Implementing a database for the analysis of ancient inscriptions: new developments in the Hesperia electronic corpus of Palaeohispanic inscriptions

Eugenio R. Luján – Eduardo Orduña
Universidad Complutense de Madrid

1. Introduction*

The purpose of the Hesperia databank of Palaeohispanic languages is to collect, organize and process all the ancient linguistic materials pertaining to the Iberian peninsula (and any related materials in southern France), with the exception of Latin, Greek and Phoenician inscriptions.[1] More than 2,000 Palaeohispanic inscriptions are known to date and, fortunately, the number increases every year. These inscriptions, which can be dated from at least the 6th century BC[3] to the beginning of the 2nd century AD, were written in at least five different languages: Iberian, Celtiberian, Lusitanian and two unidentified languages.[4] For these inscriptions several varieties of Palaeohispanic scripts were employed over time and space,[5] but we also have texts written in the Greek and Latin scripts.[6] Besides the inscriptions, for the study of Palaeohispanic languages we must also take into account the indigenous names (personal names, place names, ethnonyms, and god names) and additional information about the Palaeohispanic languages, such as glosses, that have been transmitted by a number of Greek and Latin sources, both literary and epigraphic.

* This paper is part of the research project FFI2009-13292-C03-02, which has the financial support of the Spanish Ministry of Economy and Competitiveness. We are very grateful to two anonymous reviewers for their comments and suggestions.

[1] The term 'Palaeohispanic' is employed here in the sense of 'old languages and inscriptions from *Hispania*'. The languages attested on those inscriptions are frequently referred to as 'Pre-Roman' languages of Spain. However, this name is misleading, given that, in fact, most inscriptions are later than the Roman settlement in Spain; ' Palaeohispanic' is thus preferable.

[3] The date can only be approximate, since most of the oldest inscriptions (the Southwestern inscriptions) have not been found in their original archaeological context. See De Hoz 2010, p. 358–361.

[4] Recenlty, Koch 2009 and 2011 has argued that the language of the Southwestern inscriptions—the oldest ones—is Celtic. However, most scholars working on Palaeohispanic languages and inscriptions have not accepted this proposal, since Koch's analyses and arguments are far from cogent. See De Hoz 2010, p. 386–402, for a reliable discussion of our current knowledge about the language of these inscriptions.

[5] Palaeohispanic scripts originated from Phoenician script; see De Hoz 2010, p. 485–517. Several varieties were employed over time: Southwestern script, Southern script and 'Classical' Iberian Levantine script, from which two slightlty different adaptations were made for the writing of Celtiberian. Updated tables of these scripts can be found in De Hoz 2010, pp. 618-625.

[6] There are Iberian inscriptions written in a variety of Ionian alphabet, and Celtiberian inscriptions written in the Latin script. All known Lusitanian inscriptions are written in Latin script as well.

This kind of information is fundamental, e.g., for our knowledge of the languages related to Basque in Antiquity.[7]

The study of Palaeohispanic inscriptions poses problems of different kinds that we had to bear in mind when designing and developing a database that would allow us to digitize and process them in an appropriate way.[8] Two of the identified languages belong to the Indo-European family of languages (Lusitanian and Celtiberian), one of them being specifically Celtic. The third, Iberian, remains an isolate one, in spite of different proposals. All of them are only partially understood, especially as far as longer documents are concerned, and many aspects of their grammar and vocabulary are still unclear. This difficulty is amplified in the case of Iberian, for which the recourse to comparison to other genetically related languages is not possible. As for the scripts, the Iberian Levantine script is totally deciphered, as well as its adaptations for Celtiberian. However, the values of a certain number of signs of the Southern and Southwestern varieties remain highly controversial.[9]

Therefore, one of the primary goals of the database must be to contribute to the advancement of our knowledge of these languages and scripts, allowing for complex searches that may shed light on the grammar of the Palaeohispanic languages and the interpreation of the texts, providing a quick and reliable means of testing hypotheses about debated aspects of these languages, and making it easier to confirm or reject possible phonetic values for the still undeciphered signs of some of the Palaeohispanic scripts, among other various possibilites. The database must also be useful for a better understanding of certain phenomena in the history and distribution of these languages, such as identifying lexical isoglosses, patterns of spread of some of the epigraphic features, and regional or diachronic biases of certain phenomena, in order to gain a deeper insight into the dynamics of these languages and inscriptions.

