

Poetry and Digital Humanities making interoperability possible in a divided world of digital poetry: POSTDATA project

Elena Gonzalez-Blanco¹, Salvador Ros², Pablo Ruiz Favó², Maria Luisa Díez Platas²,
Helena Bermudez², Agustín Caminero², Clara I. Martínez Cantó³, Luciana Ayciríex²
¹Coverwallet

Principal Investigator of POSTDATA ERC Project.
elena@coverwallet.com

²Dep. Sistemas de Comunicación y Control-Laboratorio de Innovación en
Humanidades Digitales (LINHD).

Universidad Nacional de Educación a Distancia (UNED).

{sros, accaminero}@scc.uned.es

{ml.diezplatas, helena.bermudez, pablo.ruiz, luciana}@linhd.uned.es

³Dep. Literatura española y Teoría de la Literatura-Laboratorio de Innovación en
Humanidades Digitales (LINHD).

Universidad Nacional de Educación a Distancia (UNED).

cimartinez@flog.uned.es

ABSTRACT

POSTDATA (Poetry Standardization and Linked Open Data) aims at bridging the digital gap among traditional cultural assets and the growing world of data. It is focused on poetry analysis, classification and publication, applying Digital Humanities methods of academic analysis -such as XML-TEI encoding- (Dombrowski & Denbo, 2013), (Flanders & Hmali, 2013) in order to look for standardization, as well as innovation by using semantic web technologies (Cigarrán-Recuero et al 2014) to link and publish literary datasets in a structured way in the linked data cloud. The advantages of making poetry available online as machine-readable linked data are threefold: first, the academic community will have an accessible digital platform to work with poetic corpora and to contribute to its enrichment with their own texts; second, this way of encoding and standardizing poetic information will be a guarantee of preservation for poems published only in old books or even transmitted orally, as texts will be digitized and stored; third: datasets and corpora will be available and open access to be used by the community for other purposes, such as education, cultural diffusion or entertainment.

POSTDATA will be materialized in the creation of a digital semantic web-based platform for poetry analysis and edition, to study, publish and share digital collections in a virtual research environment using digital humanities open standards combined with traditional philological academic analysis. The environment will be open to any language and type of poetry and accessible for multiple users with different profiles, and it will provide access to digital resources on poetry linked together through data repositories. These data will be subsequently indexed by a search engine and finally provided simultaneously on a Website interface, an API and a single point of access to enriched and open information in accordance with RDF formalism.

The proposal is based in three pillars: Semantic modeling, poetry lab and infrastructure deployment.

a) Semantic Modeling: Linked Open Data (LOD)

The effort of gathering data with an encyclopedic spirit was the origin of poetical repertoires (Horváth 1991). Since then a huge number of them has been developed (see repertoires and digital database references). Interoperability among poetic repertoires is not simple, as there are not only technical issues involved, but also conceptual and terminological problems: each repertoire belongs to its own poetical tradition and each tradition has developed an idiosyncratic analytical terminology in a different and independent way for years. There are only a couple of studies which deal with some of the above mentioned aspects (Bootz & Szoniecky 2008 and Zöllner-Weber 2009), but there is not a conceptual model of ontology referred to metrics and

poetry yet. The result of this uncoordinated evolution is a bunch of varied terminologies to explain analogous metrical phenomena through the different poetic systems, whose correspondences have been hardly **studied** (González-Blanco & Seláf 2014, González-Blanco et al. 2014a and 2014b). These technologies have shown to be flexible enough to reflect poetic needs, but it is necessary to establish a common conceptual frame to order and classify the philological information in two levels: first an abstract level with general metadata for describing all the classes (such as “poem”, “stanza”, “line” or “accent”) and their properties (such as “has number of syllables” or “has rhyme”), and second an individual “thesaurus” or set of terms to build controlled vocabularies to name each particular phenomenon in any different literary corpora.. As no previous model of such a poetic conceptualization existed before, this model will be one of the main and most innovative contributions of the project.

To achieve POSTDATA goals it has been necessary to face some challenging situations. On the one hand, in order to propose a standard of poetry data model it has been necessary to face the interoperability problems between existent poetic repertoires. This has been solved by creating a common data model for the analysis of European poetry. For this purpose, a metadata application profile (MAP), a semantic model in the Linked Open Data (LOD) framework, has been built and at the time of writing, the validation process is finishing. This MAP will allow the communication of existing data that couldn't be shared before expanding the frontiers of knowledge and research.

