Digital repertoires of poetry metrics: towards a Linked Open Data ecosystem

Mariana Curado Malta[1][2], Elena Gonzalez-Blanco[1], Clara Martinez[1], and Gimena del Rio[3]

[1] LINHD-UNED, Madrid, Spain
mariana.malta@linhd.uned.es,{egonzalezblanco,cimartinez}@flog.uned.es
http://linhd.uned.es
[2] CEOS.PP, Polytechnic of Oporto, Portugal
mariana@iscap.ipp.pt
http://www.iscap.ipp.pt
[3] CONICET - Instituto de Investigaciones Bibliográficas y Crítica Textual, Buenos Aires, Argentina
gdelrio.riande@gmail.com
http://www.conicet.gov.ar

Abstract. This paper presents work-in-progress of the POSTDATA project. This project aims to provide means to solve the interoperability issues that exist among the digital poetry repertoires. These repertoires hold data of poetry metrics that is locked in their own databases and it is not freely available to be compared and to be used by intelligent machines that could infer over the data. The POSTDATA project will use Linked Open Data (LOD) technologies to overcome the interoperability problems. POSTDATA is developing a metadata application profile (MAP) for the digital poetry repertoires, a construct that enhances interoperability. This development follows the method for the development of MAP (Me4MAP). A MAP for the digital poetry repertoires will open doors for this repertoires to be able to structure the data with a common model in order to publish it as Linked Open Data. This paper presents how this MAP is being developed so far.

Keywords: Digital humanities, Linked Open Data, interoperability, metadata application profile, poetry, digital repertoires

1 Introduction

Metrics is a discipline that establishes what features configure the structure of a verse. A verse is a line in a poem. When analyzing poems it is necessary to count syllables, accents, rhythm and rhymes as essential elements to define the poem structure, its musicality and, sometimes, a given structure even determines the type of contents. The way in which this metrics has been conceptualized in the different traditions is however different, as there are many different ways of encoding and understanding all these metrical systems.

In the beginnings, repertoires were printed books in which we could find information listed in a way similar to an address book. Now, a digital repertoire of poetry metrics is a tool that gives account of metrical and rhythmical schemes of either a poetical tradition or school or period. It may gather a long corpus of poems which are defined and classified by their main characteristics.

The lack of interoperability between the different digital repertoires dealing with poetry metrics across the different languages, literatures and traditions is a problem that needs to be addressed. POSTDATA is a project financed by a Starting Grant of the European Research Council[4] that aims to solve this problem.

The reason for this absence of interoperability is twofold: 1) there is a lack of standardization in the philological field due to the independent evolution of each different cultural tradition; 2) the technological solutions used for building each poetic digital repertoire or database are very different, and tailored following a different model without taking into account, in most of the cases, the standards used in Digital Humanities. The basis of POSTDATA is building of a semantic system which will serve as bridge to mind the gap between the technological and philological worlds. It aims to develop a metadata application profile that will give a semantic model for all the existing poetic digital repertoires that are currently available on the Web of Documents[5]. With this common model all these repertoires will be able to publish its data as Linked Open Data and become interoperable among them.

The goal of this paper is to present how POSTDATA addresses the interoperability problem among the Digital Poetry Repertoires.

This paper proceeds as follows: Section 2 presents briefly the POSTDATA project and the quest for interoperability of the Digital Repertoires of poetry metrics; section 3 presents the metadata application profile (MAP) construct as a way to achieve interoperability, and a method to develop MAPs; section 4 reports on the first steps of the development of the MAP for European poetry. The last section presents conclusions and future work.

2 POSTDATA

POSTDATA aims at shortening the digital gap among poetry and technology, looking for interoperability solutions. This project has several dimensions as we can see in FIG.1.

