provided from 15 different archives from all over Europe. While most of them are providing records based on EpiDoc (an application profile of TEI, today the de-facto standard for describing inscription), some archives are supplying records in "personalized" formats. EAGLE aggregates data also from two other different sources: Mediawiki pages, containing transla-

tions of inscriptions, and "Trismegistos records", containing information about inscriptions that appear in more than one collection.

The need for expressing queries against such heterogeneous material has led to the definition of a data model being able of relating separate concepts and objects in a seamless way, thus allowing both the scholarly research and the general public to achieve results which could hardly be obtained with the existing EpiDoc archives.

The EAGLE data model [2] consists of an abstract root entity (the Main Object) from which four sub-entities can be instantiated: (*i*) Artefact (capturing the physical nature of an epigraphy); (*ii*) Inscription (capturing the textual and semantic nature of a text region possibly present on an artefact); (*iii*) Visual representation (capturing the information related to the "visual nature" of a generic artefact); (*iv*) Documental manifestation (capturing the description of an inscription's text in its original language and its possible translations in modern languages). All the information to be aggregated in EAGLE will find its place into one or multiple instances of such sub-entities.

The EAGLE Aggregation Infrastructure is built on top of the D-NET software [3], developed by ISTI-CNR in the course of its participation in a number of European projects. D-NET is an open source solution specifically devised for the construction and operation of customized infrastructures for data aggregation, which provides a service-oriented framework where data infrastructures can be built in a LEGO-like approach, by selecting and properly combining the required services [4]. For EAGLE, D-NET has been extended with image processing services to support the Mobile Application(see bottom of Fig. 1). The aggregation process consists of four main phases (see top of Fig. 1):

1. *Metadata mapping*. The definition of the metadata mapping covers two key aspects: (*i*) a *structural mapping* from the archives' local schema to the EAGLE common schema and (*ii*) a *semantic mapping* from the local vocabularies to the EAGLE ones. In cooperation with domain experts from the archives providing records, the structural and semantic rules to map the incoming records into the EAGLE common metadata schema are encoded in the form of XSLT scripts.
2. *Metadata transformation and cleaning.* Metadata records are collected (via FTP, OAI-PMH or other protocols) and processed (using the transformations defined in phase 1) to generate the "EAGLE objects", thus creating the Pre-production Information Space.
3. *Metadata quality control*. The EAGLE records are inspected and validated to identify mapping errors and possible mistakes (e.g., typos). This quality control process may lead to the redefinition of the mapping rules (in Phase 1) and the repetition of Phase 2.
4. *Metadata provisioning*. The EAGLE records that pass Phase 3 are indexed and become available for querying and browsing through the EAGLE portal, or for ingestion to Europeana (after a further transformation to make them compliant with EDM, the Europeana Data Model) or for harvesting through standard APIs by other (external) applications.

In D-NET, data processing is specified by defining *workflows* (i.e. a graph of elementary steps, with optional fork and join nodes) and *meta-workflows* (i.e. a sequence

of *workflows*). A (meta)workflow can be easily configured, scheduled and started through a D-NET tool with a graphical user interface, while the implementation of the elementary steps is done by writing programs actually executing the needed processing. For example, the metaworkflow for processing EpiDoc documents consists of the sequence of the following workflows (where each workflow consists in a set of more elementary steps):

- collect (records from an archive);
- transform (the collected records into the EAGLE metadata format);
- clean (metadata quality control);
- index (build the index of the final records for querying and browsing);
- provision (put the final records in the format required by an OAI-PMH harvester).

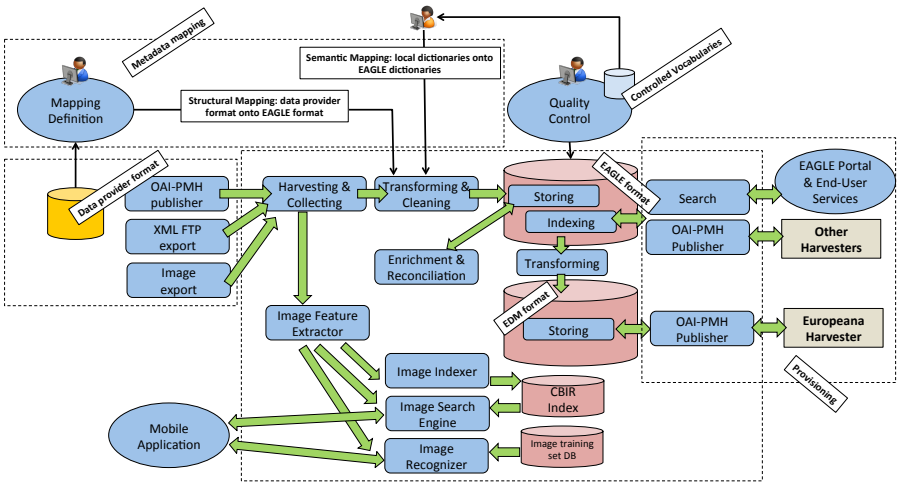

Figure 1 – The EAGLE Aggregation Infrastructure

3 References

1. The EAGLE project http://www.eagle-network.eu/
2. Casarosa, V., Manghi, P., Mannocci, A., Rivero Ruiz, E., Zoppi, F. "A Conceptual Model for Inscriptions: Harmonizing Digital Epigraphy Data Sources". EAGLE International Conference on Information Technologies for Epigraphy and Digital Cultural Heritage in the Ancient World, Paris, September 29-30 and October 1 2014.
3. D-NET Software Toolkit, http://www.d-net.research-infrastructures.eu.
4. Manghi, P., Artini, M., Atzori C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D., Pagano, P. "The D-NET Software Toolkit: A Framework for the Realization, Maintenance, and Operation of Aggregative Infrastructures". In EmeraldInsight Program 2014, vol. 48, p. 322-354, doi 10.1108/PROG-08-2013-0045.