Once a European poetry Map is developed, the data will be available in machine-readable format that allows us to apply the linked-data paradigm, Figure 1

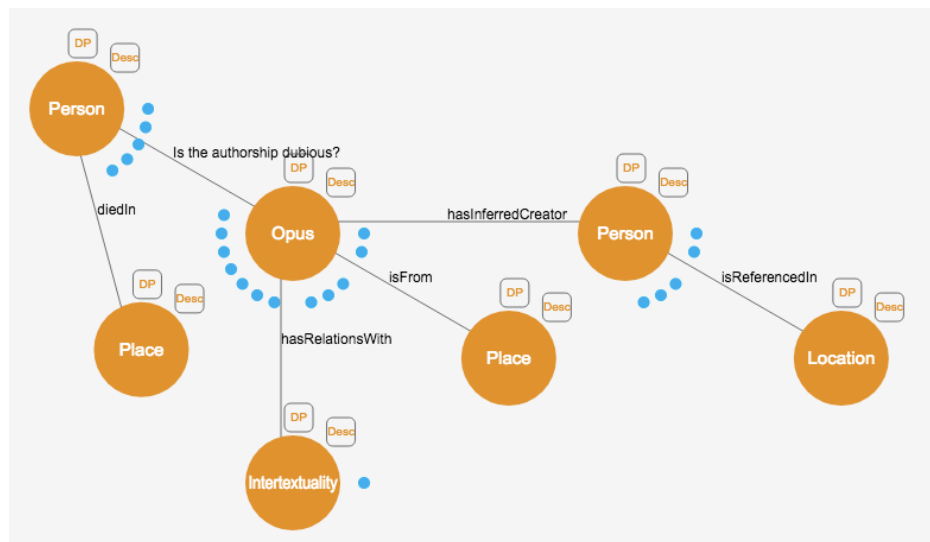


Figure 1. Excerpt of Postdata Model.

b) Poetry Lab: Natural Language Processing (NLP) and computational linguistics

There have been several interesting approaches to poetry analysis and generation from a computational linguistic perspective using automated linguistic analysis or text mining (Gervas, 2015). Nevertheless, there are two issues that are not easy to overcome: first, most teams that use these technologies are mainly formed by computer engineers with a limited knowledge of linguistic and literary issues, and second, the linguistic rules applied to poetry analysis are often imperfect, as poets usually break the “rules”. This is especially frequent when dealing with Medieval or ancient texts, as the exception to the rule are quotidian.

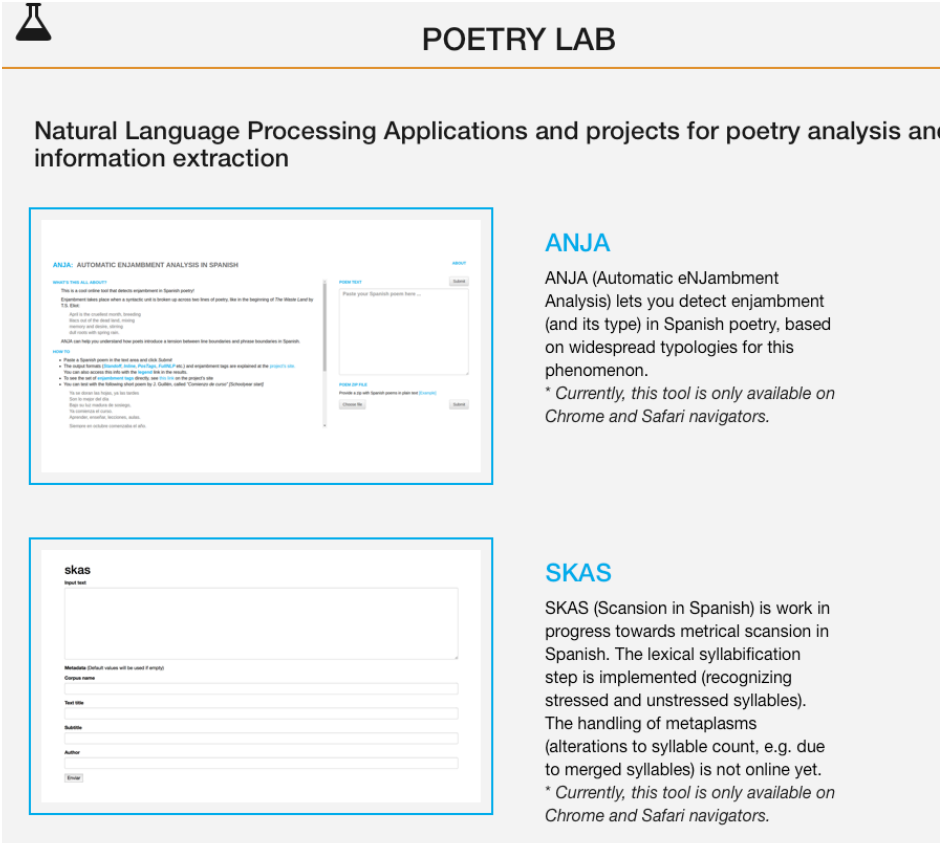
Poetry Lab it is a set of tools ranging all the different levels of poetry scholarship, from the most formal processes related with scansion to the most cognitive ones like metaphor understanding or other related with knowledge and subjective perception involving IA techniques.

