Hispanic Medieval Tagger (HisMeTag): una aplicación web para el etiquetado de entidades en textos medievales

M. Luisa Díez Platas, Ll. Tobarra, S. Ros, E. González-Blanco, A. Robles-Gómez, A.C. Caminero, Gimena del Río Riande

1. Introducción

El impacto de las nuevas tecnologías en la investigación en el ámbito de las humanidades ha supuesto una revolución en la recopilación, acceso y representación de la información así como en el procesamiento de los grandes volúmenes de datos y textos. Es en este contexto y como respuesta a las necesidades crecientes de los proyectos de investigación, en el que surge el término Humanidades Digitales (DH, *Digital Humanities*) como punto de conexión de distintas disciplinas y que establece la relación entre las nuevas tecnologías y las diferentes comunidades científicas (Del Río Riande, 2016) (Gónzalez-Blanco, 2013)

El creciente uso de la computación en la nube para facilitar la gestión de proyectos humanísticos ha dado lugar a distintas iniciativas como los Entornos Virtuales de Investigación (VRE, *Virtual Research Environments*) (Bellamy, 2012) desarrollados preferentemente en laboratorios y centros de investigación interdisciplinares especializados en Humanidades Digitales como, entre otros, el Laboratorio de Innovación en Humanidades Digitales de la UNED (LINHD)

Las Humanidades Digitales han establecido originariamente fuertes vínculos con la filología (un testimonio significativo es el consorcio TEI, creador del lenguaje de marcado que abandera esta disciplina) y, sin embargo, en los últimos años, la tendencia hacia la automatización ha provocado que el análisis textual se haya orientado hacia la lingüística computacional y el área de procesamiento del lenguaje natural (PLN), ámbitos que tradicionalmente no se asociaban con las humanidades en sí mismas. El proyecto *POSTDATA* (*Poetry Standardization and Linked Open Data*), que participa de esta tendencia, incorpora en uno de sus paquetes de trabajo el tratamiento computacional del texto en combinación con el análisis semántico de las categorías poéticas. En el seno de este proyecto ha tenido lugar la creación de una de sus primeras aplicaciones, a herramienta *Hispanic Medieval Tagger (*HisMeTag*)* .

HisMeTag permite localizar entidades nombradas en textos escritos en español medieval, mediante un proceso automático de reconocimiento de entidades nombradas (NER) y técnicas de PLN para el procesamiento lingüístico y la generación de las distintas variantes que existieron en la época medieval. Localiza, etiqueta términos conocidos y propone nuevos términos para su validación.

La accesibilidad, las distintas formas de visualización, la posibilidad de edición y la capacidad de presentar resultados para su análisis son elementos clave para asegurar el éxito y la funcionalidad de HisMetag, para lo que se ha considerado oportuno construir una plataforma web, funcional y usable.

Sin la intervención explícita del usuario, se transforma un documento textual inicial, bien en formato de texto sencillo o en XML-TEI (TEIConsortium, 2015) en otro documento en formato XML-TEI, pero enriquecido con información sobre los nombres propios, junto con diversos archivos de visualización (en formato HTML o CSV) que incluyen información relevante y totalmente preparada para su futuro análisis, o incluso para ser insertados en proyectos de investigación directamente, si fuera necesario.

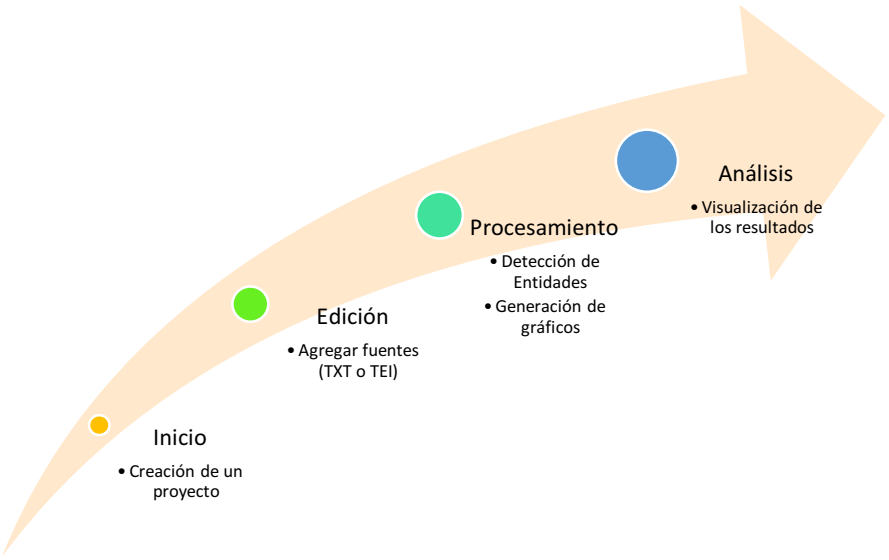
2. *Hispanic Medieval Tagger (*HisMeTag *)*