It aims at building a digital research environment to create poetry collections and repertoires as well as poetic library of treatises, where users can consume

[4] ERC-2015-STG-679528

[5] "Web of Documents" is a term used in contrast with the term Web of Data. The Web of Documents is made of documents read by human beings that navigate between documents located in servers through hyper-links, it is the Web that everyone uses in a daily basis. The Web of Data or Linked Data or even the Semantic Web, three ways of expressing similar concepts, have technologies that "enable people to create data stores on the Web, build vocabularies, and write rules for handling data"[2]

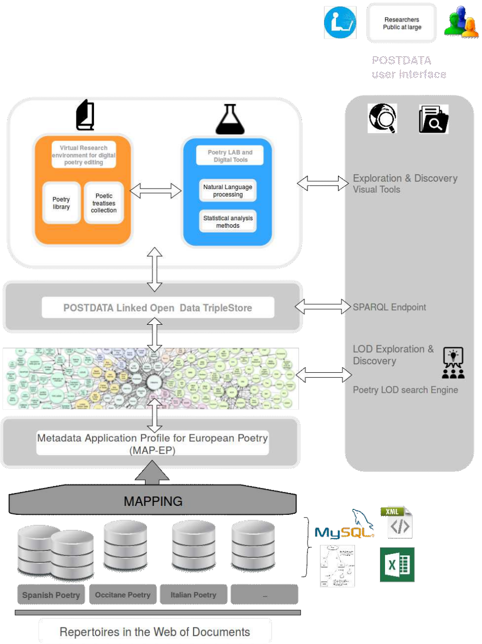
Researchers
Public at large
POSTDATA
user interface
Virtual Research environment for digital poetry editing
Poetry library
Poetic treatises collection
Poetry LAB and Digital Tools
Natural Language processing
Statistical analysis methods
Exploration & Discovery Visual Tools
POSTDATA Linked Open Data TripleStore
SPARQL Endpoint
LOD Exploration & Discovery
Poetry LOD search Engine
Metadata Application Profile for European Poetry (MAP-EP)
MAPPING
MySQL
XML
Spanish Poetry
Occitane Poetry
Italian Poetry
...
Repertoires in the Web of Documents

Fig. 1. POSTDATA project explanation schema.

information or contribute for the corpora of the library uploading their texts and analysis. Users will be able to use the service of Exploration & Discovery Visual Tools to visualize syntactical structures, perform word frequency analysis and textual patterns in poems in order to reflect metrical and rhythmical varieties. This visualization will use automated methods for poetry analysis combined with other technologies such as Natural Language Processing or Computational stylistics, combined with TEI-XML Encoding[6]. POSTDATA will develop tools to apply to the first level of poem analysis Natural Language Processing algorithms, such as Name-Entity Recognition systems to extract information, classify elements in text into pre-defined categories such as the names of persons, organizations, locations, and later revision of the results such as corrections and additions of information will also be possible. POSTDATA will also develop tools to perform statistical analysis of the poem, or of the corpora, to provide final users with relevant information. These analysis will be feed by both the data of the local repertoire as well as the data available in the Digital Poetry LOD.

There is already a very relevant set of digital poetic repertoires on the Web of Documents; there is also a certain number of local databases. All this resources constitute a rich kaleidoscope of multilingual virtual poetry. As examples we can refer repertoires in French: French lyrical collections (Nouveau Naete-

[6] See `http://www.tei-c.org` - Retrieved October 11, 2016

bus)[7], in Italian: Bibliografia Elettronica dei Trovatori (BedT)[8], in Hungarian: The Répertoire de la poésie hongroise ancienne (RPHA)[9], in Ancient Latin: The Corpus Rhythmorum Musicum[10], in Galaico-Portuguese: The Cantigas de Santa María[11], in Castellano: The Repertorio Métrico Digital de la Poesía Medieval Castellana (ReMetCa)[12], in Dutch: Dutch Song Database[13], in Occitane: Occitaine Répertoire métrique de la poésie lyrique occitane des troubadours à leurs héritiers[14], in Catalan: Repertori d'obres en vers[15], in Skaldic: The Skaldic Project[16], in German: The Lyrik des Minnesänger[17] and in English: the Digital Edition of the index of Middle English Verse[18], among many others. It is not in the aim of this paper to present all the repertoires, but only to show how alive the Digital Humanities Community of Poetry (DHCP) is, and how diverse and immense is the DHCP information available on the Web of Documents[1]. This data is at the moment locked in the silos of information of each repertoire, not available freely to be compared and to be used by intelligent machines that could infer many things over the data.