On the other hand, the automatization of poetic analysis is , as well, a really challenging situation. Since the poetic analysis is diverse (different languages, different poetry tradicitons). Poetry Lab is a space where researchers would be able to implement the most up-to-date language technologies and computational methods to process poetry data. Since no set of tools to address

basic poetry issues existed before, the Poetry Lab will contribute to the researcher community and users by democratizing technology and user experience.

At this moment tree tools are available in Postadata Poetry Lab ,Figure 2:

1. ANJA devoted to automatic enjambment analysis in Spanish
2. SKAS a first step in metrical Spanish scansion. (lexical syllabification recognizing stressed and unstressed syllables)
3. HISMETAG a Entity recognition framework for Medieval Spanish



The image shows a screenshot of the Poetry Lab website. At the top, there is a logo of a flask and the text "POETRY LAB". Below this, the main heading reads "Natural Language Processing Applications and projects for poetry analysis and information extraction".

Two tool interfaces are highlighted with blue boxes:

- ANJA: AUTOMATIC ENJAMBMENT ANALYSIS IN SPANISH**: The interface includes a text input field, a "Submit" button, and a "Feedback" link. It also features a "Feedback" section with a "Submit" button. The tool is described as detecting enjambment in Spanish poetry based on widespread typologies.
- skas**: The interface includes a text input field, a "Submit" button, and a "Feedback" link. It is described as a tool for metrical scansion in Spanish, implementing lexical syllabification.

Both tools include a note: "* Currently, this tool is only available on Chrome and Safari navigators."

Figure 2. Poetry tools: <http://prototipo-postdata.linhd.uned.es>

c) Research Infrastructures: Social impact and user perception

Some important digital humanities initiatives have been developed with different targets (Schreibman & Hanlon 2010): annotation (MIT, 2018), transcription (FP, 2018), collaborative (MAR, 2018) and (Marta, 2018), edition of a corpus (TB, 2018), editorial platforms (Elaborate, 2018) and (Muruca, 2018), CMS based on metadata, (Omeka, 2018) and working groups (CAI, 2018). Nevertheless, there is not a single platform devoted to poetry analysis, edition, visualization and publication, user-friendly and based on a linked open data system, such as the one proposed in this project.

The third pillar of this proposal is focused on the creation of a digital platform for poetry edition oriented to different kind of users: scholars with academic purposes who want to work on critical digital editions, non-experienced users 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. It will have the interoperable capacity that allows us "recycling" and 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.

CONCLUSIONS

As it has been described, the state of digital poetry world is still very idiosyncratic, unconnected and uneven from the technological readiness levels point of view. POSTDATA project aims at becoming a reference in

terms of philological digital humanities standardization, interoperability by using linked open data and a digital platform user-centric which will go far beyond research to enhance the digital poetic user experience. This project is based on a crowdsourcing philosophy, as it will create a virtual environment in which scholars and users may add, analyze, publish and reuse poems and data.