La aplicación web *HisMeTag* está orientada al análisis de textos medievales teniendo en cuenta las variantes lingüísticas y sus especiales características sintácticas para obtener información sobre las entidades nombradas mencionados en el texto.

Para esta primera demostración de la herramienta se han seleccionado dos textos medievales castellanos del Siglo XIII de distinta naturaleza: un documento notarial de la época alfonsí y un fragmento del *Libre dels tres reys d'Orient* (Gago Jover, 2011).

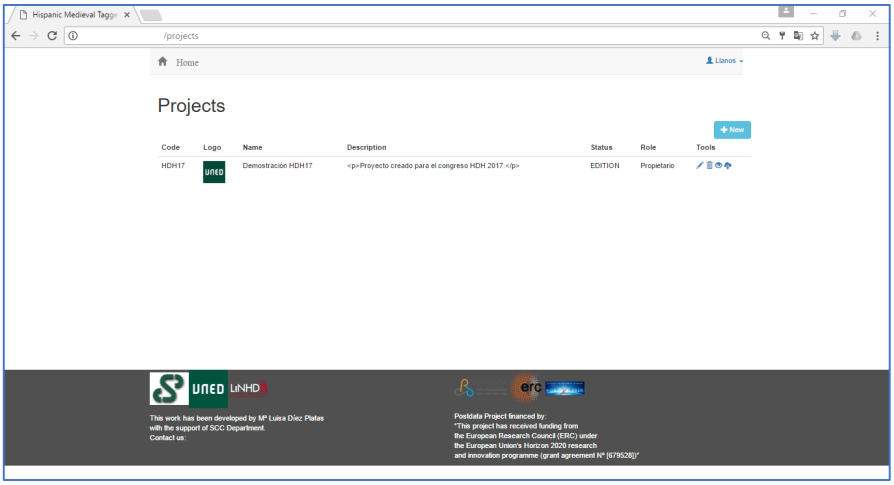
La metodología de desarrollo de la aplicación web ha seguido el patrón modelo-vista-controlador desarrollada en *Python* y basada en el *microframework Flask* (Ronacher, 2017). La interfaz se ha creado mediante plantillas Jinja2 (Ronacher, Jinja, 2017) y la biblioteca *Bootstrap* para desplegar un sitio web basado en HTML5, accesible y adaptable.

El flujo de trabajo de la aplicación web es el siguiente:



Las funcionalidades que proporciona la herramienta se describen a continuación:

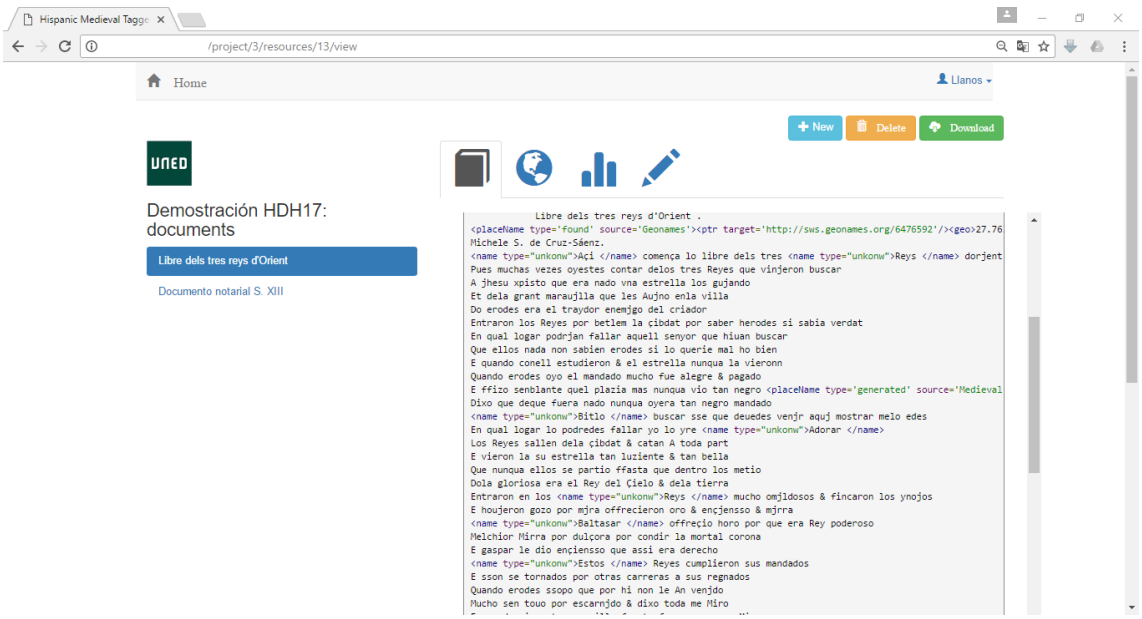
- Acepta archivos de entrada de texto TXT y archivos XML-TEI.
- Permite agrupar diferentes documentos con diferentes versiones para el análisis conjunto o independiente. Los corpus de textos se pueden organizar mediante proyectos para cada usuario registrado.



- Permite incorporar información descriptiva sobre textos que será incorporada al archivo XML-TEI de salida como metadatos
- Reconoce entidades nombradas de distinta naturaleza mediante el uso de *Gazetteers,* diccionarios y lexicones
- Realiza reconocimiento de variantes de nombres mediante un proceso de generación de variantes asociadas y reconocidas aplicando transformaciones morfológicas y fonéticas propias de la norma y la evolución lingüística de la época medieval (por ejemplo, Seulia sería asociado a Sevilla tras la trasformación)
- Cataloga términos no reconocidos de acuerdo con el contexto semántico de aparición
- Resuelve términos ambigüos usando contextos semánticos

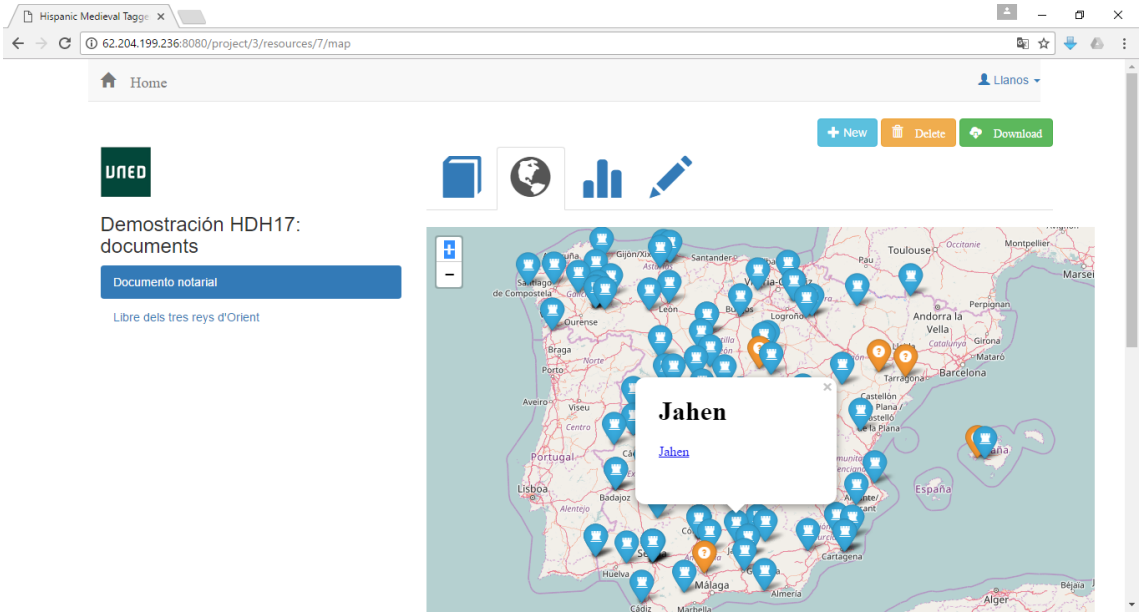
Los resultados se visualizan de diferentes formas:

- Archivo XML-TEI con todas las entidades etiquetadas.

