ANÁLISIS

DATOS ENLAZADOS DE PUBLICACIONES, PROYECTOS Y HERRAMIENTAS PARA INVESTIGADORES EN HUMANIDADES DIGITALES: CATÁLOGO PILOTO DEL CENTRO *CLARIN IULA-UPF*

Silvia Arano-Poggi y Núria Bel

Silvia Arano-Poggi es técnica de investigación en el *Proyecto Clarin/Feder* en la *Universitat Pompeu Fabra* (*UPF*), y colaboradora del grupo de investigación *Tecnologías de los recursos lingüísticos*. Sus líneas de investigación son organización del conocimiento, comportamiento informacional, aprendizaje virtual, y humanidades digitales. También es colaboradora docente en la *UOC*.
http://orcid.org/0000-0001-9436-3301

silvia.arano@upf.edu

Núria Bel es profesora agregada en la *Universitat Pompeu Fabra* (*UPF*) y dirige el grupo de investigación *Tecnologías de los recursos lingüísticos*. Sus líneas de investigación son las tecnologías lingüísticas, especialmente la generación de recursos para tecnologías como la traducción automática, extracción y minería de información. Ha participado en proyectos financiados por la UE como: *Panacea* (2010-2012, 7FP-ITC-248064) y *Metanet4u* (2010-2013, CIP-PSP-270893), este último proyecto de la red de excelencia europea *Multilingual Technology Alliance*, *Meta-net*.
http://orcid.org/0000-0001-9346-7803

nuria.bel@upf.edu

Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada
Roc Boronat, 138. 08018 Barcelona, España

Resumen

Los investigadores en Humanidades Digitales tienen dificultades para acceder y utilizar herramientas informáticas que les asistan en la explotación de los textos objeto de sus estudios. Esto se debe al hecho de que, en la mayoría de los portales y registros especializados, la información sobre dichas herramientas no está enlazada con la información sobre dónde encontrarlas y cómo utilizarlas. El *Centro de Competencias Clarin IULA-UPF* compila e interrelaciona la información necesaria en un catálogo de datos enlazados para ofrecer a los investigadores una forma integrada de acceder a toda la información. En este artículo se presentan los detalles del diseño y de la selección de materiales e instrumentos de descripción utilizados en la elaboración de dicho catálogo.

Palabras clave

Humanidades digitales, Catálogos, Centros de competencias, Datos enlazados, Servicios web, *Clarin IULA-UPF.*

Title: Linked data of publications, projects and computational tools for digital humanities researchers: the pilot catalogue of the *Clarin IULA-UPF Center*

Abstract

Digital humanities researchers experience some difficulties in accessing and using computational tools which are meant to assist them when carrying out queries of texts analysis in their studies. This is due to the fact that information concerning tools in this field and their effective usage is not linked to the information about their location and utilization. *Clarin IULA-UPF Center* aims at compiling and relating required information in a catalogue of linked data in order to offer researchers an integrated way of accessing it. This article presents details about the design and selection of the materials and description tools included in the catalogue.

Keywords

Digital humanities, Catalogues, Competence centres, Linked data, Web services, *Clarin IULA-UPF*.

Artículo recibido el 09-05-2014
Aceptación definitiva: 28-08-2014

Arano-Poggi, Silvia; **Bel, Núria** (2014). "Datos enlazados de publicaciones, proyectos y herramientas para investigadores en humanidades digitales: catálogo piloto del centro *Clarin IULA-UPF*". *El profesional de la información*, v. 23, n. 6, noviembre-diciembre, pp. 633-642.

http://dx.doi.org/10.3145/epi.2014.nov.11

1. Introducción

Las humanidades digitales (DH) han aumentado el interés por la utilización de herramientas informáticas para el análisis y explotación de textos, y en particular por las tecnologías de la lengua (**Juola**, 2008; **Schreibman**; **Hanlon**, 2010; **Brier**; **Hopp**, 2011; **Wiedemann**, 2013). Sin embargo el acceso a información sobre la disponibilidad y uso de estas aplicaciones puede resultar arduo para los investigadores del área. Las investigaciones están accesibles en bases de datos de artículos científicos o proyectos, portales de asociaciones profesionales, etc. (como *Scopus*, *Cordis*, *Association for Computational Linguistic - Anthology*); pero las descripciones de las aplicaciones se difunden en portales y registros especializados (por ejemplo *Bamboo DiRT - Digital Research Tools*; *Text Analysis Portal for Research*: *TAPoR 2.0*; *Virtual Language Observatory*). Esto es, no hay relación entre la documentación relativa a la aplicación de las herramientas a diferentes objetos de estudio y la documentación de las mismas.

http://www.elsevier.com/online-tools/scopus
http://cordis.europa.eu/fp7/projects_es.html
http://www.aclweb.org/anthology
http://dirtdirectory.org
http://www.tapor.ca
http://catalog.clarin.eu/vlo

En el marco de la iniciativa europea *Clarin*[1] que pretende facilitar el uso de aplicaciones informáticas en la investigación en humanidades y ciencias sociales (HSC), y gracias a financiación *Feder*[2] y *Universidad Pompeu Fabra*, se puso en marcha en España el *Centro de competencias Clarin IULA-UPF*[3]. Su misión es dinamizar el uso de las tecnologías de la lengua en la investigación en humanidades, y asesorar a los investigadores en materia de acceso a datos y aplicaciones informáticas para llevar a cabo investigación experimental, tanto a nivel académico como industrial, que tenga como objeto la extracción de datos en textos en cualquier lengua.