All these repertoires face a challenge of interoperability. POSTDATA addresses this issue by using LOD technologies. It will add a semantic layer to its repertoire (the set of all poetry collections) in order to be able to publish Poetic related data as Linked Open Data, and be interoperable with other entities that may do the same. POSTDATA will also provide a SPARQL endpoint for its dataset.

POSTDATA will not achieve anything without the contribution of the DHCP. The repertoires of this community have data that is trapped in the Web of Documents, and needs to be released, i.e., this data needs to be published as LOD. Making poetry available on line as machine-readable data will open a world of possibilities of linking, indexing and extracting new information through the combination of the different datasets. In order for this data to be interoperable POSTDATA needs to build a common model to structure DHCP data all in the same way. This "common model" is in fact a metadata application profile (MAP), a construct that enhances interoperability [3].

[7] `http://nouveaunaetebus.elte.hu` – Retrieved September 27, 2016
[8] `http://www.bedt.it/BEdT_04_25/inf_home_crediti.aspx` – Retrieved September 27, 2016
[9] `http://rpha.elte.hu/` – Retrieved September 27, 2016
[10] `http://www.corimu.unisi.it` – Retrieved September 27, 2016
[11] `http://csm.mml.ox.ac.uk/` – Retrieved September 27, 2016
[12] `http://www.remetca.uned.es` – Retrieved September 27, 2016
[13] `http://www.liederenbank.nl/` – Retrieved September 27, 2016
[14] `http://icalia.es/troubadours/ca/` – Retrieved September 27, 2016
[15] Local database
[16] `http://www.abdn.ac.uk/skaldic` – Retrieved September 27, 2016
[17] `http://www.lhm-online.de` – Retrieved September 27, 2016
[18] `http://dimev.net` – Retrieved September 27, 2016

3 Development of Metadata Application Profiles

A profile is a term used to refer to a document that shows how standards and specifications can be used to deploy a particular application. A metadata application profile is a construct that when used by a certain community enhances interoperability [4]. The Dublin Core Metadata Initiative (DCMI)[19], a well-known and influential global initiative concerned with metadata, defined the rules to build a MAP in a recommendation called "The Singapore Framework for Dublin Core Application Profiles" (see [4]). This recommendation says that a MAP is composed by:

- functional requirements,
- domain model,
- description set profile,
- usage guidelines (optional),
- syntax guidelines (optional).

The functional requirements state what kind of things the community of practice wants to do with the data. The domain model presents a way to model the concepts (abstract and not abstract) and respective properties that data represents.

A MAP targets a community, meaning that all the different members of that community must feel represented in the domain model described by that MAP. This representativity has to do with the fact that each member of the community must be able to describe its resources using the MAP defined by the community. If the MAP fails to serve a specific member of the community, this member will be "excluded" in the sense that its data will not be interoperable with the rest of the community of practice. If this exclusion happens it might mean that the MAP was not very well developed because it does not respond to the needs of all members of the community that integrated the development.

In LOD a certain community of practice served by a MAP might have other type of communities of practice that "live" in the boundaries of the community of practice the MAP serves. Both the "boundary community" and the community of practice might be interested in sharing part of the data, that is, might want to have a certain level of interoperability between them. During the MAP development process developers should be aware of these "boundary communities" and try to integrate, when possible, part of their characteristics. LOD is a wide and open ecosystem. The more "boundary communities" are "touched", the more probable is that the data is used.

