Lagarto: paleographic transcription using crowdsourcing and simple IT

M.T. Carrasco Benitez - 3 March 2025

Overview

Lagarto is a set of auxiliary techniques and tools for paleographic transcription, mainly crowdsourcing and very simple IT tools for palaeographers. The rationale is to facilitate the simultaneous work of many people and lowering the IT barrier for palaeographers by using a simple text editing in a wiki or editor, as opposed to specialised tools. Priority is given to quickly produce and publish transcriptions at segment level (roughly a sentence). Subsequent passes would take care of finer markups.

El Libro

The name Lagarto is in honor to the Nave of the Lagarto of the Seville Cathedral where the Institución Colombina [IC] is hosted. The Lagarto website is `http://lagarto.top` and references to this website are inserted in curly brackets; example {format} refers to `http://lagarto.top/format`

Lagarto is developed by starting from a concrete case: *El Libro de los Epítomes* (The Book of Epitomes, referred as *El Libro*) [LE] that contains summaries of books of the library of Ferdinand Columbus. In the Repertorios of Hernando Colón {rep} it corresponds to numbers 4 and 5.

The facsimile mostly used in this project is the incomplete AM 377 fol. in the Arnamagnæan Collection (Copenhagen) [LE-C], a hand written clean version in a reasonable clear calligraphy, mostly in Medieval Latin with at least one epitome in Spanish {led/mix/global.html#508}; it contains 1877 epitomes. There is also a draft version in the Institución Colombina (Seville) [LE-S] with additional and overlapping epitomes; no facsimile is publicly available and the only one included comes from secondary publication {led/mix/global.html#1229}. In total "over 2150" epitomes [BEE]. The Catálogo Concordado [CC] is mainly used for metadata, as well as other sources {link} [LOM].

The projects *Book of Books* [BB] and *Hernando Colon's Book of Epitomes* [BE] also work in the transcription and translation of El Libro.

Verification of the Lagarto approach

The Lagarto approach (crowdsourcing and simple IT) was applied to the transcription of El Libro by ten students from the Complutense University of Madrid, called *sumistas* in honor of the original creator of the epitomes; additional sumistas were in the process of being recruited. The work suddenly stopped after a couple of weeks as sumistas were directed to stop working on the project by teaching staff at their university. Even this short period was sufficient to confirm that the approach works, both the crowdsourcing and the IT techniques.

The metadata Lagarto subproject producing the *List of all epitomes* {led/mix/global.html} is very active.

Basic description

In essence, the task of the sumistas is writing transcribed text segments and at least a translation, one per line. For example:

```
t1:Georgius ſymler Vuimpinenſis natione theuthonic         [transciption]
es:Georg Symler, natural de Wimpfen, de la nación teutona  [Spanish translation]
```

Transcribing XVI Latin is a task only for people with the appropriate training. Sumistas start a transcription by following the instructions {ins}, essentially editing in the Lagarto wiki [LW]; that's all, and from there onwards Lagarto takes care. Among others, a wiki facilities tracking previous versions, and scale well for organising hackathons [HAC]. Having a bespoke system for transcribing would be ideal, but repurposing existing software is faster and costs less.

The *Lagarto Processor* takes as input *Lagarto Format* {format} data and produces as output the *Libro de los Epítomes Dossier* (LED), a web-based package heavily interlinked, viewable online {led} and downloadable as one zip file {led.zip}.

Lagarto abstract data type (ADT)

An informal skeleton overview:

```
- book                          [root element - the whole Libro]
     - section                  [section 1 - ordered list - section = epitome]
          - head                [header - metadata - full list of elements at {format}]
               - id             [section identifier - epitome number]
          - body                [aligned segments]
               - align          [multilingual aligned segments - align 1 - number implicit by the order]
                    - t0        [segment - clean or consolidated texts of t1 and t2]
                    - t1        [segment - transliteration from source 1 - there might be only one source]
                    - t2        [segment - transliteration from source 2 - up to N]
                    - es        [segment - Spanish translation]
                    - note-t0   [note - for segment 0]
                    - note-t1   [note - for segment 1]
                    - note-t2   [note - for segment 2 - up to N]
                    - note-es   [note-  for segment es]
               - align          [multilingual aligned segments - align 2 - up to N]
     - section                  [section N]
```

Many elements are optional. It can be implemented as several formats, one of them is the Lagarto Format. A *document type definition* [DTD] might be written later.

