A Common Conceptual Model for the Study of Poetry in the Digital Humanities

Mariana Curado Malta
mariana@iscap.ipp.pt
Instituto Politécnico do Porto, Portugal

Elena González-Blanco
egonzalezblanco@flog.uned.es
Universidad Nacional de Educación a Distancia, Spain

Clara Martínez Cantón
cimartinez@flog.uned.es
Universidad Nacional de Educación a Distancia, Spain

Gimena Del Rio
gdelrio.riande@gmail.com
CONICET, Argentina

Many of the DH approaches to poetry have focused on its metrical aspect (Birnbaum & Thorsen, 2015a, 2015b; Pue, Teal & Brown, 2015; González-Blanco, Martínez Cantón & Rio Riande, 2015). Following Chatman & Levin (1967, p. 142) we could say that meter is a "systematic literary convention whereby certain aspects of the phonology are organized for aesthetic purposes." In this sense, versification is an abstraction of linguistic phenomena in which words (in their formal and semantical aspect) relate to rhythm and rhyme for artistic purposes. Although many theories about versification and metrics have been developed for different languages and traditions, our work is interested in a structural and formal approach that looks at poetry into discrete units, categories, and their relationships. Thus, we are involved in analysing how metrical repertoires in digital form model those structures.

A digital repertoire of poetry metrics is a catalogue that gives account of the metrical and rhythmical schemes of either a poetical tradition, a period or school, gathering a corpus of poems that are defined and classified by their main characteristics. This kind of repertoires may sometimes contain the text of the poem and information related to authors, manuscripts, editions, music, and other features, all of them related to the poems. In the beginning, repertoires were printed books in which we could find information listed in a way similar to an address book. The digital era has changed the way in which information is displayed allowing the user to perform complex and multiple searches. In all these cases there is an ontological leap when the data is put in digital format.

There are a vast number of European digital metrical repertoires available online in open access – e.g. French lyrical collections (Nouveau Naetebus), Italian (BedT), Hungarian (RPHA), Medieval Latin (Corpus Rhythmorum Musicum, Annalecta Hymnica Digitalia), Classical Latin (Pedecerto), Galician-portuguese (Oxford Cantigas de Santa María, MedDB2), Castilian (ReMetCa), Dutch (Dutch Song Database), Occitan (BedT, Poèsie Neotroubadouresque, The last song of the Troubadours), Catalan (Repertori d'obres en vers), Skaldic (Skaldic Project), or German (Lyrik des Minnesänger), among many others. Each one of these metrical repertoires was developed in a specific technology and stores data in its own database. This data is locked in the information silos of each repertoire, not available freely to be compared and used by intelligent machines that could infer over the data. The lack of interoperability between the different digital repertoires dealing with poetry across different languages, literatures and traditions is a problem that needs to be addressed (González-Blanco & Seláf, 2014; González-Blanco, Rio Riande, Martínez Cantón & Martos Pérez, 2014; González-Blanco, E. G., Rio Riande, G.del & Martínez Cantón, C., 2016a). There is an interoperability problem since the technological solutions used for building the database of each digital repertoire employ a different data model.

POSTDATA is a project funded by a Starting Grant of the European Research Council whose main goal is to study European poetry in a comparative and wide context. Its starting point is the analysis of digital metrical repertoires as digital objects that have been developed to expose European poetry on the Web of Documents in open access. POSTDATA is an idea that resulted from previous work developed by a network of partners that own digital metrical repertoires and that have been working together for some years.

The POSTDATA project focuses mainly on the possibilities of Linked Open Data (LOD) technologies in order to publish the data available on information silos in the Web of Data as LOD. A side effect of this project is that not only this network will be able to publish its own data in LOD but also any other entity that owns data related to poetic metrics. Once the data is structured the same

way and published in the Web of Data, anyone will be able to deploy new software that uses this data. It will be possible to compare, extrapolate and create new views of the data, opening doors to new knowledge in i) literary research communities since POSTDATA results will enable comparisons of poetic traditions and the creation of new repertoires, and ii) literature learning environments for e.g. the development of editorial projects following the common conceptual model.

LOD technologies are very powerful since one can implement intelligent agents to infer over the data (W3C, 2010). Another advantage of LOD over other paradigms is the infinite possibilities of linking data, allowing agents to rely also on other sources of information to build knowledge.

The first step of POSTDATA is the development of a Metadata Application Profile (MAP) for the community of practice of Digital Humanities that deal with poetry. An MAP is a semantic model, a construct that enhances interoperability (Nilsson, Baker, & Johnston, 2008).

The development of an MAP is a crucial task for any community of practice. This development should be structured, and has to integrate –since the early phases of development– elements of representative members of the community. Commonly the organizations of a community of practice differ in organization-type, location, culture and in the language they speak. In poetry studies, all these are key factors, since metrics and their features are based largely on the nature of the languages. In addition, in the case of the Digital Humanities Poetry metrics community, the way metrics has been conceptualised in the different traditions is however diverse, as there are multiple ways of encoding and understanding metrical systems. Finding a common ground of understanding in such an environment becomes a huge challenge. These digital metrical repertoires are the key to build an inclusive MAP for poetry, since they reflect different approaches of conceptualising poetry and their formal features.