Acknowledgment

Authors would like to acknowledge the support of European research project ERC-2015-STG-679528 POSTDATA, research project (2014I/PPRO/ 031) from UNED. Furthermore, authors thank the staff of LINHD – the Digital Humanities Innovation Lab at UNED

REFERENCES

Repertoires and digital databases

- Alberni, Anna, *The Last Song of the Troubadours*, <http://icalia.es/troubadours/en/home>.
- Asperti, Stefano, Fabio Zinelli et al. *BedT, Bibliografia Elettronica dei Trovatori*. www.bedt.it.
- Brea, Mercedes, et al., 1994-2014. *MedDB: Base de datos da Lírica profana galego-portuguesa*. <http://www.cirp.es/bdo/med/meddb.html>.
- Colombi, Emanuela, *Pedecerto. Metrica Latina Digitale*, <http://www.pedecerto.eu/>
- Fumerton, Patricia, *English Broadside Ballad*, <http://ebba.english.ucsb.edu/>
- González-Blanco, Elena, et al. 2013. *ReMetCa: A Digital Repertoire on Medieval Spanish Metrics*. www.remetca.uned.es.
- Grijp, Louis, et al., *Dutch Song Database*. <http://www.liederenbank.nl/index.php?lan=en>.
- Halvor Undlien, Jan, *Henrik Ibsen Skrifter*. <http://www.ibsen.uio.no>
- Horváth, Iván, et al. 1991-2014. *Répertoire de la poésie hongroise ancienne*. <http://rpha.elte.hu/>.
- Leonardi, Lino, *Repertorio della tradizione poetica italiana dai Siciliani a Petrarca*. <http://www.mirabileweb.it/>.
- Mooney, Linne R., *Digital Index of Middle English Verse*. <http://www.dimev.net/>
- Parkinson, Stephen, et al., *The Oxford Cantigas de Santa Maria Database*. <http://csm.mml.ox.ac.uk/>.
- Plecháč, Petr, *Czech Versification Research Group*. <http://www.versologie.cz/en/index.html>.
- Rauner, Erwin, *Analecta Hymnica Medii Aevi Digitalia*. http://webserver.erwin-rauner.de/crophius/Analecta_conspectus.htm.
- Seláf, Levente, *Le Nouveau Naetebus. Répertoire des poèmes strophiques non-lyriques en langue française d'avant 1400*. www.nouveaunaetebus.elte.hu.
- Stella, Francesco. *Corpus Rhythmorum Musicum*, <http://www.corimu.unisi.it>.
- Willis, Tarrin, *Skaldic poetry of Scandinavian Middle Ages*: <https://www.abdn.ac.uk/skaldic/db.php>

Metrical repertoires published in paper

- Antonelli, Roberto, *Repertorio metrico della scuola poetica siciliana*, Palermo, Centro di Studi Filologici e Linguistici Siciliani, 1984.
- Betti, Maria Pia, *Repertorio Metrico delle Cantigas de Santa Maria di Alfonso X di Castigli*, Pisa, Pacini 2005.
- Brunner, Horst, Burghart Wachinger and Eva Klesatschke, *Repertorium der Sangsprüche und Meisterlieder des 12. bis 18. Jahrhunderts*, Tübingen, Niemeyer, 1986-2007.
- Frank, István, *Répertoire métrique de la poésie des troubadours*, Paris, H. Champion, 1953-57.
- Gorni, Guglielmo, *Repertorio metrico della canzone italiana dalle origini al Cinquecento (REMCI)*. Firenze, Franco Cesati, 2008.
- Gómez Bravo, Ana María, *Repertorio métrico de la poesía cancioneril del siglo XV*. Alcalá de Henares, Universidad, 1999.
- Mölk, Ulrich, and Friedrich Wolfzettel, *Répertoire métrique de la poésie lyrique française des origines à 1350*, München, W. Fink Verlag, 1972.
- Naetebus, Gotthold. *Die Nicht-Lyrischen Strophenformen Des Altfranzösischen. Ein Verzeichnis Zusammengestellt Und Erläutert*. Leipzig, S. Hirzel, 1891.
- Pagnotta, Linda, *Repertorio metrico della ballata italiana*. Milano, Napoli, Ricciardi, 1995.
- Parramon i Blasco, Jordi, *Repertori mètric de la poesia catalana medieval*. Barcelona, Curial, 1992.

- Pillet, Alfred, and Henry Carstens, *Bibliographie der Troubadours*, Halle, M. Niemeyer, 1933.
- Raynaud, Gaston, *Bibliographie des chansonniers français des xiii. [treizième] et xiv. [quatorzième] siècles*, Paris, Vieweg, 1884.
- Solimena, Adriana, *Repertorio metrico dei poeti siculo-toscani*. Centro di studi filologici e linguistici siciliani in Palermo, Palermo, 2000.
- _____, *Repertorio metrico dello Stil novo*, Roma, Presso la Società, 1980.
- Tavani, Guiseppe, *Repertorio metrico della lingua galego-portoghese*, Roma, Edizioni dell'Ateneo, 1967.
- Touber, Anton H., *Deutsche Strophenformen des Mittelalters*, Stuttgart, Metzler, 1975.
- Zenari, Massimo, *Repertorio metrico dei 'Rerum vulgarium fragmenta' di Francesco Petrarca*. Padova, Antenore, 1999.