The development of a MAP is though a crucial task for a community of practice. This development should be structured and integrate, since the early phases of development, elements of all representative members of the community of practice.

The DHCP organizations differ in organization-type, location, culture and in the language they speak. To find a common ground of understanding in such an

[19] `http://dublincore.org` - Retrieved October 6, 2016

environment becomes a huge challenge. This circumstance is not new for a MAP development. In fact such a development is often done in complex settings that are very open, in contrast with the development of software that serves a certain organization that is protected inside its walls of context, culture and language, where requirements can be elicitated using very well known techniques. In a MAP development, developers will never know in fact the total reach of the MAP, the community of practice that the MAP serves can be very well defined but there will be always a degree of uncertainty - to elicitate requirements is not easy in such uncertainty.

The authors think that the existence of a method for the development of a MAP may help to address all the referred challenges.

Recent studies say that there is no method for the development of MAPs (see [8]), in order to address this issue [5,6] have been working on the definition of a method for the development of metadata application profiles (Me4MAP). POSTDATA is using Me4MAP[20] to develop the MAP-EP.

4 MAP-EP Development process

The development of the MAP-EP faces the challenge to serve at least 14 repertoires that are presently active in the Web of Documents[21]. There are other initiatives that make part also of the community of practice, but are not core community. The poetic repertoires can be defined as the core community of DHCP, other initiatives such as the LOD project of "Biblioteca Nacional de España"[22], the data project of "Museo del Prado"[23], Pelagios[24], Biblissima[25], Claros[26], among others, are the "boundary communities" as previously called. These projects do not deal with poetry but with information concerning bibliographic records, arts in general, and geographical places and persons connected to the resources described (manuscripts, pieces of art, objects in general). POSTDATA also wants to have a certain degree of interoperability with these initiatives. The "Vision Statement" of the MAP-EP should clearly state what is the core domain and should also open doors to other boundary domains. POSTDATA Vision statement is still being defined.

As defined in Me4MAP the first activity is the first Singapore Stage (S1) which develops the Functional Requirements.

[20] Only draft versions are published so far. The first version of Me4MAP was submitted to an international research journal and is waiting for approval. POSTDATA team is using this first version of Me4MAP not yet published.

[21] This number is changing at the moment of writing this chapter since the project is a work-in-project

[22] See `http://www.datos.bne.es` - Retrieved October 7, 2016

[23] `https://www.museodelprado.es/modelo-semantico-digital/el-prado-en-la-web` - Retrieved October 11, 2016

[24] `http://commons.pelagios.org/` - Retrieved October 8, 2016

[25] `http://www.biblissima-condorcet.fr` - Retrieved October 8, 2016

[26] `http://www.clarosnet.org/` - Retrieved 8 October, 2016

According to Me4MAP the functional requirements can be elicitated using the technique of developing uses-cases. The development of POSTDATA use case model is build with the study of the: (i) functionalities of the digital repertoires that are on the Web of Documents; (ii) local repertoires that are being build by researchers, at the same time as the project is being developed, and that want to use POSTDATA tools to be able to share and use data. So far there are two of such local repertoires working with POSTDATA.

POSTDATA will also implement a survey to end users of the repertoires in order to understand what kind of things such users would like to do with the data. This survey will run on line. All POSTDATA partners (responsible of the repertoires) will help POSTDATA to disseminate the survey. The answers will be analyzed and a set of functionalities defined.

From all this work POSTDATA team will define a use case model that will explicit the Functional Requirements.

POSTDATA team is also already collecting information about the data models of the databases, that together with the functional requirements, will be used to define the Domain Model, the second Singapore Stage (S2) (the second activity defined by Me4MAP). This information is being collected, organized and analised. POSTDATA team contacted all the responsible of the repertoires in order to obtain documentation of the databases. To communicate with some of the responsible is not easy since many of them are not database experts so do not "speak the same language". This results in information that is not understandable or that it is not enough to get a data model. Many information is re-created with the help of philologists of the team, they analyse the Websites and their functionalities in order to understand the meaning of some fields. From the 21 repertoires we have collected so far information from 15 (see Table 1).