Lagarto Format

It is an implementation of the ADT {format} where each `section` is in a separate wiki page or file, both representations contain the same plain text record-jar format [RJ]: a serie of records separated by "`%%`" (two percentage characters) at the beginning of the line, where each line is a *key-value pair* separated by "`:`" (column character). The first record is the `head` and the rest of the records the `body`, where each record is an `align`, aligned *multilingual parallel segments* and the corresponding notes; sumista decide on where to break the text to create the segments. Partial illustrative example of {led/epi/142/raw.txt}:

```
id:142
title:Observationes de arte grammatica
%%
t1:Georgius ſymler Vuimpinenſis natione theuthonic <lb/> grâmatices obseruationes côpilauit
la:Georgius Symler Vuimpinensis natione teutonica grammatices obseruationes compilauit
es:Georg Symler, natural de Wimpfen, de la nación teutona, compiló unas "Observaciones de gramática",
%%
t1:quib cuncta <lb/> fere grâmaticalia fundamenta potius agregata quâ <lb/> digesta uidentur
la:quibus cuncta fere grammaticalia fundamenta potius aggregata quam digesta uidentur,
es:en las que prácticamente todos los fundamentos gramaticales parecen entremezclados, antes que estructurados,
%%
t1:Vnde pro uectiorib potius quam <lb/> nouis tyrunculis illud lectitand censuerit
la:unde pro uectioribus potius quam nouis tirunculis illud lectitandum censuerit
es:de lo que habría aconsejado, más en favor de los más avanzados que de los principiantes jóvenes, leerlo repetidas veces,
%%
```

The `align` elements are implicitly numbered by the order of appearance, numbers are shown after processing {led/doc/parallel.html}. Only a few epitomes have been transcribed. On the other hand, many headers have been partly filled (all in LE-C) and used to generate the List of all epitomes {led/mix/global.html}, a heavily hyperlinked dashboard of El Libro.

ADT can also be implemented in XML {book.xml} or a database management system though it would require a bespoke interface. Without specialised tools, it is easier to write RJ than XML, though there are some work arounds {section.html}. Efforts should be in the direction of generalising the particular case of El Libro to other books. Indeed, with minor changes, the Lagarto techniques can be apply to other books.

Libro de los Epítomes Dossier (LED)
It contains different views and presentations.

* Comprehensive dashboards:
- All segments in parallel: it could be considered *transcribing-aid tool*; it facilitates the side by side comparison of texts with the corresponding notes {led/doc/parallel.html}.

- List of all epitomes: see description below.

* Whole Libro:
- original text Medieval Latin
- translation into Classical Latin
- translation into other languages such English and Spanish
- all segments in parallel in HTML and CSV

* Per epitome:
- index, example: {led/epi/142}
- metadata
- original Medieval Latin (HTML, TXT, TEI) and word frequency
- translation into Classical Latin (HTML, TXT, TEI) and word frequency
- translation into other languages such English and Spanish (HTML, TXT, TEI) and word frequency
- all segments in parallel in Lagarto Format, HTML and CSV
- raw metadata and pristine data created by the sumista{led/epi/142/raw.txt}

The granularity of Text Encoding Initiative [TEI] corresponds to the level of marking in the Lagarto Format; at segment level, it might be coarser than expected in TEI.

Formally, LED is a container format based on Xdossier [XDOSSIER]. The content is mostly textual. Other formats should be viewable by most web browsers without dependence on external programs. LED can be explored with the standard tools of any operating system, graphicals or command-lines, such as Windows File Explorer or Linux command-line. It is also designed for longterm digital preservation with a two-pronged approach: textuality and many distributed copies by encouraging downloads of the zip file {led.zip}.

As all the pristine data produced by the sumistas is included in LED, other parties can process these data in a different ways. At present, the first translation is to Classical Latin. If required, other presentations can be added to LED.

Aligning several sources and missing epitomes

Sources must be of the same work. For example, aligning the two incomplete sources of El Libro: the clean Copenhagen [LE-C] and the draft Seville [LE-S]. Some epitomes are just in one of them, some in both though they might differ. Metadata from missing epitomes in both sources is also considered to reconstruct as much as possible of El Libro. The Lagarto Format can deal with several sources. Example:

```
id:0
title:Lorem ipsum
%%
t0:Lorem ipsum   [consolidated texts of t1 and t2]
t1:Lorem ipsum   [text from LE-C]
t2:Lorem ipsum   [text from LE-S]
la:Lorem ipsum   [translation from t0]
es:Lorem ipsum   [translation from t0]
%%
```

Translation must be done from `t0`. If each transliteration segment requires different translations, notes, etc; a more complex data structure must be used. For example, as in the *Diálogo de la Lengua* [DILE].

