From syllables, lines and stanzas to linked open data: standardization, interoperability and multilingual challenges for digital humanities

Elena González-Blanco García
LINHD, National Distance Education University (UNED)
Paseo Senda del Rey, 7
Madrid, 28040, Spain
0034 91 3986873
egonzalezblanco@flog.uned.es

Mara Manailescu
LINHD, National Distance Education University (UNED)
Paseo Senda del Rey, 7
Madrid, 28040, Spain
0034 91 3987808
mmanailescu@linhd.uned.es

Salvador Ros
School of Computer Science, National Distance Education University
Calle Juan del Rosal,16
Madrid, 28040, Spain
0034 91 3987205
sros@scc.uned.es

ABSTRACT

This proposal presents the challenges and first results of POSTDATA ERC Starting Grant project, which aims at bridging the digital gap among traditional poetry collections and the growing world of data. It is focused on poetry analysis, classification and publication, applying Digital Humanities methods of academic analysis in order to look for standardization. The context of the project is the corpora of European poetry, with a special focus on poetic materials from different languages and literary traditions. Interoperability problems between the different poetry collections are solved by using semantic web technologies to link and publish literary datasets in a structured way in the linked data cloud.

This paper will present the current situation in the field of digital humanities analyzing poetry as the "study case" and the application of different technologies used in the field of digital humanities to provide new and innovative results. It will also introduce LINDH, the Digital Humanities Innovation Lab at UNED, a pioneer Digital Humanities center in Spain and its role as a facilitator of different technologies to be applied to the study of traditional humanistic problems with the most updates technologies in the field.

Keywords

Digital humanities, Linked Open Data, TEI-XML, poetry, semantic web

1. INTRODUCTION

Digital Humanities are experiencing a growing interest in Spain, with important results, especially during the last five years in many different field, but especially linked to philological traditions, such as linguistics and literature. This paper presents a clear example of this transformation of a traditional humanistic discipline into a new technological world with poetry as the central problem to be analyzed. For this reason, this paper is focused on a real case of a project: POSTDATA, an ERC Starting Grant project developed at LINHD, the Digital Innovation Lab at UNED and the first Digital Humanites Centre in Spain. It is a groundbreaking project with ambitious targets which seeks to unify poetical analysis between cultures and existing research databases by using web semantic technologies. POSTDATA aims to offer a standardized model in the philological field and a metadata application profile (MAP) for poetry in order to build a common classification of all these poetic materials. The final goal of the POSTDATA project is: i) to be able to publish all the data that is now locked in LOD, where any agent interested will be able to build applications over the data in order to serve final users; ii) to build a Web platform where: a) researchers, students and other final users interested in poetry will be able to access poems (and their analyses) of all databases; b) researchers, students and other final users will be able to upload poems, the digitalized images of manuscripts, and fill in the information concerning the analysis of the poem, collaboratively contributing to a LOD dataset of poetry (González-Blanco, 2016).

2. STATE OF THE ART AND OUR APPROACH

2.1 Current situation

The need for information exchange has made it necessary to create international standards in most fields. This phenomenon has been especially important for scientific disciplines, with widely adopted standards or even regulations for data sharing. In this sense, humanities have followed a more independent way of evolution, as there are important factors, such as history, creativity or self-identity, which influence each particular tradition with different results.

POSTDATA project aims at shortening the digital gap among poetry and technology, looking for interoperability solutions and an interdisciplinary work with innovative results beyond the current state-of-the art. It is based on the application of semantic web technologies to link and publish poetic datasets in a structured way in the linked data cloud. Making poetry available online as machine-readable data opens a great world of possibilities of linking, indexing and extracting new information through the combination of the different datasets. The addition of a "semantic" layer of data to existing different databases and digital resources devoted to poetry boosts interoperability among them and lets scholars develop innovative comparative studies, which were not possible to carry out before.