> Las HD han aumentado el interés en la utilización de herramientas informáticas para el análisis y explotación de textos, en particular de las aplicaciones de las tecnologías de la lengua

El centro ofrece acceso a herramientas útiles para el análisis y la explotación de textos en forma de servicios web. La descripción de las mismas se relaciona, gracias a la tecnología de datos enlazados abiertos (*linked open data*, LOD), con información y enlaces a publicaciones y proyectos.
http://www.clarin-es-lab.org

Como forma de contribuir a la colaboración interdisciplinaria y a la reutilización y difusión de los resultados de investigación, se elaboró un catálogo con valor añadido para potenciales usuarios de las DH. Se ha considerado valor añadido no sólo la interrelación de información de publicaciones, proyectos y herramientas, sino también la puesta en práctica de una política singular de selección para el crecimiento del catálogo. La política, que tiene en consideración a los posibles usuarios, se basa en dos ideas fundamentales:

a) promover el razonamiento en casos análogos: que como resultado de la navegación en el catálogo, los usuarios puedan reproducir, emular o adaptar las experiencias relatadas en artículos, comunicaciones o proyectos con diferentes textos, similares herramientas o en otras disciplinas;

b) enfatizar el carácter "operativo" de la información: el concepto de información operativa (*actionable information*[4]), que se toma prestado de la gestión del conocimiento, hace referencia al uso dado a la información. La información operativa se presenta en un formato fácilmente interpretable que busca despertar la capacidad exploratoria del usuario con la finalidad de que utilice o no dicha información para la toma de decisiones.

Una vez incorporados los nuevos materiales seleccionados bajo esta política, el usuario será capaz de construir a partir de la navegación en el catálogo y gracias a su estructuración utilizando recursos de la web semántica, su propio itinerario para, por ejemplo, descubrir cómo usan otros una aplicación informática a partir de la lectura de publicaciones, y experimentar con una herramienta similar.

Por último se destaca el carácter experimental del proyecto. La versión del catálogo presenta unos contenidos y una estructura que no son definitivos, ya que está en período de pruebas por parte de usuarios potenciales[5]. El catálogo piloto muestra el trabajo de modelado de datos y la vinculación de información relativa a una de las posibles tareas incluida en los procesos de análisis de texto: el reconocimiento de entidades.

2. Metodología

Para la elaboración del catálogo piloto se han seleccionado artículos y descripciones de proyectos, sobre las aplicaciones informáticas más utilizadas en la investigación actual en DH, y quién y cómo las han utilizado. Durante la búsqueda y análisis preliminar de información se constata que en general los proyectos y las publicaciones refieren cómo unas herramientas particulares ayudan a los investigadores a identificar, anotar, extraer y visualizar la información requerida para un caso de estudio. La tarea más frecuente que desempeñan dichos productos es el reconocimiento de entidades (*named entity recognition*, NER): donde se identifican y clasifican segmentos de texto que son nombres de per-

sonas, organizaciones o instituciones, lugares geográficos, expresiones temporales, cantidades, valores monetarios, etc. Si bien es una tarea básica dentro del procesamiento del lenguaje natural, es recurrente en estudios que explotan *corpora* digitales (revistas, periódicos, blogs, tuits, libros, etc.) en diversas disciplinas de las HSC y las DH.
http://es.wikipedia.org/w/index.php?title=Reconocimiento_de_nombres_de_entidades&oldid=65336440

Para la búsqueda de publicaciones se consultan bases de datos y repositorios (*Scopus*, *Dialnet*, *ACL Anthology*, *arXiv.org)*.
http://www.elsevier.com/online-tools/scopus
http://dialnet.unirioja.es
http://www.aclweb.org/anthology
http://arxiv.org

Los proyectos se identificaron en la base de datos *Cordis*, pero también en webs de instituciones y programas de financiación específicos de investigación en HSC:

- *German Federal Ministry of Education and Research* (Alemania);
- *Joint Information Systems Committee* (*JISC*, Reino Unido);
- *National Science Foundation* (Estados Unidos);
- *Economic and Social Research Council* (Reino Unido);
- *National Endowment for the Humanities* (United States);
- *Social Sciences and Humanities Research Council* (Canadá);
- *Digging Into Data Challenge* (de carácter internacional);
- *Ministerio de Economía y Competitividad* (España); y
- *Ministerio de Educación Cultura y Deporte* (España).

En las búsquedas se utilizaron como palabras clave: análisis textual, análisis de texto/de texto asistido por ordenador/automático de texto, datos textuales, reconocimiento/extracción/identificación de entidades, minería de textos/de datos, herramientas informáticas/digitales, software, y aplicaciones informáticas. Todos ellos en combinación con los términos: ciencias sociales, humanidades o humanidades digitales, pero también con nombres específicos de disciplinas (historia, ciencia política, literatura, psicología, etc.).

Dado que el objetivo era permitir que el usuario estableciera un itinerario de información desde el descubrimiento de la herramienta hasta cómo utilizarla, se establecieron como criterios delimitadores:

a) rango temporal (2009-2014) para que fueran investigaciones recientes;

b) que hubiera información explícita sobre el uso de una aplicación informática identificada y se explicara para qué y cómo se había utilizado.

Este segundo criterio fue determinante para desestimar un gran número de trabajos, poniendo de manifiesto que muchos investigadores no mencionan explícitamente esta información en la comunicación de sus resultados.

> El centro *Clarin IULA-UPF* ofrece acceso a herramientas útiles para el análisis y la explotación de textos en forma de servicios web

3. Características del catálogo

La aplicación del modelo LOD a un catálogo genera un producto abierto de información al vincular datos que se encuentran distribuidos en la Web. En el catálogo piloto del centro *Clarin IULA-UPF*, los datos enlazados abiertos, gracias a su estructura de grafo, posibilitan al usuario durante la navegación, por ejemplo, el descubrimiento de un informe final de proyecto donde se usan aplicaciones informáticas para revisar y extraer información de los textos, pero también le permiten acceder a otros trabajos del mismo autor u otros proyectos en los que ha participado al estar estos datos relacionados en forma de enlaces visitables.