Cited references

- Bootz, P. & S. Szoniecky, "Towards an ontology of the field of digital poetry", paper presented at *Electronic Literature in Europe*, 2008. Full text available at <http://elmcip.net/node/415>
- Cigarrán-Recuero, Juan, Gayoso-Cabada, Joaquín, Rodríguez-Artacho, Miguel, Romero-López, Dolores, y Sarasa-Cabezuelo, Antonio, "Assessing semantic annotation activities with formal concept analysis", *Expert Systems with Applications* 41, 2014, 5495–5508.
- CAI, *Cost Action Interedition* www.interedition.eu, Last access 15th june 2018
- Dombrowski, Q. & S. Denbo, "TEI and Project Bamboo", *Journal of the Text Encoding Initiative*, Issue 5, 2013. <http://jtei.revues.org/787>
- Dublin Core Metadata Initiative <http://dublincore.org/> (ISO 15836:2009). Last accessed 20th January 2015. Elaborate <http://www.e-laborate.nl/en/>, Last access 15 th june 2018
- Flanders, J. & S. Hamlin, "TAPAS: Building a TEI Publishing and Repository Service", *Journal of the Text Encoding Initiative* 5, 2013, <http://jtei.revues.org/788>
- FP, *From the Page* <http://beta.fromthepage.com/> Last access 15 th june 2018
- Gervás, Pablo, "A Logic Programming Application for the Analysis of Spanish Verse", *First International Conference on Computational Logic*, Imperial College, London, 2000. <http://nil.fdi.ucm.es/sites/default/files/GervasCL2000.pdf>. Last access 20 January 2015.
- González-Blanco, Elena, & L. Seláf, 2014. "Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires", en *Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world*, eds. Soriano, L., et al, Oxford, New York, Wien, Peter Lang, 2014.
- González-Blanco, Elena, M. G del Río, C. I. Martínez & M. D. Martos, "Una propuesta de integración del sistema de formularios de bases de datos MySQL con etiquetado TEI: ReMetCa, Repertorio Digital de la Métrica Medieval Castellana", *Janus Digital*, Annex I, *Humanidades Digitales: Desafíos, logros y perspectivas de futuro*, ed. López Poza, S. & N. Pena Sueiro. 2014, 209-219. <http://www.janusdigital.es/anexos/contribucion.htm?id=19>
- González-Blanco, Elena, M. G del Río, C. I. Martínez & M. D. Martos, "La codificación informática del sistema poético medieval castellano, problemas y propuestas en la elaboración de un repertorio métrico digital: ReMetCa", in *Visibilidad y divulgación de la investigación desde las humanidades digitales*. ed. Baraibar, A., Pamplona, U. de Navarra, 2014. BIADIG Collection, 22, 185-203. URL: <http://hdl.handle.net/10171/35718>
- González-Blanco, Elena, L. Seláf, M. G del Río, C. I. Martínez & M. D. Martos, "Building a metrical ontology as a model to link digital poetic repertoires", Long paper presented at the DH2014 International Conference in Lausanne 2014. Abstract published at: <http://dharchive.org/paper/DH2014/Paper-674.xml>
- MARL: *Mapping the Republic of Letters* <http://republicofletters.stanford.edu/> Last access 15 th june 2018
- Martha Berry Digital Archive <https://mbda.berry.edu/> Last access 15 th june 2018
- MIT Annotation Hyperstudio <http://hyperstudio.mit.edu/projects/annotation-studio/>, Last access 15 th june 2018
- Muruca <http://www.muruca.org/>, Last access 15th june 2018
- Omeka <http://omeka.org/>, Last access 15th june 2018
- Schreibman, S. & A. M. Hanlon, "Determining Value for Digital Humanities Tools: Report on a Survey of Tool Developers", *Digital Humanities Quarterly* 4:2, 2010. <http://www.digitalhumanities.org/dhq/>
- TB, *Transcribe Bentham* <http://www.transcribe-bentham.da.ulcc.ac.uk/>, Last access 15 th june 2018