As this project is conceived from an interdisciplinary point of view, it is structured in three different dimensions to reflect the need of building bridges between the disciplines, users and methodologies: a) academic and philological, b) technological, and c) social and user-oriented. The state of the art, objectives and methodology of the project will be explained following this triple structure.

2.2 Technological standardization

The lack of unified criteria is translated into many different uncoordinated technologies when research data are transformed to build digital projects and do not even follow a standard, in most cases. From the technological point of view, the variety of projects and solutions used is so great, that it is not easy to exchange information among them. This is also due to the fact that the current technological state of the art reflects the changes experienced by digital humanities themselves in the last years.

The multiplicity of technologies used includes SQL databases, TEI and XML markup languages, semantic web technology standards (RDF, OWL, SKOS), natural language processing systems (NLP) and visualization tools. Relational databases have been deeply used by the first digital poetic repertoires combining an ER (Entity-Relationship) model, together with the data model based on records for the logical implementation (Elmarsi, 2011). The problem of representing ER composition model is that the result shown is data centred, but it is not enough to mark textual items that need to be analyzed from a metrical point of view. There are other projects based on XML solutions, as TEI has a specific module for poetry analysis, "Verse", with a rich set of tags to describe metrical schemas, rhymes, accentual structure and syllabic varieties. However, this model is not widely used by the different projects, and the lack of philological unified criteria makes it difficult to translate literary schemas into XML tags, making researchers create new tags or express nuances with customized attributes for each project.

The key for interoperability both in philological and technological fields is a common reference system, for which semantic web technologies are a powerful solution. Building a linked data model by adding a semantic layer of metadata to the existing databases does not alter their internal structure. This solution requires, however, to assume unified criteria on the philological model that serves as a reference.

Although semantic web technologies have had success in archives libraries and museums (group known as LODLAM http://lodlam.net/ (http://lodlam.net/)), its application to poetic corpora is very different, as there are only a couple of studies dealing with some of the above mentioned aspects ((Bootz, 2008) and (Zöllner-Weber, 2009)), but there is not a standard conceptual model of ontology referred to metrics and poetry. The closest works related to this topic are probably the conceptual model of CIDOC, the controlled vocabularies of English Broadside Ballad Project (Fumerton) and the linked data relations offered by the Library of Congress (http://id.loc.gov), which do not offer enough information on metrics vocabulary. There are also interesting computational approaches which use automated linguistic analysis or text mining, based on the morphological and phonetic structure of each language. Results have been impressive, as one of the greatest advantages is the speed of the analysis of big amounts of text (Gervás, 2000). Nevertheless, the integration of these technologies with the previous models described is not easy, and solutions are often customized for the variation of natural language used, most times standard English.

From the technological point of view, the main objective of POSTDATA will be to translate the philological standardization described into existing digital humanities standards in order to exchange information (mainly TEI-XML for tagging, SQL for structuring, and OWL for publishing datasets as Linked Open Data). The secondary objectives of this pillar will be:

- To study and revise the digital standards which have been created or applied to poetry: TEI-XML verse module, some Dublin Core and Cidoc-CRM elements, FRBRoo, and Europeana Data Model, among others.

- To design a model to analyze and represent the concepts described in the philological conceptualization. ReMetCa structure will be taken as a prototype, thanks to its combination of an the ER (Entity-Relationship) data model, translated into the creation of SQL databases for structuring contents and combined with the design of TEI-XML schemas for poetry encoding.

- To build a common ontology using semantic web technologies and W3C standards (such as OWL) and publish the metadata extracted from the philological conceptualization as linked open data (LOD), ready to be shared, linked and improved by the community of users. This ontology will have as a starting point the ReMetCa ontology that I created with my research group for the Spanish project.

- To develop a framework to link data with external existing datasets, such as the datahubs of the Spanish or French National Libraries, the British Museum LOD portal, or the Getty Vocabularies.