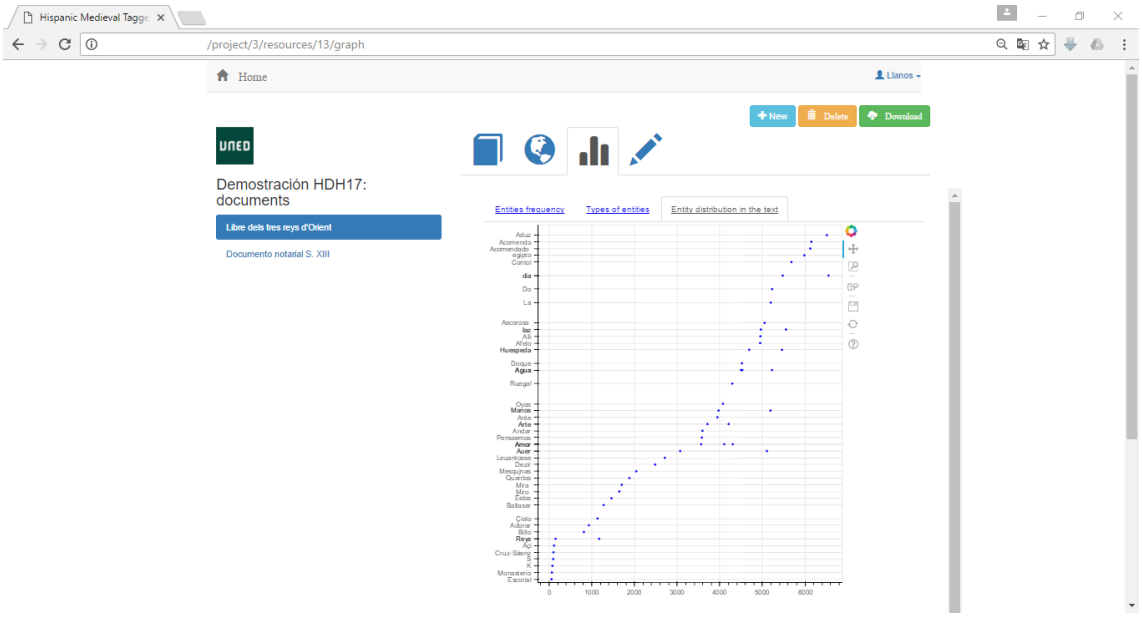


- Archivos CSV con las entidades reconocidas y características
- Mapa geográfico de las localizaciones identificadas. Para crear el mapa la aplicación utiliza el servicio cartográfico *Leaflet* (Vladimir & OpenMaps, s.f.), que

permite generar mapas anotados. En el mapa se muestran los lugares que se han identificado con su ubicación actual. Existen otros términos geográficos que no es posible representar porque se ignora su ubicación actual; por ejemplo, la Atlántida. También se incluyen términos ambiguos que podrían interpretarse tanto como nombres de persona como de lugar; como por ejemplo el término Moya, que corresponde tanto a un pueblo como a un apellido. Este mapa se almacena en un fichero HTML para su posterior exportación.



- Gráficas que se generan usando el paquete *Python Bokeh* (Continuum Analytics, 2015). Esta biblioteca genera gráficos interactivos con opciones como zoom o selección de solo un fragmento, posibilitando descargar los gráficos resultantes en una imagen. También se genera una página web HTML que facilita la exportación de los gráficos con las herramientas que proporciona *Bokeh*. En estas gráficas se presentan:
    - la frecuencia absoluta para todos los términos descubiertos,
    - el número de términos localizados por categorías,
    - la distribución de apariciones de cada entidad en el texto.