The Hesperia databank is now a joint project of several research teams of the Complutense University at Madrid, the University of the Basque Country, and the Universities of Zaragoza and Barcelone.[10] It is hosted on a dedicated server of the Faculty of Philology of the Universidad Complutense at Madrid and it is accesible through the web site of the project (http://hesperia.ucm.es/).

The Hesperia databank was originally a FileMaker database developed by Fernando Quesada, from whose design it still retains its graphical aspect—viz., most of the original fields and the distribution in tabs with a heading. Now Hesperia is

[7] The fundamental work to date is Gorrochategui 1984. For a recent review of what we know about Basque in Antiquity see Gorrochategui 2009.

[8] The standard edition for these inscriptions is to date Untermann's *MLH*.

[9] This is one of the basic problems with Koch's proposals (see fn. 4). His interpretations are based on values of the signs of the Southwestern script that are not assured.

[10] For a presentation of the project in previous stages see Luján 2005.

implemented internally as a MySQL database containing several tables, corresponding to the different sections: Epigraphy, Numismatics, Onomastics, Lexicon, and Bibliography. The graphical user interface is an HTML form,[11] built using the PHP programming language, each form corresponding to a MySQL table. Only the Numismatics section has two tables, and we will refer to them later. Each section, except Lexicon and Bibliography, has tabs in order to keep all the fields in the screen without the need to scroll down to see all of them.[12]

Some of the databases are now in an advanced state of development. However, given the diversity of the materials that had to be integrated into the databank and the technical complexity of dealing with all the different types of materials (e.g., fonts for the transliteration of the scripts, photographs and drawings of the inscriptions, cartography, and so on), the various parts of the databank will be fully operational in different phases. The goals of the databank have also evolved since its initial conception and we now have more ambitious prospectives. This has led us to make certain improvements in the databank by integrating additional tools and implementing new ways of dealing with the information that we have compiled. We thus believe that the Hesperia database can provide interesting methodological cues for other digitized corpora of ancient inscriptions.

We have provided general descriptions of the databank and all of its sections in previous works (Orduña – Luján – Estarán 2009, Orduña – Luján in press),[13] so we will focus here in more detail on some parts of the Epigraphy section that can be seen as more innovative in certain aspects, or that have a deeper complexity. We will focus especially on those features related to the storing and representation of transcriptions of Palaeohispanic texts and their critical apparatus, the search engine, and the dynamic generation of maps. We will also provide some more detail about certain new improvements in the Hesperia databank that were not described in our previous works—in particular the new way of presenting bibliographic references, as well as the bibliographic impact searcher. We will also describe the new way to display the lexical references that correspond to a certain text, as this was still work in progress at the time of our prior publications.

[11] Here we use 'form' in the technical sense of the HTML tag. In the rest of this paper we will prefer 'section' or 'tab', as our graphical user interface is divided in tabs, each one containing one HTML form. Each 'section' corresponds to a MySQL table.

[12] The entry of new data into the database is only allowed for users with privileges, so in this paper we will mainly focus on what a normal user will see.

[13] The original posters presented at the Lisbon conference can be accessed on the Internet (http://eprints.ucm.es/8672/) and provide a general overview of the main features of the databank.

2. Browsing the database

We will first discuss a simple but, in our view, important feature of the Hesperia databank: it is easy to navigate. Unlike many databases on the Internet, in which the user must start from a search engine,[14] in Hesperia it is possible to browse all the records in the same way as in other desktop applications such as Filemaker. Once the users enter the database, they find the first record, and they can browse through the records with the 'next/previous' buttons with which all tabs are provided, thus navigating through all of the records in the order that they were entered. Every user can thus have a glimpse of the database without any previous knowledge, as would be required if they were confronted first with a search engine. It is also possible to filter the records according to simple criteria, for example 'language = Iberian' or 'archaeological site= Empúries'. Once the filter has been selected, the navigation buttons act only over the matching records.