2.3 Philological standardization

During the Middle Ages and the Renaissance, the powerful influence of Latin made scholars inherit the terminology of Classical poetry treatises and apply it to Romance languages, regardless of their different linguistic traits and verse structures. When vernacular theories started to arise, each literary school set up its own terminology and classification system. This multiplicity led to complex situations, such as the creation of conceptual genres that only exist in some traditions.

Later, the musical analysis system applied by (Navarro Tomás, 1956) was followed as a valid system through many years, using concepts like anacrusis. In the last years, there have been many different approaches to explain the Spanish panorama, as it is shown by the semantic comparative model designed by the Czech (Bělič, 2000).

The international context is richer, especially in English, with two prominent schools: 1) A traditional approach based on stress and classical feet; and 2) A generative approach based on the terminology and concepts shown through text grids that take into account word boundaries, with a strong impact on poetic theories (Gerber, 2013)

The models described are just an example of the idiosyncrasy that can be observed in each literary tradition. Although the current ICT infrastructures are prepared to harvest different types of collections and models, it is necessary at a first stage to standardize metadata and map vocabularies and terminology at the philological level in order to build a consistent able to be shared between the different traditions.

From the philological point of view, the main objective of POSTDATA proposal is to develop a conceptual standardization to

build an abstract model for poetry representation based on existing philological concepts taken from projects, handbooks and corpora from the different traditions. This objective will be divided into three secondary objectives:

- To carry out a comparative analysis of digital projects and repertoires from different poetic traditions to extract the main conceptual elements and metadata shared by most of them (such as author, accent, syllable, caesura).
- To analyze the evolution of manuals and academic studies on poetry in order to understand the evolution of theories and to choose the most adequate concepts for its digital representation. This analysis will require to compare the early grammars of Medieval Latin and early Romance literatures, but also rhetoric books and stylistic treatises in order to discover how the discipline has been reshaped along the years.
- To create an abstract model of representation with the essential common elements extracted in the first two objectives. This model will also represent their properties, attributes and relationships in a logical, a conceptual and a material level, distinguishing between common abstract metadata (shared by all the traditions, like "author" or "poem"), and particular controlled vocabularies (such as the names of stanzas in each tradition).
- To work on Spanish, French and Italian corpora and bibliography to extract information and to test the technological applications developed.

2.4 Social and user-oriented state of the art

The third pillar of this project is focused on the creation of a virtual research environment for poetry scholarly editing oriented to different kind of users: researchers with academic purposes who want to work on critical digital editions, non- experienced uses that want to read, share and learn more about poetic traditions and also companies who will use this resource for different application in fields like education, psychology, tourism or cultural purposes. One of the most important aspects of this virtual environment is that the previously described technologies need to be hidden to most users, as one of the keys of the platform's success is accessibility and the majority of users do not have technical computing abilities.

The platform will be built by integrating previously existing tools that have been developed by other research teams at previous projects. Innovation lies in the application context for this combination of tools, which specifically oriented to poetry analysis. Such a specialized platform represents a very innovative concept and would be very difficult to build from scratch. However, the advances made in related areas, such as digital editions will save us time and effort and get challenging results when applying them to poetry.

From the social and user-oriented point of view, the main target of POSTDATA is be to build a user-friendly environment to manage textual editions and documentary collections using existing standards based on XML, together with other advanced programming languages, query languages and database tools, suitable for different kinds of users (academic, professional and social) without high computing abilities. The secondary objectives of this pillar will be:

- To create a resource that let user make his own choices following a three step process: 1. a selection of text encoding options based on TEI-XML tags to build DTDs or schemas, 2. A complex query system, like XQuery and 3. A UX user interface to visualize and publish the text edition in different formats with search possibilities, using XSLT and other technologies (such as Javascript, PHP and CSS) to boost visualization possibilities, social sharing options and downloadable and exportable texts.

- To design a publication environment suitable for different types of users: academic users, social users (users with not academic purposes, but literary interests), and professional users (like teachers, musicians, tourist operators, Apps developers, etc.).

- To train academic researchers and professional users in digital humanities tools and possibilities, by offering, together with the environment, webinars, tutorials and learning guides and materials on the technologies used.