En el catálogo se publican datos generados específicamente para el mismo, como la descripción de un proyecto o de una herramienta en particular, pero también se reutiliza información de otros recursos ajenos al establecer relaciones con fuentes externas, por ejemplo utilizando el *Virtual international authority file* (*VIAF*) para identificar un autor, o recurriendo a la base de datos *Faceted digital bibliography and library project* (*DBLP*) para establecer el identificador único de una comunicación a un congreso.
http://viaf.org
http://dblp.l3s.de

En los siguientes apartados se explica con más detalle el contenido del catálogo y los instrumentos de descripción utilizados (metadatos y vocabularios).

Tabla 1. Proyectos seleccionados

N.	Disciplina	Acrónimo	Nombre	Fecha
1	Historia	*Chalice*	*Connecting historical authorities with linked data, contexts and entities*	2010 – 2011
2	Historia	*ChartEx*	---	2011 –
3	Historia	*DbyD*	*Digging by debating*	2012 – 2013
4	Multidisciplinar	*DMCI*	*Data mining with criminal intent*	2010 – 2011
5	Historia	*DPRM*	*Digital prosopography for renaissance musicians*	2013 – 2014
6	Historia	*DVE*	*Dynamic variorum editions*	2010 – 2011
7	Historia	*Isher*	*Integrated social history environment for research*	2012 – 2013
8	Historia	*MRL*	*Mapping the republic of letters*	2009 - ?
9	Historia	*MT*	*Mapping texts*	2010 – 2012
10	----	*Scrutiny*	----	2009 – 2010

3.1. El contenido

El catálogo relaciona información de proyectos, publicaciones y herramientas informáticas[6]. Cabe recordar que los contenidos presentados en este trabajo no son definitivos ya que el catálogo está en constante expansión, las cifras de publicaciones, proyectos y herramientas son efímeras, y en el momento de la consulta real al catálogo la información puede haber cambiado debido a la incorporación periódica de nuevos items.

Proyectos

Incluyen proyectos de investigación que utilizan o han utilizado tecnologías lingüísticas. En esta primera selección se incorporan diez que trabajan con el reconocimiento de entidades con varias finalidades (tabla 1).

A continuación se presentan dos ejemplos:

- ***ChartEx***: mediante el uso de tecnologías de procesamiento del lenguaje natural (NLP), que incluyen el NER (con la herramienta *Brat*), el proyecto propone aumentar el saber sobre la vida doméstica y relaciones entre las personas en la Edad Media a través de la extracción de información sobre lugares, personas o eventos en contratos de compra/venta de propiedades en formato digital.

  http://www.chartex.org
  http://brat.nlplab.org

- ***Isher***: tiene como objetivo investigar la aplicación de software para detectar, vincular y visualizar eventos, tendencias, nombres de personas y organizaciones, así como cualquier otra entidad de interés para la historia social. Utiliza la plataforma *U-Compare* para el procesamiento de textos.

  http://www.nactem.ac.uk/DID-ISHER
  http://u-compare.org

Publicaciones

Incluye información sobre artículos, comunicaciones a congresos, informes, etc., sobre las herramientas y cómo se han utilizado. Para ilustrar la tarea de reconocimiento de entidades se incorporan de momento 14 publicaciones (tabla 2).

A continuación se describen sintéticamente algunas de estas publicaciones[7] a modo de ejemplo:

En el campo de la historia, **Torget** *et al.* (2011) y **Yang**, **Torget** y **Mihalcea** (2011) informan sobre los resultados del proyecto *MT*. En las publicaciones se explica cómo el proyecto explora la combinación de la minería de textos con la cartografía geoespacial para estudiar los periódicos históricos gracias a la frecuencia de palabras, el reconocimiento de entidades con nombre (con *Stanford NER*), y *Topic models*[8] (con *Mallet*).
http://nlp.stanford.edu/software/CRF-NER.shtml
http://mallet.cs.umass.edu

En el campo de la ciencia política **Shor** *et al.* (2014) estudian la cobertura informativa de las mujeres políticas en periódicos de tendencia liberal y conservadora. En este caso se utiliza la identificación de nombres de personas, con el software *Lydia text analysis*[9], como parte del proceso de investigación.

Waila, **Singh** y **Singh** (2013) presentan el análisis de blogs con contenido político, mediante el análisis textual asistido por ordenador en el modelado de temas (*topic models*), el análisis de sentimientos, y el reconocimiento de entidades

Tabla 2. Publicaciones seleccionadas

N.	Disciplina	Título	Autores	Año
1	Ciencias políticas	*Is there a political bias?*	**Eran Shor** *et al.*	2014
2	Cinematografía	*The movie mashup application MoMa*	**Jean-Marc Finsterwald** *et al.*	2012
3	Literatura	*Advanced visual analytics methods for literature analysis*	**Daniela Oelke**; **Dimitrios Kokkinakis**; **Mats Malm**	2012
4	Literatura	*Names in novels*	**Karina Van-Dalen-Oskam**	2013
5	Literatura	*The geographic imagination of civil war era American fiction*	**Matthew Wilkens**	2013
6	Literatura	*From speaker identification to affective analysis*	**Elias Iosif**; **Taniya Mishra**	2014
7	Literatura	*Structure based clustering of novels*	**Mariona Coll-Ardanuy**; **Caroline Sporleder**	2014
8	Literatura/Sistemas información geográfica	*Mapping the English lake district*	**David Cooper**; **Ian Gregory**	2011
9	Literatura/Sistemas información geográfica	*Visual GISting*	**Ian Gregory**; **Andrew Hardie**	2011
10	Historia	*Topic modeling on historical newspapers*	**Tze-I Yang**; **Andrew Torget**; **Rada Mihalcea**	2011
11	Historia	*Data mining with criminal intent*	**Dan Cohen** *et al.*	2011
12	Historia	*Mapping texts*	**Andrew Torget** *et al.*	2011
13	Sociología	*Analizing androcentric focus & deshumanization in news headlines using advanced exploitation tools*	**Marta Villegas**, **Maite Melero**, **Núria Bel**	2011
14	Sociología/Ciencias políticas	*Blog text analysis using topic modeling, named entity recognition and sentiment classifier combine*	**Pranav Waila**; **V. K. Singh**; **M. K. Singh**	2013

nombradas. Utilizan la herramienta *Alchemy API named entity extraction* para reconocer nombres de personas, organizaciones, lugares y fechas.
http://www.alchemyapi.com/products/features/entity-extraction