POSTDATA is using Me4MAP, a method for the development of metadata application profiles (see Curado Malta & Baptista 2013a, 2013b), to develop the referred MAP. According to Me4MAP, the first activities in such a development are "S1 – Developing the Functional Requirements" and "S2 – Developing the Domain Model," these two activities result, respectively, in the Functional Requirements deliverable and in the Domain Model deliverable. S1 activity feeds S2 activity.

A Domain Model is a conceptual model that is used to explicit the concepts that exist in a certain universe of discourse. The concepts have properties to show how they are defined and related.

As described in Curado Malta, Centenera & Gonzalez-Blanco (2017), the work-team is developing the Domain Model at the same time S1 activities occur, slightly changing the order of S1 and S2. This change is possible since Me4MAP is also under evaluation and testing in a Design Science Research methodological approach –for more information about this process see Curado Malta (2014). This Me4MAP testing is leaded by one of the authors.

The goal of this paper is to present a preliminary version of the Domain Model of the MAP for Poetry – Domain Model V0.1. The following paragraphs give details on how this version of the Domain Model was developed.

The Domain model V0.1 was developed based on the analysis of eleven data models of databases of metrical repertoires (see Table 1) using a reverse engineering technique –for more details about this technique see Curado Malta et al. (2017).

Project	URL
Base de Datos da Lírica profana galego-portuguesa (MedDB)	https://www.cirp.gal/meddb
Cantigas de Santa Maria	http://csm.mml.ox.ac.uk/
Corpus of Spanish Golden-Age Sonnets	https://github.com/bncolorado/CorpusSonetosSigloDeOro
Corpus Rhythmorum Musicum	http://www.corimu.unisi.it
Kalevala	http://dbgw.finlit.fi/skvr/
Lyrik des hohen Mittelalters	http://www.lhm-online.de
Métrique en Ligne	http://www.crisco.unicaen.fr/verlaine/
Repertorio Métrico Digital de la Poesía Medieval Castellana (ReMetCa)	http://www.remetca.uned.es
Répertoire métrique de la poésie lyrique occitane des troubadours à leurs héritiers (xiiie-xve siècles)	Local Database
Reperrtoire de la Poésie Hangroise Ancienne (RPHA)	http://rpha.elte.hu/
Versologie	http://metro.ucl.cas.cz/kveta

Table 1. Repertoires which databases were used as basis for Domain Model V0.1

These repertoires have different technologies and paradigms: ten are in the Web of Documents and one is deployed locally. The technologies in which the databases are implemented are: 1) MySQL databases, 2) XML databases and files, 3) Perl scripts and, 4) Worksheet files.

The Domain Model of the MAP is represented as an Unified Modeling Language (UML) class diagram that expresses POSTDATA proposal for a common conceptual model for the European Poetry. The Domain Model V0.1 is available at the permanent link https://doi.org/10.5281/zenodo.437827. The technique used to represent the domain model is UML since it is a standard technique (ISO/IEC 19505-1 and 19505-2) to model businesses and processes very well known in the software engineering community –see Rumbaugh, Jacobson and Booch (2004).

This first version will be tested in a Workshop held in March where partners of the project will follow a hands-on session to test the Domain Model V0.1 using a resource of its own database. POSTDATA will give to each partner a template testing sheet and information about the mapping between the database in question and the Domain Model. The conclusions of the testing and of other workshop discussions will feed the process of development of the Domain Model. The work-team programs to deliver a version 0.2 out of this worshop conclusions. This process of development is highly iterative; it will be fed by the following activities:

- development of S1 activities (as already referred before);
- Analysis of the results of a survey to final users: POSTDATA wants to know the needs of final users of repertoires in order to understand their problems and how they can be solved. Part of the survey is created based on the existent models. The survey also has open questions for users to fill in freely.
- Analysis of other poetic repertoires: there are still at least eleven databases' data models to be analysed. POSTDATA is still looking for more poetic repertoires in order to have a wider representation of other languages and traditions (plewease see the geographical view of the repertoires) Analysis of the information of two case studies: two researchers are building a digital repertoire at the same time the MAP is being developed; they will inform the work-team about their needs.

Once a stable version of the Domain Model is available, POSTDATA will follow the activities described by Me4MAP. One important activity is the mapping of the Domain Model to a semantic model, where every property of the Domain Model will be matched by a RDF vocabulary term, other constraints such as cardinality, domain and range must also be defined. In order to enhance interoperability, this matching has to be done taing care that the most popular vocabularies of the LOD ecosystem are used. A study of the most used vocabularies will be performed so as to find the best options for each term. Examples of RDF vocabularies that might be used are Dublin Core Metadata Terms, Digitised Manuscripts to Europeana (DM2E), Friend of a Friend (FOAF) vocabularies, among others. This semantic model will be tested in an iterative process of development, using resources from metrical repertoires that were not used as cases in the metadata application profile process development.

Bibliography

Birnbaum, D. J., & Thorsen, E. (2015a). Markup and meter: Using XML tools to teach a computer to think about versification. In *Balisage: The Markup Conference*.