Another way of navigating the database is by using the search engine, which allows the user to enter search criteria for every field, in a multi-filter way that could be labelled as 'faceted browsing'[15]: all the matching results are displayed as a table with the contents of some important fields (e.g. 'text'), and each result has a link to the complete record. The results are sorted in ascending or descending alphabetical order according to the field selected by the user.

Last, but not least, it is possible to navigate 'geographically': the MapServer program, which we will discuss later, allows the user to see the points corresponding to all the archaeological sites that have provided Palaeohispanic inscriptions. The user can then click on them to see the list of the inscriptions found at that site, with links to the relevant records.

We will now turn specifically to the Epigraphy section, focusing on its more innovative aspects. As a general background, it will suffice to say that it is organised in tabs, with a heading always visible over them that contains some general information about the inscription, such as the name of the archaeological site, its location, or its references in epigraphic corpora. There are six tabs, which contain other general information about the epigraphic object itself, the text and its critical apparatus, illustrations, epigraphy and palaeography, archaeological context, and bibliography. The Illustrations tab allows the uploading of an unlimited number of image files of the inscriptions, and they appear as a table containing the images as thumbnails, which are links to the full size images (Figure 1) and are automatically generated whenever an image is uploaded. These image files are not stored in the

[14] This is the case, e.g., with the "Epigraphische Datenbank Heidelberg" (http://www.uni-heidelberg.de/institute/sonst/adw/edh/index.html.en), which contains Latin inscriptions.

[15] 'Faceted browsing' is a technique for accessing information that allows users to browse that information applying multiple filters at the same time.

MySQL database, but in folders or directories whose names correspond to the record id numbers, so that the database files remain small in size and are easy to back up.

3. Text and critical apparatus

In the Epigraphy section there is a tab devoted exclusively to the transcription of the Palaeohispanic text and its variant readings. Palaeohispanic inscriptions are, for the most part, unique and we thus lack different versions of the same text, so in this context 'variant' is intended to refer to an alternative reading proposed by a scholar that is not the one selected for the main text. In Palaeohispanic epigraphy it is especially important to provide the variants since, as we have remarked in §1, we still have a limited knowledge of these languages and in many cases we lack cogent reasons to definitively reject certain proposals. It is thus crucial to have a way of presentig the alternative readings in such a fashion that, on the one hand, they do not intrude upon the user when reading the text, and on the other hand, they are not neglected because the user would need to look away from the main text every time that there is a variant for a word.

We have found a solution that, in our view, fulfills both requirements: each section of text with alternative readings is marked in blue, and a floating bubble appears over this section, containing the variant spellings, all in a single window (Figure 2).

We refer to sections of text which have alternative spellings as 'words', in the sense of a section of text separated by punctuation marks. We use them as a natural way of segmenting the text, since this is how the Iberian and Celtiberian scribes proceeded themselves, whether or not it corresponds to our concept of a word; this is convenient because in this way variant spellings are easily found by the search engine. Let us suppose that we have selected as the preferred transcription for a text *tautintibas*, while other scholars have read *boutintibas*. If we marked only *$ta&utintibas* (the $ and & signs enclose the text with an alternative reading in our notation)[17], corresponding to *$bo&* in the critical apparatus, this would be useless: a user looking for *boutintibas*, or *boutin*, would not be able to find it, as the field containing the variants would have only *bo*. As we can see, the longer the texts that we mark as variants, the easier it will be for a user to find these variants. Punctuation marks are usually a good splitting point when they exist, but sometimes it is better to cut shorter sections.

We have said that the text and its variants both appear in a single window. In fact, this is not a true HTML form field, but a white table cell that allows formatting.

[17] These signs were chosen arbitrarily at the first stages of the project because they do not correspond to any other sign used in the transcription of Palaeohispanic languages, and they can thus be safely ignored by the search engine.

The real HTML form, only accessible to the team members,[18] contains two fields, one for the transcription selected as the preferred reading, and the other for the critical apparatus. Once the main text is entered, the editor selects in order each word or section of text with variants, and by pressing a button, this word is enclosed between the $ and & marks and then automatically copied into the Critical Apparatus field, leaving the cursor ready to enter the alternative readings. A semicolon is used at the end of all the variant spellings for one word in order to separate them from the next word.

