Analyzing poetry databases to develop a metadata application profile. Why each language uses a different way of modelling?

Patricia Garrido, LINHD-UNED and UCM, pgarrido@linhd.uned.es

Introduction

This lightning talk is a description of a work-in-progress which explains my collaboration in the POSTDATA[1] project where I am working as a student in practices, contributing to the Project with my knowledge in philology, and learning how to use DH tools and methodologies to analyze traditional philological problems.

POSTDATA project and its process

My contribution to this project belongs to the first of its work packages: "semantic web and ontology development", which deals with the development of a metadata application profile for poetry. It is reverse engineering process, as we analyze the logical models of different databases and create particular conceptual models in order to create a final and common conceptual model to all the existing ones. For the accomplishment of this work, a classification of the different databases has been made taking into account the language in which poetry is written. At the moment, I am working in specific repertoires and databases devoted to Latin poetry from different provenances and universities: *Pedecerto*, the *Corpus Rhytmorum Musicum Analecta Hymnica Digitalia and Analecta carminum medii aevi*, and comparing them with other repertoires of German, English, French, Spanish and Portuguese poetry.

First, it is necessary to analyze the logical model of each database in order to understand the concepts that are represented by each table making a description of the different terms that were chosen by the designers. An example of this procedure can be well explained using *Pedecerto*[2] as a case study, a digital instrument for the analysis of Latin verses. It is a repertoire which is composed by two different databases, sending user information from both of them. For example, the word "sistema" appears in the model without any contextualization and it becomes difficult to interpret it. For that reason, it is necessary to go back to the website and look for disambiguation. In the case of "sistema", the conclusion is that this term describes "the type of behavior in the metric system" (If there is a "D" it is a dactylic system; if "E" it is an elegiac couplet; if "N" we have hexameter and pentameter meters mixed with other kind of meter...)

A similar phenomenon happens to the *Corpus Rhytmorum Musicum*[3], which is a musical and textual philological database of the earliest Medieval Latin Songs. This one is more related to music and manuscripts, so I find terms such as "NRMano" and exploring the website as I have explained below, I can describe the term as the "number of hands which have written a determinate manuscript".

It is necessary to build an abstract model in which the terms used for describing general concepts, such as "manuscript", "poem" or "literary work" have identical or very similar meaning across the different databases.

There is a second phase in this process, which consists of the analysis and grouping of the controlled vocabularies from each different literary tradition, which are collected by the search tools

1 The POSTDATA project: http://postdata.linhd.es/

2 The Pedecerto repertoire, supported by the University of Udine, has its own website: http://www.pedecerto.eu/

3 The The Corpus Rhytmorum Musicum is supported by the University of Siena in Arezo and its website is: http://www.corimu.unisi.it/

of the repertories. The study of controlled vocabularies can be focused from different perspectives, but we first classify the term looking later for groups and hyperonyms. The execution of this task is very positive for the review of the previous one, since in the logical entity we find terms that refer to controlled vocabularies and must not appear in the conceptual model. As many databases do not show a regular work on controlled vocabularies, it is sometimes not easy to identify and extract their terms and keywords. In this sense, *ReMetCa* Project is a repertoire of special relevance, as it has developed a great effort to study controlled vocabularies using external tools, as Tematres.

So, this Lightning Talk will describe all these methods to compare and analyze poetry databases, but also will reflect on the idiosyncrasy of classifying poetry and the differences of conceptualization among the different languages, literatures and traditions and its representation in the digital world.

References

González-Blanco García Elena and Rodríguez Gómez, José Luis, "ReMetCa, an integration proposal of MySQL and TEI-Verse"." Issue 8 del Journal of the Text Encoding Initiative (2015)
González-Blanco García, Elena, del Rio Riande, Gimena, and Martínez Cantón, "Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires", LREC 2016 Proceedings (2016)
González-Blanco García, Elena, "Un nuevo camino hacia las Humanidades Digitales: El Laboratorio de Innovación en Humanidades Digitales de la UNED (LINHD)", Signa, Revista de la Asociación Española de Semiótica, 25 (2016): 79-93.

Repertoires and projects

Corimu: http://www.corimu.unisi.it/
POSTDATA: http://postdata.linhd.es/
Pedecerto: http://www.pedecerto.eu/
ReMetCa: http://www.remetca.uned.es/index.php?lang=es
Analecta Hymnica Digitalia and Analecta carmine medii aevii: http://webserver.erwin-rauner.de/crophius/Analecta_conspectus.htm