En literatura y lingüística **Oelke**, **Kokkinakis** y **Malm** (2012) analizan un subconjunto del *Banco literario sueco* centrándose en la extracción de nombres de persona, su género y su forma normalizada, incluyendo menciones de seres teístas (dioses y figuras mitológicas), y examinan su aparición en 13 novelas con la herramienta *Swedish FS* para el reconocimiento de entidades en lengua sueca[10].

Coll-Ardanuy y **Sporleder** (2014) investigan las redes sociales que surgen desde el personaje principal (Elizabeth) de la novela *Orgullo y prejuicio* de Jane Austen. Como primer paso para iniciar el trabajo se propone el reconocimiento de los nombres de los personajes de la novela utilizando la herramienta *Stanford NER*.
http://www.tekstlab.uio.no/nn/foredrag/fefor03/dimitris_fefor03.pdf

Otros autores aplican el análisis geoespacial a los textos literarios. Es el caso de **Wilkens** (2013) que estudia el uso de los lugares geográficos reales en la literatura de ficción norteamericana del siglo XIX. Utiliza también el *Stanford NER* para la identificación y extracción de nombres de lugares geográficos que posteriormente identifica en mapas con sistemas de información geográfica.

Herramientas

Incluyen aplicaciones de tecnologías lingüísticas para el análisis de texto. En esta primera versión del catálogo se decide incorporar solamente servicios web alojados en el centro *Clarin IULA-UPF*, aunque se está valorando la integración de productos de otros proveedores de servicios que también los ofrezcan en acceso libre. Para ejemplificar la aplicabilidad al reconocimiento de entidades se incorporan dos servicios web:

1) *Freeling*: se integran dos módulos en formato servicio web del kit de productos de análisis lingüístico *Freeling* que incluyen el desempeño de la tarea de reconocimiento de entidades nombradas: el *Analizador morfosintáctico Freeling* y el *Etiquetador morfosintáctico Freeling*.
http://ws04.iula.upf.edu/soaplab2-axis/#morphosintactic_tagging.freeling3_morpho_row
http://ws04.iula.upf.edu/soaplab2-axis/#morphosintactic_tagging.freeling3_tagging_row

2) *ContaWords*: aplicación web que analiza las palabras de un archivo de texto, clasificándolas por categorías (nombres, verbos, etc.), asignándoles su correspondiente lema, para después contar su frecuencia de aparición en el texto. También realiza el reconocimiento de entidades con nombre, por ejemplo nombres de personas, de empresas, de organizaciones, geográficos, etc.
http://contawords.iula.upf.edu

> El catálogo relaciona información relativa a proyectos, publicaciones y herramientas informáticas

3.2. Instrumentos de descripción

Metadatos

Para la descripción de los diversos items se adopta una combinación de esquemas de metadatos que permite proporcionar siempre una descripción genérica, pero también una especializada para aquellos casos que la requieran.

En la descripción genérica se utilizan metadatos del conjunto básico de elementos del esquema *Dublin core (DC)*: identificador, autor, colaborador, título, descripción y materia. También del mismo esquema se aplican metadatos de especificación para incluir información sobre la citación bibliográfica (en formato *Vancouver*) y las relaciones entre publicaciones, herramientas y proyectos. Se identifican con el prefijo "dc" y "dcterms" respectivamente.
http://dublincore.org

Debido a que se apuesta por la inclusión de aplicaciones informáticas que no necesitan instalación, en formato de servicio web, se usa el esquema *Biocatalogue* para describirlas Éste es el esquema de metadatos elaborado para la plataforma del mismo nombre donde se describen servicios web del ámbito de las ciencias de la vida. Los metadatos específicos que se han utilizado de este esquema son: versión, tecnología, tarea ejecutada, persona de contacto, proveedor del servicio, dirección web de acceso, dirección de ejecución (*wsdl*), ejemplo de texto de entrada y ejemplo de resultado. Se identifican con el prefijo "bio".

Tabla 3. Vocabularios utilizados

Acrónimo	Nombre	Web
BIBO	*Bibliographic ontology*	*http://bibliontology.com/specification*
BIO	*Biocatalogue*	*https://www.biocatalogue.org*
Faceted DBLP	*Faceted digital bibliography and library project*	*http://dblp.l3s.de/?q=&newQuery=yes&resTableName=query_resultucocuv*
Dbpedia	*Dbpedia*	*http://dbpedia.org/About*
DCMI Types	*DCMI type vocabulary*	*http://dublincore.org/documents/dcmi-type-vocabulary*
FOAF	*Friend of a friend*	*http://www.foaf-project.org*
ISOcat	*Data category registry*	*http://www.isocat.org*
LC/NAF	*Library of Congress name authority file*	*http://id.loc.gov*
VIAF	*Virtual international authority file*	*http://viaf.org*

http://purl.org/ms-lod/BioServices.ttl
https://www.biocatalogue.org

En previsión de la futura incorporación y vinculación con recursos lingüísticos se integra además el esquema de metadatos *MetaShare* (*MS*)[11] para una descripción más acotada de los mismos. Posibles ejemplos de aplicación son: la codificación de caracteres (UTF8, ascii, etc.), forma de creación (manual o automática), nombre del proyecto que ha financiado su construcción, etc. Se identifican con el prefijo "ms".