4. Lexical references

One of the tables upon which Hesperia is internally built is the Lexicon table, which stores the lexical entries and is accessible through the corresponding section. The lexical entries are 'words' in the sense described in the previous paragraphs. Due to the current state of knowledge of the Palaeohispanic languages—which is very limited in terms of semantic meaning, and in the case of the Iberian language, practically nil—the contents of these lexical entries have little to do with what one would expect in the case of better known languages. These entries typically contain little more than bibliographic references[19] or a list of other possibly related words or segments.[20]

One advanced feature of Hesperia is the simplicity with which one may consult this lexicon through the text itself: one link in the Epigraphy and Palaeography tab opens a window with the transcription of the text, in which each word, in the sense already explained, is a link to its lexical entry (Figure 3). This is made possible by means of a function that stores the text in an array, using the punctuation marks as separators, so that each word is an array element. Before comparing each array element with the entries in the lexicon, some functions are internally used in order to allow the match even if the transcription system has not been the same in the lexicon and in the text, or if there are some other anticipated differences. For example, the text transcription or the lexical entry can be outdated regarding the transcription of a sign previously thought to be **bo**, the correct value of which is, however, **ta**.

The links in the text open a side window with the lexical entry, in which every word in bold (which we usually employ only to transliterate Iberian words) is transformed also into a link to this entry. In this case we have provided a simple option of creating a link to this word without previously checking if this entry exists or not, so that it is possible that the link sometimes points only to a 'There is no entry

[18] The same approach is used in other tabs.

[19] These bibliographic references are manually entered, in the author-year format, pointing to the full reference included in the general bibliography, which is also manually entered.

[20] See De Hoz 2011, p. 290, n. 127 for a survey of the existing Iberian lexicons and its problems. As he points out, they do not usually propose a segmentation of the lexical entry.

for…' message. This occurs when the lexical entry mentions suffixes or other text strings smaller than those used as lexical entries. On the other hand, this system is very useful when the lexical entry simply points to another entry and also in order to keep different readings or transcriptions of a word in the lexicon. Such entries are also employed in the few cases in which we are certain enough to separate a lexeme from one or several affixes: e.g., for the frequent Iberian word *śalir* (most probably 'silver') we have an entry *śalir*, and the entries *śalirg, śalirban, śalirnai* simply point to the first one.

5. Bibliography

In Hesperia we have developed a new way to present bibliographic references. There is a Bibliography table that contains all the bibliographic records cited in the different sections of the database. One of the fields of this table contains the short citation form, in the author-year style that will be used to cite each bibliographic record. There is also a link for each record that displays the number of records in the Epigraphy, Onomastics or Numismatic sections that cite this bibliographic entry, followed by the list of links to each of these records.

In the Bibliography tab of the Epigraphy, Numismatics or Onomastics section, we enter only the short form of the references, usually followed by page number(s), and the user can see these short references in form of a link that shows a floating window with the complete reference (Figure 4).

6. The search engine

All the sections in the Hesperia databank are provided with a search engine that supports searches on matching criteria entered in any field of the database—or in several fields at a time—with the option of looking for records that match the criteria provided in any one of the fields, or alternatively in all of them (Figure 5). For each field it is also possible to look for an exact match or for a 'regular expression' match. As we have already said, the results appear in the form of a table, with links to each record.

In its more common use, the regular expression search simply involves the ability to look for parts of the contents, for example *Puig* instead of *El Puig de Sant Andreu*, the name of an archaeological site. But, in fact, it is a much more powerful engine that allows for very complex searches. Our search engine can be accessed in three ways: (a) simple search, with a single window to enter a text that will be searched for in every field; (b) easy search, which allows for entering search criteria in only some fields; and (c) advanced search, which contains all the fields. If the "text" field is specified, the engine also automatically searches the critical apparatus. This search engine, in all three modalities, uses special PHP functions that emulate the Perl

regexp engine and use the same regexp language.[21] One important point of this particular regular expression search feature is its ability to ignore Unicode diacritic marks (combining characters) such as 'underline' or 'dot under', used in our transcriptions for marking uncertain signs or signs only partially preserved but with assured reading. The match is thus possible without using these marks in the search string.