When defining the Domain Model, it will very important to be aware of standard conceptual models that exist in the same community of practice. POSTDATA team has in mind to study the FRBRoo[27] with the aim to integrate it in the domain model since it has become a very important conceptual model in the Galleries, Libraries, Archives and Museums (GLAM) community. FRBRoo is in fact an object-oriented formulation of the FRBR model[28] as an extension of CIDOC CRM[29].

TEI, the Text Encoded Initiative[30] that has a module for the description of poetry related resources, is a data model that should be taken in account. This data model is not yet deployed in the Semantic Web, and it is widely used by the DHPC (using XML related technologies).

Me4MAP defines another activity - to be developed in parallel with S2 - called Environmental Scan. According to Me4MAP, an Environmental Scan is

[27] See `http://archive.ifla.org/VII/s13/wgfrbr/FRBRoo_V9.1_PR.pdf` - accessed October 10, 2016

[28] See `http://www.ifla.org/publications/functional-requirements-for-bibliographic-records` - accessed October 10, 2016

[29] See `http://www.cidoc-crm.org` - accessed October 10, 2016

[30] See `http://www.tei.org` - accessed October 12, 2016

Table 1. Type of information sent by the responsible of the repertoires

Type	How many	Observations
MySQL dump script	5	Able to open in phpMyAdmin and able to analyse the Logical Model.
MWD file	2	Able to open with MySQL Workbench and able to analyse the Logical Model.
XML data files	2	Able to load the files to a XML parser and able to extract the XML Schema.
XML dtd file	1	Able to extract the XML Schema using a XML parser.
Perl script with data	1	Able to open the file with a plain text editor and able to analyse the file.
Excel file with data	1	Able to open the file with OpenOffice software and able to analyse the tables on the file.
Documentation	3	Pdf files with text explaining the tables and fields, some with ER diagrams of the database. Able to analise the pdf file - no possiblities to check inconsistencies.

a report that contains a review of the metadata schemas that are available in any serialization of the Semantic Web (e.g. RDF/XML, turtle, etc.) and that may serve the needs of the Domain Model The POSTDATA team is aware of the importance of using standard or/and the most used RDF vocabularies to achieve good levels of interoperability with other communities of practice. The study of these vocabularies is done in the Environmental Scan.

The development of the Environmental Scan of MAP-EP has already started but is still in the very beginnings of development. Nevertheless POSTDATA team has the following considerations:

- standards should be the most used, so dcterms[31] will be always a first choice to terms and classes
- Digital Manuscripts to Europeana (DM2E)[32] is a very important initiative that will be used to describe concepts related to manuscripts;
- the BIB FRAME vocabulary[33] and BIBO[34] ontology are also strong candidates to describe bibliographic records of the POSTDATA domain model.

[31] See `http://dublincore.org/documents/dcmi-terms/` - Retrieved October 10, 2016

[32] See `http://dm2e.eu/` - accessed 8 October, 2016

[33] See `https://www.loc.gov/bibframe/` - accessed October 17, 2016

[34] See `http://bibliontology.com/` - accessed 8 October, 2016

Since the domain of MAP-EP has names of persons and locations related to the bibliographic records, authority repertoires such as the geonames ontology[35], DBpedia[36] and VIAF directory[37] are planned to be used.