Vocabularios

Permiten asignar valores a los metadatos seleccionados para la descripción. El detalle de los que se utilizan en el catálogo se muestra en la tabla 3.

> A través de la navegación entre los diferentes elementos interconectados del catálogo el usuario puede establecer un recorrido de consulta que puede ser meramente informativo o dar lugar directamente a la experimentación

4. Ejemplo de consulta del catálogo

La gestión de la información en LOD se realiza con el programa *TopBraid*, gestor de ontologías, que permite la integración de vocabularios y esquemas de metadatos. Todos los datos son almacenados en un punto de acceso *Sparql*. Para facilitar el primer contacto con el catálogo piloto se crea una página introductoria con información contextual sobre sus objetivos, contenidos y características. La consulta del catálogo se implementa en un buscador local.
http://www.topquadrant.com/tools/ide-topbraid-composer-maestro-edition
http://www.w3.org/TR/rdf-sparql-query

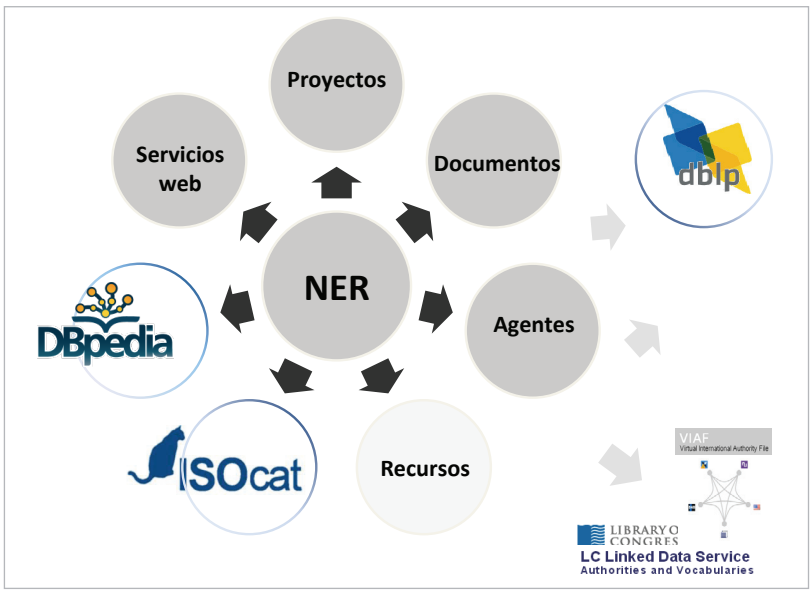

Figura 1. Modelo de vinculación LOD del catálogo

http://ws02.iula.upf.edu/corpus_data/webTest/home.html

El catálogo relaciona información sobre:

- Proyectos (de investigación que utilizan o han utilizado tecnologías lingüísticas);
- Herramientas (servicios web de tecnologías lingüísticas);
- Documentación (artículos, comunicaciones, etc., sobre las herramientas y cómo se han utilizado)[12].

Para acercar la consulta de los materiales a sus potenciales usuarios, facilitar el razonamiento analógico y la utilización de la información, se complementan los elementos anteriores con tres items más:

- Áreas (disciplinas y temas de investigación);
- Agentes (personas y organizaciones);
- Tareas (que realizan las herramientas).

En la versión actual del catálogo la tarea de reconocimiento de entidades nombradas actúa como hilo conductor de la interrelación de elementos (figura 1).

Por esta razón la relación entre los elementos se puede describir de la siguiente forma:

- los documentos explican el uso de *NER* (*named entity recognition*) en investigaciones que realizan análisis de textos;
- los proyectos hacen referencia a la utilización de *NER* en investigaciones de análisis de texto;
- los servicios web ejecutan *NER* sobre los textos que propongan los usuarios;
- los agentes (personas) son autores de documentos, crean servicios, o participan en proyectos;
- los agentes (instituciones) participan en proyectos o representan la filiación de las personas;
- los recursos son utilizados como corpus para detectar y extraer nombres de entidades (relación no implementada).

NER está presente como tema (si se trata de documentos o proyectos) y como tarea si está relacionada con los servicios.

¿Cómo se inicia el itinerario del usuario potencial?

Ya desde la página inicial de la web del centro de competencias se invita al usuario a realizar un itinerario de navegación, marcado por la analogía, al plantear como invitación *¡Mira qué hacen los demás!* como punto de acceso a la entrada al catálogo (figura 2).
http://iula02v.upf.edu/corpus_data/webTest/index.html

Como se ve en la figura 2, el usuario se encuentra con una nube de temas cuya finalidad es que éste se sienta identificado con alguno de ellos debido a su ejercicio profesional, estudios o investigación.

Esta nube de términos también guía su entrada en el catálogo ya que la primera pantalla que se visualiza es una lista de áreas

Figura 2. Web centro *Clarin IULA-UPF*

(en sentido amplio: se incluyen disciplinas pero también temas de investigación) a través de las cuales puede iniciar su itinerario de navegación (figura 3).

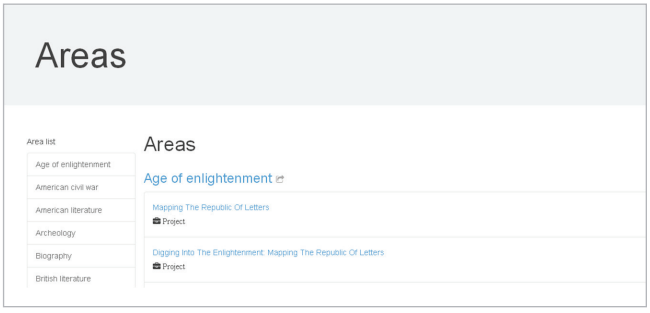

Figura 3. Catálogo: página de acceso por áreas

Una vez en esta página, el usuario puede recorrer la lista de áreas para seleccionar alguna de su interés, como no tener ninguna preferencia y visitar los primeros enlaces que se ofrecen. Ejemplo: selección del área de *American literature* (figura 4).