The regular expression search engine provides the user with a set of wildcard symbols such as \b (boundary), \w (word character, i.e., an alphabetical character), \s (white space), etc., which can form very complex expressions when combined with text strings.[22] One example will suffice to show the power of the regexp language: the expression *(baś)?b?i[td]?[ei][rŕ]ok(an|ar|e)(te)?(tine)?* can match all the different forms of a probable Iberian verbal paradigm whose lexeme is still uncertain, regardless of the different spelling of the vibrants, sibilants and occlusives, and the different possible suffixes: *biteŕoketetine, biteŕokan, basbiteŕoketine, eŕokar, bideŕokan,* etc.

As we have already said, the results of a search are displayed as a table that includes some of the fields (archaeological site, reference, and text). If the user is looking for a text string, the matching parts of the text are marked in red (Figure 6). All the records, or some of them, can be selected in order to be included in a PDF file, or to be represented in a map. In the case of the PDF file, the user can also select the fields to be included.

7. Dynamic map generation with MapServer

One of the most interesting features of our database is its ability to show dynamically generated maps with MapServer. This is an open source CGI program, developed by the University of Minnesota, that takes geographical information from a 'shapefile'[23] or a database and represents it on a map, allowing access to all the information contained in the 'shapefile' or the database related to a geographical point. The program requires certain files for every MapServer application: one or several HTML templates that act as a graphical user interface, and a 'mapfile' or configuration file that tells the program where to find the geographical information, the templates, or other files such as icons, and contains the layers that will appear in the map. These layers can be set by default or selected by the user. In our case, a physical map of the Iberian Peninsula always appears by default, and the user can select the other layers: point layers showing the distribution of archaeological sites, for example, or raster

[21] PERL is considered the programming language with the most powerful regexp engine.

[22] The standard book about regular expressions is Friedl 2002.

[23] A 'shapefile' is a multifile format consisting of three files, a .shp file containing the geographical coordinates of the objects, a .shx one whith the indexes, and a .dbf one with the attributes of the objects.

WMS (Web Map Server) layers obtained from external servers showing aerial orthophotos.

We use this program in two different ways: statically, to show layers with all the inscriptions, or dynamically, to generate maps that contain all or some of the results of a search.

The first method is accessible through the link 'MapServer', which provides a map in which the user can select different layers, some of them corresponding to external WMS servers that provide satellite photos or cartographic maps, and the rest corresponding to the different sections (i.e., tables) of the database: Epigraphy, Onomastics, and Numismatics. Selecting one of these layers, the user can see all the points that have provided Palaeohispanic inscriptions, Latin inscriptions with Palaeohispanic personal names, or the distribution of Palaeohispanic mints. In the case of epigraphic points, there are different icons according to the writing system employed at that site (Figure 7).

In the second method, the search engine generates a table with all the matching records, and each one has a check-box used to select the records to be included in a PDF file or in a map. There is also a button that provides direct generation of the map with all the results. A map is generated only with the points corresponding to the matching results, and the list of records with their corresponding links is shown below the map.

Regardless of the method used, the generated map has different controls, and two main working modes: navigation or information. In the navigation mode the user can zoom or move the map. Starting from the first level of zoom, labels appear next to the points showing the name of the place: that is, the archaeological site in the epigraphical layer, the town or city in the onomastic layer, and the mint name in the numismatic layer (Figure 8). In the information mode the user may click on the icon of a point to obtain the list of all the records located in that point and the links to each record. It is very easy to use MapServer in this way to browse the database in a geographical approach.

Both the static point layers and the dynamic search result point layers are preserved by default while using the zoom feature or after clicking for information on a point. In this way we can see the points over the aerial orthophotos of the archaeological sites, which can have a resolution of up to 25cm/pixel (Figure 9), so that when, in the near future, we have more precise geographic coordinates for each inscription inside an archaeological site, we will be able to see the distribution of the inscribed objects over the site.