5 Conclusions and Future Work

This paper presents preliminary work of a research project financed by the European Research Council (ERC). This project (POSTDATA) wants to solve the interoperability problems that exist among the digital poetry repertoires. These repertoires are present in the Web of Documents or are local files holding data of poetry metrics, that is, information about poetry analysis. This data is trapped in every database, structured in many different ways, and it is not shared among repositories. The aim of POSTDATA is to liberate this data in a way that it can be shared and open, in order to be used by intelligent machines that can compare the data and infer over it arriving to new dimensions of knowledge. The solution to solve the interoperability issue referred is to use Linked Open Data technologies and to publish the data as LOD. The data needs to be structured with a common model, that is, a metadata application profile (MAP), a construct that enhances interoperability. POSTDATA is using Me4MAP, a method for the development of application profiles do develop a MAP for European Poetry (MAP-EP). This paper presents the way MAP-EP is being developed, showing how POSTDATA team is using: 1) the Websites and the logical models of the repertoires; 2) use-cases of work of researchers that are collecting poetry data and discussing with the POSTDATA team the things they want to do with the data and 3) a survey to final users of the existent repertoires asking about the things they would like to do with the data; to define the functional requirements and the domain model of the MAP. At the same time POSTDATA team is already developing the Environmental Scan, a report that states all the RDF vocabularies that may serve the domain model.

At the end of the project all repertoires will be able to map its relational models with the MAP-EP. And will be ready to publish data in LOD.

As future work the POSTDATA team has to follow the path defined by Me4MAP to develop MAP-EP. According to the plan, a first version of MAP-EP will be ready by the end of 2017.

During this development process a research team will be interested in monitoring the use of Me4MAP in order to validate it. Me4MAP was developed using a Design Science Research Methodology (see [10]). The evaluation of the use of Me4MAP will inform the construction moments of DSR in order to create a revised version of Me4MAP.

Acknowledgments. The authors would like to thank all the responsible of the repertoires for their availability in sharing information and discussing database

[35] See `http://www.geonames.org/ontology/` - Retrieved October 10, 2016

[36] `http://dbpedia.org` - Retrieved October 12, 2016

[37] `http://viaf.org` - Retrieved October 12, 2016

issues with the POSTDATA team.

Mariana Curado Malta thanks ISCAP.IPP for her 3 year's leave which opened the possibility to work in POSTDATA, a wonderful and challenging professional experience in Madrid.

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 Horizon2020 research and innovation programme, (http://postdata.linhd.es/).

References

1. González-Blanco García, E. and Seláf, L.: Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires, pp. 321-332. Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world, eds. Soriano, L., Coderch, M., Rovira, H., Sabaté, G. and Espluga, X. Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien: Peter Lang, (2014).
2. Semantic Web, https://www.w3.org/standards/semanticweb/
3. Interoperability Levels for Dublin Core Metadata, http://dublincore.org/documents/interoperability-levels/
4. The Singapore Framework for Dublin Core Application Profiles, http://dublincore.org/documents/singapore-framework/
5. Baker, T., Dekkers, M., Heery, R., Patel, M., Salokhe, G.: What terms does your metadata use? Application profiles as machine-understandable narratives. Journal of Digital information 2, 2 (2001)
6. Curado Malta, M. and Baptista, A. A.: Me4DCAP V0.1: A method for the development of Dublin Core Application Profiles. In Proceedings of the 17th International Conference on Electronic Publishing - Mining the Digital Information Networks, pp. 33 – 44. IOS Press (2013)
7. Curado Malta,M. & Baptista, A. A.: A Method for the Development of Dublin Core Application Profiles (Me4DCAP V0.2): Detailed Description. In International Conference on Dublin Core and Metadata Applications, pp.90-103. Dublin Core (2013)
8. Curado Malta, M. & Baptista, A.A.: State of the Art on Methodologies for the Development of a Metadata Application Profile. In proceedinggs of MTSR 2012, CCIS 343, pp. 61–73. Springer-Verlag, Berlin Heidelberg (2012)
9. IFLA: Functional Requirements for Bibliographic Records. International Federation of Library Associations and Institutions (2009)
10. Hevner, A.: The three cycle view of design science research. Scandinavian Journal of Information Systems. vol. 19 (2). pp. 87 (2007)