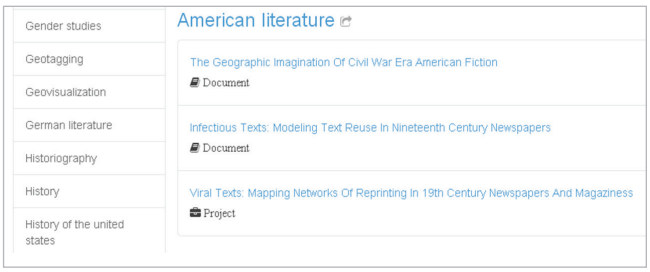

Figura 4. Catálogo: selección de área específica

Bajo cada área se listan documentos y proyectos relacionados. Aquí nuevamente el usuario decide su ruta de navegación entre los resultados que se ofrecen. Ejemplo: se escoge el primer ítem de la lista *The geographic imagination of Civil war-era American fiction* (figura 5).

Figura 5. Catálogo: ejemplo de publicación

A partir de este momento el usuario puede hacer clic en los elementos con hipervínculos (destacados en azul): autor (*creator*), materia (*subject*), identificador (*identifier*), o área (*area*). Hay dos tipos de enlaces:

- Externos: relacionan datos obtenidos en fuentes ajenas y están marcados con un icono en forma de flecha. Si el usuario visita estos enlaces saldrá del catálogo para consultar las fuentes originales de los datos externos (por ejemplo *DBpedia*).
- Internos: aquellos que vinculan información contenida en el propio catálogo (por ejemplo *named entity recognition*). Si el usuario decide visitar este enlace se traslada a otra página del catálogo donde encuentra todos los elementos que fueron etiquetados con dicha tarea (figura 6).

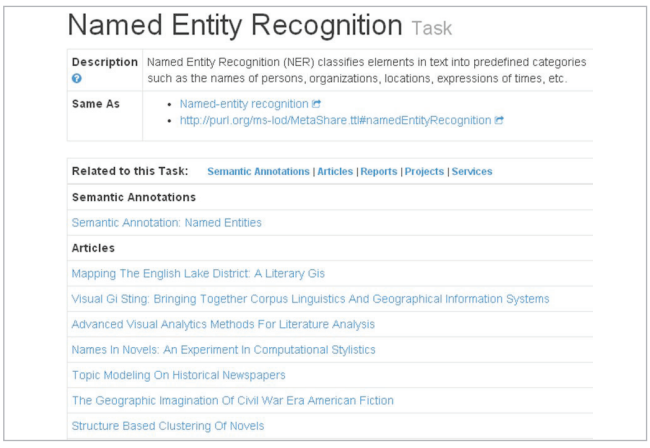

Figura 6. Catálogo: Tarea *named entity recognition*

Llegados a este punto es posible visualizar el potencial de las relaciones establecidas en las informaciones contenidas en el catálogo. Los usuarios tienen acceso a ampliar información sobre la tarea si no la conocen, con información interna del catálogo (definición (*description*) y tipificación como tipo de anotación semántica (*semantic annotation: named entities*), como también pueden profundizar dicha informa-

ción visitando fuentes externas enlazadas como la *DBpedia*. Además pueden ver que esta tarea está relacionada con artículos, informes, proyectos y herramientas para conocer de primera mano otras experiencias que la utilizan en su metodología (en pestañas *Articles, Reports & Projects*), y experimentar con herramientas similares (en pestaña *Services*).

Mediante la navegación entre los diferentes elementos interconectados del catálogo el usuario puede establecer por sí mismo un recorrido de consulta que puede ser meramente informativo o dar lugar directamente a la experimentación.

5. Conclusiones

El uso de tecnología en la práctica de las humanidades y ciencias sociales (HCS) no es una novedad (**Roberts**, 1997; **Popping**, 2000). Sin embargo, la utilización de términos como *humanidades digitales* (**Burdick** *et al.*, 2012; **Liu**, 2012), *culturomics* (**Michel** *et al.*, 2011; **Bohannon**, 2011) o *big data en humanidades, artes y ciencias sociales* (**Parry**, 2010; **Leetaru**, 2012) pone de manifiesto el interés por el uso de la tecnología en la praxis de la investigación en las HSC. Es necesario capacitar a los investigadores para que descubran trabajos existentes que sean referentes para utilizar los programas disponibles en su objeto de estudio.

Disponer de un repositorio de información especializada que incluya material de referencia sobre la utilización de aplicaciones informáticas en la investigación en HSC y DH, capacita a los investigadores para explorar prácticas alternativas a las habituales. Por otra parte, a nivel más técnico, la integración de recursos dispares en una única fuente de información beneficia la sostenibilidad, interoperabilidad e integridad de los datos, lo cual repercute en la optimización de los costos y la funcionalidad. La propuesta de un catálogo basado en LOD permite una explotación de datos simple y eficiente. Por ejemplo, a partir del ejemplo diseñado sobre *NER* (*named entity recognition*), el usuario puede con una simple consulta acceder a publicaciones, proyectos, servicios web, y expertos relacionados con ese tema. Asimismo, el uso de LOD posibilita que el usuario sea el protagonista pudiendo escoger el itinerario de descubrimiento que más le convenga.

Notas

1. *Common Language Resources and Technologies*
http://www.clarin.eu

2. Este trabajo ha sido financiado por el *Fondo Europeo de Desarrollo Regional* (*Feder*), *Programa operativo Feder de Cataluña 2007-2013*, *Objetivo 1*.