8. Other sections of Hesperia

We have already described the Bibliography and the Lexicon sections of Hesperia, and discussed how they interact with the Epigraphy section. There are two other important sections, Numismatics and Onomastics, which we have already described in detail in previous works, so we will only provide a short overview here.

Their graphical user interfaces work in a very similar manner to the one already described for the Epigraphy section, and both Numismatics and Onomastics have a complex search engine that allows for dynamic map generation and PDF file generation as well. The Numismatics section is internally organised in two tables, the main one containing mints and general information about them, and the other, the coin legends corresponding to each mint. The Onomastics section is organised in four tabs: Anthroponymy-Corpus, Anthroponymy-Analysis, Theonymy-Analysis and Toponymy, each one corresponding to a MySQL table. The first of these tabs contains all the names and the geographical coordinates studied in more detail in the next two tables, and the last one contains all the place names mentioned by ancient sources. This last section is less directly connected to the other three. Searches in each of the first three tabs allow for dynamic map generation, taking the geographical information from the first one. The 'mapfile' or configuration file used by the MapServer program allows for setting different layers from only one MySQL table, using filters that in this case take a form similar to 'type = anthroponym' or 'type = theonym', 'type' being one field in the Anthroponymy-Corpus table.

Finally, there is a Charts section that consists of a search engine similar to the one described above. Here the user may view the distribution of the results according to a given feature (such as the type of material of the inscriptions), which is presented as a pie chart showing percentages. It is thus very easy to detect certain interesting facts: for example, a chart representing the distribution of the frequent Iberian word *śalir* according to the material of the inscription clearly shows that this word, apart from coins, appears only on lead tablets.

References

J. de Hoz (2010), *Historia Lingüística de la Península Ibérica en la Antigüedad*, I. *Preliminares y mundo meridional prerromano* (Manuales y Anejos de *Emerita* 50), Madrid: Consejo Superior de Investigaciones Científicas.

J. de Hoz (2011), *Historia Lingüística de la Península Ibérica en la Antigüedad*, II. *El mundo ibérico prerromano y la indoeuropeización* (Manuales y Anejos de *Emerita* 51), Madrid: Consejo Superior de Investigaciones Científicas.

J. Friedl (2002), *Mastering Regular Expressions*, Sebastopol: O'Reilly (2nd edition).

J. Gorrochategui (1984), *Onomástica indígena de Aquitania*, Bilbao: Universidad del País Vasco.

J. Gorrochategui (2009), 'Vasco antiguo: algunas cuestiones de geografía e historia lingüísticas', in F. Beltrán et al. (eds.), *Acta Palaeohispanica X. Actas del X Coloquio sobre Lenguas y Culturas Paleohispánicas* (= *Palaeohispanica* 9), Zaragoza: Institución Fernando el Católico, p. 539–555.

J. T. Koch (2009), *Tartessian: Celtic in the South-west at the Dawn of History*, Aberystwyth: Celtic Studies Publications – David Brown.

J. T. Koch (2011), *Tartessian 2. The Inscription of Mesas do Castelinho. ro and the Verbal Complex. Preliminaries to Historical Phonology*, Oxford: Oxbow.

E. R. Luján (2005), '*Hesperia*. The electronic *corpus* of Palaeohispanic inscriptions and linguistic records', in *Review of the National Center for Digitization* (Belgrade) 6, p. 78–89, (http://elib.mi.sanu.ac.rs/files/journals/ncd/6/d007download.pdf).

MLH = J. Untermann (1975–2000), *Monumenta Linguarum Hispanicarum*, 4 vols., Wiesbaden: Reichelt.

E. Orduña – E. R. Luján – M. J. Estarán (2009), 'El banco de datos Hesperia', in F. Beltrán et al. (eds.), *Acta Palaeohispanica X. Actas del X Coloquio sobre Lenguas y Culturas Paleohispánicas* (= *Palaeohispanica* 9), Zaragoza: Institución Fernando el Católico, p. 83–92, (http://ifc.dpz.es/recursos/publicaciones/29/54/08ordunaetal.pdf).

E. Orduña – E. R. Luján (in press), 'Philology and technology in the Hesperia databank', *Journal of History, Literature, Science and Technology (JHLiST)*.