- To embed all the content available in this virtual environment into a linked-data framework to let users perform faceted searches with different levels of complexity, based on SPARQL queries, with multiple visualization possibilities.

2.5 Activities and results

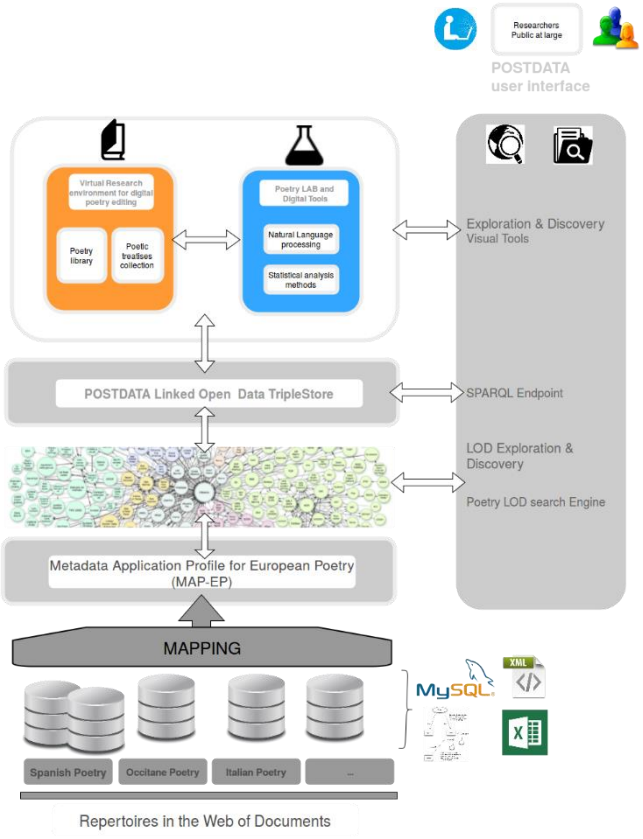

Figure 1. POSTDATA Organization Scheme

The implementation of the POSTDATA architecture is structured into four stages:

State-of-the-art Repertoires

The first stage of the project has consisted of analyzing the existing repertoires and databases devoted to poetry analysis in the different countries, literary traditions and languages. They contain metadata about the works and manuscripts, but especially data about content

and metrical features. At the current moment 21 data models have been analyzed by structuring their data models and transforming them into logical structures. Detecting common shared fields and problematic structural concept like "genre" or "topic" have been crucial to understand the difficulty of the models. A first prototype using SQL has been build. Controlled vocabularies inside each digital resource and database have been taken apart and studied in an independent way in order to look for translations and similarities with similar words in other languages.

Ontology and Semantic Web

To define a Metadata Application Profile (or Ontology) for the Poetic Community. This will help define a Metadata Application Profile (others call it Ontology) to be used by the Poetic Community. This definition contains: a Data Model and a Semantic Model, a specific RDF Vocabulary and several Vocabulary Encoding Schemes to be used by the Semantic Model. By providing a MAP, the Poetic Community will be able to publish data in the Linked Open Data ecosystem and be interoperable with many other communities (e.g. Arts libraries, geographic resources and others to be established yet). We will use Me4MAP (Curado Malta, 2013), a method for the development of Metadata Application Profiles.

Services Development

This stage of the project will support the definition and development of services (functionalities) to be used in the Platform. Although this is yet a future task in the project, a first approach on the possibilities of tools and software developments to be implemented has been made. This work includes a detailed description of the services, the roles of the users, a list per role of the services the role is allowed to access and finally the services software packages.

Integration and visualization

The final product of POSTDATA project is conceived as a easy to use interface, which contains a LOD powerful search engine, a digital edition platform and a poetry lab to play with apps and tools. A visualization layer will be included at the end with different visualization possibilities: geolocalization, timelines, graphs.