List of all epitomes

A dashboard {led/mix/global.html} of all epitomes in LE-C, LE-S or missing from both manuscripts. It contains links to many sources such as: Catálogo Concordado [CC], Material Evidence in Incunabula [MEI], Universal Short Title Catalogue [USTC], author(s), translator(s), direct links to the relevant facsimile and notes. The input data is the `head` element. This ongoing work is very much active and the result should vertebrate the transcription proper.

Complutensian Polyglot Bible

El Libro is monolingual, though epitomes might be in different languages. A far more complex task would be the paleographic transcription of the Complutensian Polyglot Bible, focusing on the three parallel texts: Hebrew, Latin Vulgate and Greek Septuagint; the Bible contains other texts. Even a trial run with a few pages could be considered. The main focus would be analysis of this printed work, that might or not be used for biblical studies.

From a purely transcription perspective it might be easier as it is printed as opposed to hand written; one might consider a first pass with AI image recognition. From an alignment perspective far harder; though based on the ADT, a new abstract data type would probably be required, particularly if segments in each language are not aligned (translation of each other) and they require different translation, though they might be consolidated later.

Acknowledgement and intellectual property rights

Catálogo Concordado (see Aviso Legal). Facsimile of the original manuscript, Copenhagen, Arnamagnæan Collection, AM 377 fol. Photo: Suzanne Reitz. Used with permission; hosted at Handrit.is. From AM02-0377-en.xml was taken and modified the elements/attributes msItem (n), title, locus (from to). Also consult the sources/author(s)/owner(s) for complementary and/or more detailed intellectual property rights.

Author

Manuel Tomas CARRASCO BENITEZ
ca@dragoman.org

References

[BB] Book of Books project
https://bookofbooks.ku.dk

[BE] Hernando Colon's Book of Epitomes
https://www.english.cam.ac.uk/research/hernandocolon

[BEE] Hernando Colon's Book of Epitomes - Edition
https://www.english.cam.ac.uk/research/hernandocolon/edition

[CC] Catálogo Concordado
https://icolombina.es/catalogo-concordado

[CSS] Cascading Style Sheets
https://www.w3.org/Style/CSS

[DILE] Diálogo de la Lengua
http://dile.dragoman.org/tab.html

[DK] Docker
https://www.docker.com

[DP] A System for Long-Term Document Preservation
http://larry.masinter.net/0603-archiving.pdf

[DSV] Delimiter-separated values
https://en.wikipedia.org/wiki/Delimiter-separated_values

[DTD] document type definition
https://en.wikipedia.org/wiki/Document_type_definition

[EIEF] Everything is a file
https://en.wikipedia.org/wiki/Everything_is_a_file

[FOSS] Free and open-source software
https://www.gnu.org/philosophy/free-sw.html

[HAC] Hackathon
https://en.wikipedia.org/wiki/Hackathon

[IC] Institución Colombina
https://icolombina.es

[KV] Key-value pairs
https://en.wikipedia.org/wiki/Attribute%E2%80%93value_pair

[LE] Libro de los Epítomes
https://en.wikipedia.org/wiki/Libro_de_los_Ep%C3%ADtomes

[LE-C] AM 377 fol., Arnamagnæan Collection (Copenhagen)
http://lagarto.top/rep#5
https://handrit.is/manuscript/view/en/AM02-0377/0#mode/2up

[LE-S] Institución Colombina (Seville)
http://lagarto.top/rep#4

[LOM] Hernando Colón y la Biblioteca Colombina
https://ultreia.ucv.es/index.php/ultreia/catalog/book/19

[LW] Lagarto wiki
http://wiki.lagarto.top
Example for epitome 142:
https://wiki.lagarto.top/index.php/142

[MEI] Material Evidence in Incunabula
https://data.cerl.org/mei

[R4810] Long-Term Archive Service Requirements
https://tools.ietf.org/html/rfc4810

[RJ] Record-Jar Format
http://www.catb.org/esr/writings/taoup/html/ch05s02.html#id2906931

[TAOUP] The Art of Unix Programming
http://www.catb.org/esr/writings/taoup/html/index.html

[TB] Transcribe Bentham
http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham

[TE] The Importance of Being Textual
http://www.catb.org/esr/writings/taoup/html/ch05s01.html

[USTC] Universal Short Title Catalogue
https://www.ustc.ac.uk

[XDOSSIER] Xdossier
http://dragoman.org/xdossier

[XML] Extensible Markup Language
https://www.w3.org/TR/2008/REC-xml-20081126