La plataforma permite descargar los resultados del procesamiento realizado en una carpeta comprimida, para todo el proyecto o sólo uno de los recursos. Así, pueden utilizarse los resultados con otras utilidades que acepten los formatos de salida descritos.

3. Conclusiones

Se presenta una versión preliminar de una aplicación web orientada al reconocimiento de entidades nombradas y análisis de textos en español medieval que arroja resultados innovadores para el procesamiento de un corpus textual que hasta ahora se analizaba con métodos manuales. Permite la gestión de proyectos de investigación agrupando los documentos asociados y mediante un potente reconocedor realiza la detección automática de términos, etiquetándolos o generando un documento TEI válido dependiendo de la entrada proporcionada. Sobre ese procesamiento inicial la herramienta ofrece diversas visualizaciones que facilitan la tarea del análisis inicial del texto en su contexto geográfico.

Agradecimientos

Este trabajo se enmarca dentro de los proyectos de investigación, proyecto europeo ERC-2015-STG-679528 POSTDATA, y los proyectos financiados por el MINECO: Acción Europa Investiga EUIN2013-50630: Repertorio Digital de Poesía Europea (DIREPO) y FFI2014-57961-R Laboratorio de Innovación en Humanidades Digitales: Edición Digital, Datos Enlazados y Entorno Virtual de Investigación para el trabajo en humanidades, dirigidos por Elena González-Blanco. Además, los autores agradecen el soporte del proyecto de investigación (2014I/PPRO/031) de UNED junto con el Banco Santander, y a la Comunidad de Madrid por el apoyo prestado mediante la Red de Excelencia e-Madrid (S2013-ICE2715). También expresan su agradecimiento al apoyo prestado por SNOLA, reconocida Red de Excelencia (TIN2015-71669-REDT) por el Ministerio de Economía y Competitividad de España. La herramienta se comenzó a desarrollar gracias al proyecto de investigación *microgrant Mediaeval Iberian* de *Pelagios Commons* (2016).

Referencias

Continuum Analytics. (2015). *Welcome to Bokeh*. Retrieved from http://bokeh.pydata.org/en/latest/

Bellamy, C. (2012). The Sound of Many Hands Clapping: Teaching the Digital Humanities through Virtual Research Environment (VREs). *Digital Humanities Quarterly, 6*(2). Retrieved from http://www.digitalhumanities.org/dhq/vol/6/2/000119/000119.html

Del Río Riande, G. (2016, Marzo 22). Humanidades Digitales. Construcciones locales en contextos globales. SEDICIBlog. Retrieved Mayo 28, 2017, from http://sedici.unlp.edu.ar/blog/2016/03/22/humanidades-digitales-construcciones-locales-en-contexto

Del Río Riande, G., González-Blanco, E., Martínez-Cantón, C., Ros, S., Robles-Gómez, A., Pastor, R., . . . Caminero, A. (2017). EVI-LINHD, A Virtual Research Environment for the Spanisk-Speaking Community. *Digital Scholarship in the Humanities (DSH) (accepted for publication)*.

Gago Jover, F. (2011). *"Libre dels tres reys d'Orient", Textos Poéticos Españoles. Digital Library of Old Spanish Texts. Hispanic Seminary of Medieval Studies.* Retrieved from http://www.hispanicseminary.org/t&c/poe/docs/text.tro.htm

Gónzalez-Blanco, E. (2013). Actualidad de las Humanidades Digitales y un ejemplo de ensamblaje poético en la red: ReMetCa. *Cuadernos Hispanoamericanos*, 53-67.

Ronacher, A. (2017). *Flask, web development, one drop at a time*. Retrieved from http://flask.pocoo.org/

Ronacher, A. (2017). *Jinja*. Retrieved from http://jinja.pocoo.org/docs/2.9/

TEIConsortium. (2015). *Text Encoding Iniciative* . Retrieved from http://www.tei-c.org/Guidelines/P5/

Vladimir, A., & OpenMaps. (n.d.). *Leaflet, an open-source JavaScript library for mobile-friendly interactive maps*. Retrieved from http://leafletjs.com/