3. El centro *Clarin IULA-UPF* se ubica en la sede del *Instituto de Lingüística Aplicada* (*IULA*) de la *Universitat Pompeu Fabra* (*UPF*).

4. *Glossary of Archival and Records Terminology*
http://www2.archivists.org/glossary/terms/a/actionable-information

5. Si se desea formar parte de los grupos de testeo del catálogo envíe un correo electrónico a: *iulatrl@upf.edu*

6. Se está implementando la inclusión de recursos lingüísticos, como corpus y léxicos, aunque todavía no se encuentran relacionados con el resto de la información.

7. La descripción de las herramientas se incluye en un anexo al final del artículo.

8. *Topic models*: en procesamiento de lenguaje natural se denomina así a un modelo estadístico para descubrir los temas implícitos en una colección de documentos a partir del análisis de las palabras que aparecen más frecuentemente en el texto.

9. Aunque la aplicación está citada en diversas publicaciones no se ha podido identificar una dirección web para su consulta.

10. Cabe destacar que en el artículo la herramienta no estaba identificada claramente salvo a través de los investigadores que la desarrollaron, información a partir de la cual se pudo rastrear el nombre de la misma. Dado el interés del tema y la escasez de resultados obtenidos se ha decidido igualmente incluirla en el catálogo. Es una clara muestra de las carencias en la descripción de la metodología en las publicaciones tal y como se remarcaba en el apartado de metodología del presente trabajo.

11. El esquema de metadatos completo está disponible en el manual *Documentation and user manual of the META-Share metadata model*.
http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf

12. Se recuerda que los recursos lingüísticos se contemplan como futura tarea en la vinculación de datos, por lo tanto también aparece en el modelado de la consulta.

6. Bibliografía

Bohannon, John (2011). "*Google books*, *Wikipedia*, and the future of culturomics". *Science*, v. 331, n. 6014, p. 135.
http://dx.doi.org/10.1126/science.331.6014.135

Brier, Alan; **Hopp, Bruno** (2011). "Computer assisted text analysis in the social sciences". *Quality & quantity*, v. 45, n. 1, pp. 103-128.
http://dx.doi.org/10.1007/s11135-010-9350-8

Burdick, Anne; **Drucker, Johanna**; **Lunenfeld, Peter**; **Presner, Todd**; **Schnapp, Jeffrey** (2012). *Digital_humanities*. Cambridge: MIT Press. ISBN: 978 0262018470
http://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf

Cohen, Dan; **Gibbs, Frederick**; **Hitchcock, Tim** *et al.* (2011). *Data mining with criminal intent*. Final white paper.
http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf

Coll-Ardanuy, Mariona; **Sporleder, Caroline** (2014). "Structure-based clustering of novels". En: *Procs of the 3rd Workshop on computational linguistics for literature* (*CLfL*), pp. 31-39.
http://www.aclweb.org/anthology/W/W14/W14-0905.pdf

Cooper, David; **Gregory, Ian** (2011). "Mapping the English Lake District: a literary GIS". *Transactions of the Institute of British Geographers*, v. 36, n. 1, pp. 89-108.

http://dx.doi.org/10.1111/j.1475-5661.2010.00405.x

Finsterwald, Jean-Marc; **Grefenstette, Gregory**; **Law-To, Julien**; **Bouchard, Hugues**; **Djalil-Mezaour, Amar** (2012). "The movie mashup application MoMa: geolocalizing and finding movies". En: *Procs of the ACM multimedia 2012 workshop on geotagging and its applications in multimedia, GeoMM'12*, pp. 15-18.

Gregory, Ian; **Hardie, Andrew** (2011). "Visual GISting: bringing together corpus linguistics and geographical information systems". *Literary and linguistic computing*, v. 26, n. 3, pp. 297-314. *http://dx.doi.org/10.1093/llc/fqr022*

Iosif, Elias; **Mishra, Taniya** (2014). "From speaker identification to affective analysis: a multi-step system for analyzing children's stories". En: *Procs of the 3rd Workshop on computational linguistics for literature* (*CLfL*), pp. 40-49. *http://aclweb.org/anthology/W14-0906*

Juola, Patrick (2008). "Killer applications in digital humanities". *Literary and linguistic computing*, v. 23, n. 1, pp. 73-83. *http://www.mathcs.duq.edu/~juola/papers.d/killer.pdf http://dx.doi.org/10.1093/llc/fqm042*

Leetaru, Kalev H. (2012). "A big data approach to the humanities, arts, and social sciences: *Wikipedia's* view of the world through supercomputing". *Research trends*, n. 30, September. *http://www.researchtrends.com/issue-30-september-2012/a-big-data-approach-to-the-humanities-arts-and-social-sciences-summary*

Liu, Alan (2012). "The state of the digital humanities: a report and a critique". *Arts and humanities in higher education*, v. 11, n. 1-2, pp. 8-41. *http://dx.doi.org/10.1177/1474022211427364*

Michel, Jean-Baptiste; **Kui-Shen, Yuan**; **Presser-Aiden, Aviva**; **Veres, Adrian**; **Gray, Matthew K.**; **The Google Books Team**; **Pickett, Joseph P.**; **Hoiberg, Dale**; **Clancy, Dan**; **Norvig, Peter**; **Orwant, Jon**; **Pinker, Steven**; **Nowak, Martin A.**; **Liebermann-Pinker, Steven** (2011). "Quantitative analysis of culture using millions of digitized books". *Science*, v. 331, n. 6014, pp. 176-182. *http://scholar.harvard.edu/files/pinker/files/michel_et_al_quantitative_analysis_of_culture_science_2011.pdf http://dx.doi.org/10.1126/science.1199644*