Birnbaum, D. J. & Thorsen, E. (2015b). Enabling the automated identification and analysis of meter and rhyme in Russian verse. In *DH 2015: Global Digital Humanities*, Sydney.

Bosch, Mela & Rio Riande, Gimena del (2016). Las Humanidades Digitales o el ornitorrinco". In *Las Humanidades Digitales en Argentina: discursos y tecnologías en cruce*. Retrieved from: http://www.centrocultural.coop/blogs/utopia/2016/11/01/cronica-de-la-actividad-las-humanidades-digitales-en-argentina/ – accessed 25 March, 2017.

Chatman, S. & Levin, S. R. (1967). *Essays on the language of literature*. Boston: Houghton Mifflin.

Curado Malta, M., & Baptista, A. A. (2013a). A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2): detailed description. In M. Foulonneau & K. Eckert (Eds.), *Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013* (pp. 90–103). Lisbon: Dublin Core Metadata Initiative. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/178/81 – accessed 23 March, 2017.

Curado Malta, M., & Baptista, A. A. (2013b). Me4DCAP V0. 1: A method for the development of Dublin Core Application Profiles. In *Information Services and Use* (Vol. 33, pp. 161–171). Dublin Core Metadata Initiative. Retrieved from http://doi.org/10.3233/ISU-130706 – accessed 23 March, 2017.

Curado Malta, M., & Baptista, A. A. (2013c). State of the Art on Methodologies for the Development of Dublin Core Application Profiles. *International Journal of Metadata, Semantics and Ontologies*, *8*(4), 332–341. Retrieved from http://doi.org/http://dx.doi.org/10.1504/IJMSO.2013.058416 - accessed 23 March, 2017.

Curado Malta, M. (2014). *Contributo metodológico para o desenvolvimento de perfis de aplicação no contexto da Web semântica*. Universidade do Minho, Escola de Engenharia. Programa Doutoral em Tecnologias e Sistemas de Informação. Retrieved from http://hdl.handle.net/10171/35718 – accessed 23 March, 2017.

Curado Malta, M., Centenera, P. & Gonzalez-Blanco, E. (2017). Using Reverse Engineering to define a Domain Model: The case of Development of a Metadata Application Profile for the European Poetry.

In Curado Malta, M., Baptista, A. A. and Walk, P. (Eds.), *Developing Metadata Application Profiles* (pp. 146-180). Hershey PA: IGI Global. DOI: 10.4018/978-1-5225-2221-8.ch007.

González-Blanco García E. & Seláf, L. (2014). Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires". In L. Soriano, M. Coderch, H. Rovira, G. Sabaté & X. Espluga (Eds.), *Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world* (pp.321-322). Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien: Peter Lang.

González-Blanco García, E., Rio Riande, G. del, Martínez Cantón, C. & Martos Pérez, M. D. (2014). "La codificación informática del sistema poético medieval castellano, problemas y propuestas en la elaboración de un repertorio métrico digital: ReMetCa". In A. Baraibar (Ed.), *Visibilidad y divulgación de la investigación desde las humanidades digitales. Experiencias y proyectos* (pp.185-203). Pamplona, Servicio de Publicaciones de la Universidad de Navarra. Colección BIADIG (Biblioteca Áurea Digital), 22 / Publicaciones Digitales del GRISO. Retrieved from http://hdl.handle.net/10171/35718 - accessed 23 March , 2017.

González-Blanco, E. G.,Martínez Cantón, C. & Rio Riande, G. del (2015). Making visible the invisible: metrical patterns, contrafacture and compilation in a Medieval Castilian Songbook. DH 2015: Global Digital Humanities, Sydney. Retrieved from http://dh2015.org/abstracts/ - accessed 23 March, 2017.

González-Blanco, E. G., Rio Riande, G. del & Martínez Cantón, C. (2016a). Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires. In John P. McCrae, Christian Chiarcos et al. (Eds.), *Proceedings of the LREC 2016 Workshop*, *LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-LDL2016_Proceedings.pdf - accessed 25 March, 2017.

González-Blanco, E. G., Rio Riande, G. del & Martínez Cantón, C. (2016b). "DH Poetry Modelling: a Quest for Philological and Technical Standardization". DH 2016, Krakow. Retrieved from http://dh2016.adho.org/abstracts/73 – accessed 23 March, 2017.

Nilsson, M., Baker, T., & Johnston, P. (2008). The Singapore Framework for Dublin Core Application Profiles. Retrieved from http://dublincore.org/documents/singapore-framework/- accessed 23 March, 2017.

Pue, S. A.; Teal,T. K. & Brown, C. T. (2015). Using Bioinformatic Algorithms to Analyze the Politics of Form in Modernist Urdu Poetry. DH 2015: Global Digital Humanities, Sydney.

Rumbaugh, J., Jacobson, I. and Booch, G. (2004). T*he Unified Modeling Language Reference Manual* (2nd Edition). Boston: Addison-Wesley.

W3C (2010). RDF - W3C standards. http://www.w3.org/RDF/ - accessed 23 March, 2017.