3. CONCLUSIONS

At the present moment, many Digital Humanities and culture-based projects shown shared problems, and they offer similar or different unstandardized solutions, but most of them are not interoperable, as bridges are needed to communicate the existing models at different levels, as we have already seen: first, at the conceptualization level, as it is necessary to map and translate philological concepts to make information exchangeable, and second, at a technological level, as there is no standard solution to deal with poetical problems and depending on the focus and aim of the projects, results are proven to be different. However, following some protocols, like using open data formats, APIs, exchange languages and standardized metadata it is more probably to be on the way towards getting interoperability for poetical projects.

POSTDATA does not propose a new method for analyzing poetry, but an abstract model based on a working methodology supported by a double standardization system, both at philological and technological levels. The advantage of using semantic web technologies as an interoperability solution is that they do not alter the internal structure of the existing projects and add to them a semantic layer expressed in RDF which makes able to identify and link each object, expressed as URI identifiers, making it searchable and linkable through the web. Semantic content are easily joinable to database systems, and they have especially been used to deal with library collections, due to their need of exchanging information, but also with museums, as the CIDOC model shows, and are being explored in relationship with TEI projects and with other important digital humanities textual projects, such as (Perseus), (Pelagios) or (Biblissima).

Our proposal aims to set up a procedure to combine philological criteria to map vocabularies and concepts which might have common means and properties in the different traditions and to insert them into an abstract framework in which each of these elements can fit as individuals of an ontology which gathers the main poetics concepts shared by most traditions.

4. ACKNOWLEDGMENTS

This paper has been developed thanks to the research projects funded by MINECO and led by Elena González-Blanco: Acción Europa Investiga EUIN2013-50630: Repertorio Digital de Poesía Europea (DIREPO) and FFI2014-57961-R. Laboratorio de Innovación en Humanidades Digitales: Edición Digital, Datos Enlazados y Entorno Virtual de Investigación para el trabajo en humanidades, and the Starting Grant research project: Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528), funded by European Research Council (ERC) under the European Union´s Horizon 2020 research and innovation programme, (http://postdata.linhd.es/).

5. ADDITIONAL AUTHORS

Aditional authors: Mariana Curado Malta (POSTDATA Team, mariana.malta@linhd.uned.es), Gimena del Rio Riande (SECRIT-IIBICRIT CONICET, Argentina, gdelrio@conicet.gov.ar), Clara I. Martínez Cantón (cimartinez@flog.uned.es).

6. REFERENCES

Bělič, O. (2000). *Verso español y verso europeo: introducción a la teoría del verso español en el contexto europeo, .* Santafé de Bogotá, Instituto Caro y Cuervo.

Biblissima. (n.d.). *http://www.biblissima-condorcet.fr/en.*

Bootz, P. &. (2008). *"Towards an ontology of the field of digital poetry", .* Paper presented at Electronic Literature in Europe, Full text available at http://elmcip.net/node/415.

Curado Malta, M. &. (2013). *A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2).*

Elmarsi, R. &. (2011). *Fundamentos de Sistemas de Bases de Datos,.* Madrid, Pearson, Addison Wesley.

Fumerton, P. (n.d.). *English Broadside Ballad, http://ebba.english.ucsb.edu/.*

Gerber, N. (2013). *Stress-Based Metrics Revisited: A Comparative Exercise in Scansion Systems and their Implications for Iambic Pentameter. Thinking Verse III,.*

Gervás, P. (2000). *"WASP: Evaluation of Different Strategies for the Automatic Generation of Spanish verse",*. Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science, University of Birmingham, .

González-Blanco, E. d. (2016). *Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires.* "LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources".

http://id.loc.gov/ . (n.d.).

Navarro Tomás, T. (1956). *Métrica española, Barcelona, Labor*.

Pelagios. (n.d.). *http://commons.pelagios.org/*.

Perseus. (n.d.). *http://www.perseus.tufts.edu/hopper/*.

Zöllner-Weber, A. (2009). *"Ontologies and Logic Reasoning as Tools in Humanities?", Digital Humanities.*