Oelke, Daniela; **Kokkinakis, Dimitrios**; **Malm, Mats** (2012). "Advanced visual analytics methods for literature analysis". En: *Procs of the 6th Workshop on language technology for cultural heritage, social sciences, and humanities*, pp. 35-44. *http://dl.acm.org/citation.cfm?id=2390357.2390364*

Parry, Marc (2010). "The humanities go *Google". The chronicle of higher education*, May 28. *http://chronicle.com/article/The-Humanities-Go-Google/65713*

Popping, Robert (2000). *Computer-assisted text analysis*. London: Sage. ISBN: 0761953795

Roberts, Carl W. (ed.) (1997). *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Mahwah: Lawrence Erlbaum. ISBN: 0805817352

Schreibman, Susan; **Hanlon, Anne** (2010). "Determining value for digital humanities tools: report on a survey of tool developers". *Digital humanities quarterly*, v. 4, n. 2. *http://digitalhumanities.org/dhq/vol/4/2/000083/000083.html*

Shor, Eran; **Van-de-Rijt, Arnout**; **Ward, Charles**; **Askar, Saoussan**; **Skiena, Steven** (2014). "Is there a political bias? A computational analysis of female subjects' coverage in liberal and conservative newspapers". *Social science quarterly*. *http://dx.doi.org/10.1111/ssqu.12091*

Torget, Andrew J.; **Mihalcea, Rada**; **Christensen, Jon**; **McGhee, Geoff** (2011). *Mapping texts: combining text-mining and geo-visualization to unlock the research potential of historical newspapers. A white paper for the National Endowment for the Humanities. http://mappingtexts.stanford.edu/whitepaper/MappingTexts_WhitePaper.pdf*

Van-Dalen-Oskam, Karina (2013). "Names in novels: an experiment in computational stylistics". *Literary and linguistic computing*, v. 28, n. 2, pp. 359-370. *http://dx.doi.org/10.1093/llc/fqs007*

Villegas, Marta; **Melero, Maite**; **Bel, Núria** (2014). "Metadata as linked open data: mapping disparate XML metadata registries into one RDF/OWL registry [forthcoming, accepted paper]". En: *Ninth intl conf on language resources and evaluation* (*LREC-2014*), 26-31 May 2014, Reykjavik, Iceland. *http://www.lrec-conf.org/proceedings/lrec2014/pdf/664_Paper.pdf*

Waila, Pranav; **Singh, V. K.**; **Singh M. K.** (2013). "Blog text analysis using topic modeling, named entity recognition and sentiment classifier combine" En: *Procs advances in computing, communications and informatics*, pp. 1166-1171. *http://dx.doi.org/10.1109/ICACCI.2013.6637342*

Wiedemann, Gregor (2013). "Opening up to big data: computer-assisted analysis of textual data in social sciences". *Forum qualitative sozialforschung/Forum: qualitative social research*, v. 14, n. 2. *http://www.qualitative-research.net/index.php/fqs/article/view/1949*

Wilkens, Matthew (2013). "The geographic imagination of Civil war-era American fiction". *American literary history*, v. 25, n. 4, pp. 803-840. *http://dx.doi.org/10.1093/alh/ajt045*

Yang, Tze-I; **Torget, Andrew J.**; **Mihalcea, Rada** (2011). "Topic modeling on historical newspapers". En: *Procs of the 5th ACL-HLT Workshop on language technology for cultural heritage, social sciences, and humanities*, pp. 96-104. *http://dl.acm.org/citation.cfm?id=2107636.2107649*

7. Anexo. Descripción de herramientas

Se muestran breves descripciones de las aplicaciones y/o plataformas utilizadas tanto en los proyectos como en las publicaciones seleccionadas para la presente versión del catálogo. Éstas no se incluyen de momento en el catálogo. Todas permiten ejecutar la tarea de reconocimiento de entidades.

Alchemy API named entity extraction: interfaz de programación de aplicaciones (*API, application programming interface*) para la identificación de nombres de personas, empresas, organizaciones, ciudades, accidentes geográficos, etc., en textos digitales o webs. Emplea algoritmos estadísticos y tecnología de procesamiento del lenguaje natural para analizar y extraer la información. Es una aplicación comercial y dispone de una versión de prueba.

Brat: aplicación web para crear anotaciones para reconocimiento de entidades y relaciones binarias útiles para tareas de extracción de información, entre otras.

Lydia text analysis*:* sistema de procesamiento de lenguaje natural para el análisis y la extracción de información especializado en medios de comunicación online (periódicos, blogs, etc.). Permite relacionar personas, lugares y temas a partir del procesamiento del lenguaje natural y el análisis estadístico de frecuencias nombres de entidades y colocaciones. Actualmente está integrado en el buscador especializado en entidades *Textmap*. *http://www.textmap.com/index.htm*

Mallet: paquete de aplicaciones *java* para procesamiento estadístico del lenguaje natural, clasificación de documentos, agrupamiento (*clustering*), modelado de temas (*topic models*), extracción de información, y otras aplicaciones de aprendizaje automático para textos.

Swedish FS: aplicación para el reconocimiento de entidades específica para la lengua sueca. Está desarrollada sobre la base de gramáticas de estado finito sensitivas al contexto (*context-sensitive finite-state grammars*). No se ha encontrado web específica.

Stanford NER: aplicación *java* para el reconocimiento de nombres de entidad.

U-Compare: sistema integrado de minería de texto y procesamiento del lenguaje natural desarrollado en entorno *UIMA* (*Unstructured information management applications*, entorno especializado en análisis de grandes volúmenes de información no estructurada). Se compone de un conjunto de módulos interoperables y flujos de trabajo a través de una interfaz única. Los módulos que incluye son: lectores de corpus, visualizadores, editores de texto, y herramientas sintácticas, semánticas (donde se incluye el reconocimiento de entidades), y estadísticas.