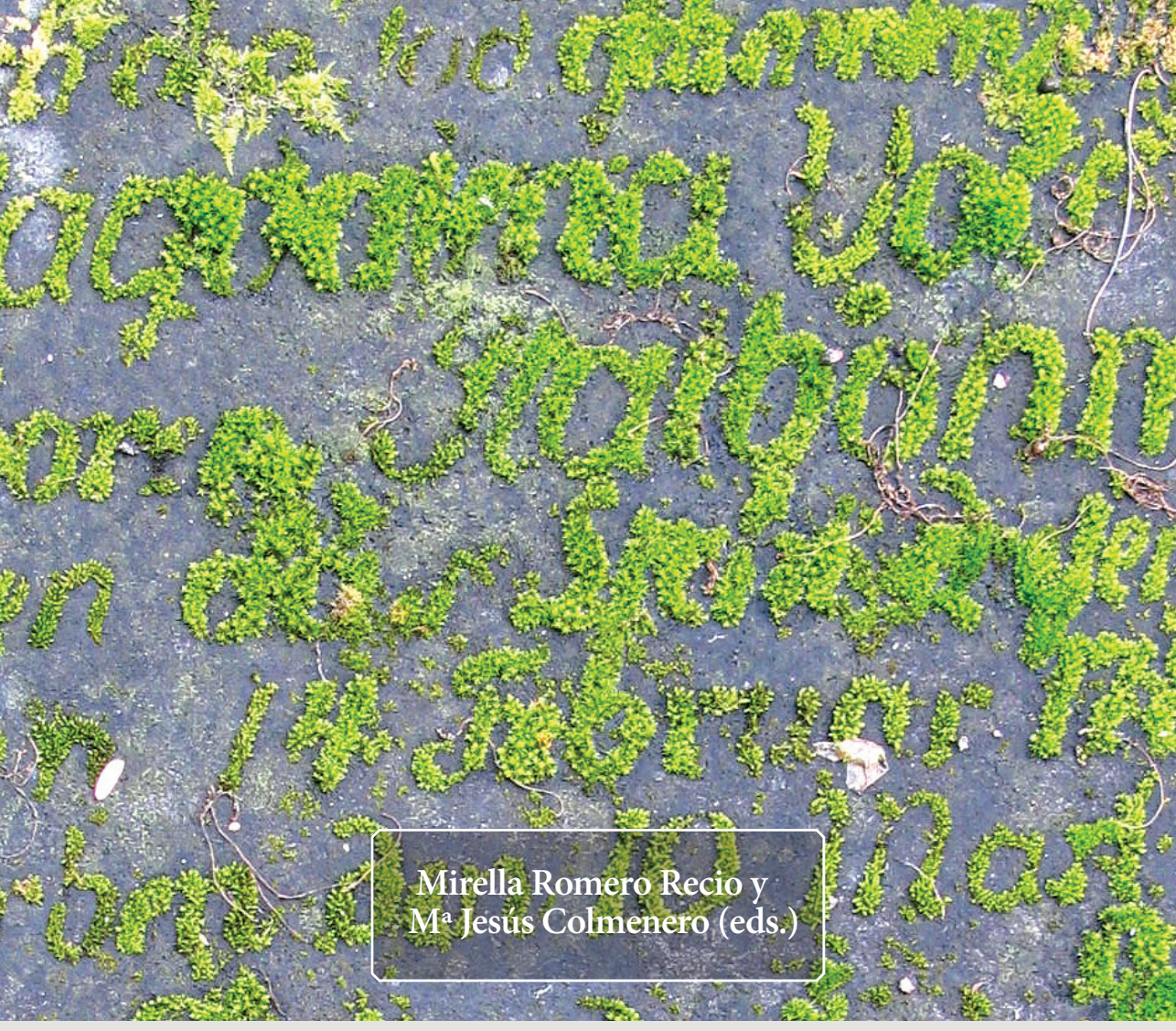


Mirella Romero Recio y
Mª Jesús Colmenero (eds.)

Historiografía digital

proyectos para almacenar y construir la Historia

Anejos de la Revista de Historiografía nº4

Anejos de la Revista de Historiografía nº4

Directora
Mirella Romero Recio

ISBN 978-84-16829-01-9
Depósito legal M-34989-2016

mail
revhisto@uc3m.es

Diseño y maquetación
Syntagmas (www.syntagmas.com)

Madrid, 2016

Instituto de Historiografía
Julio Caro Baroja

Universidad Carlos III
de Madrid

VIII

Aplicación de linked data y la web semántica a la investigación histórica[1]

Jorge Morato
Universidad Carlos III de Madrid

1. Introducción

Las Humanidades Digitales agrupan un conjunto de herramientas que tratan de disminuir la carga de trabajo y agilizar las tareas de investigación mediante métodos tecnológicos. Un subdominio de esta área es la denominada Historia Digital (*Digital History* o e-historia). La Historia Digital trata de emplear las tecnologías de la información para facilitar el estudio de la historia[2]. Una parte relevante de la Historia Digital consiste en unir datos mediante relaciones significativas. Existe un creciente número de investigaciones que tratan de descubrir y evidenciar nuevos hechos mediante la conexión con otros conjuntos de datos. En este documento se revisarán y analizarán diversos enfoques utilizados en investigaciones en Historia Digital.

1. Este trabajo se inscribe en el proyecto de investigación «El Almacén de la Historia. Repositorio de historiografía española (1700-1939)» (HAR2011-27540), financiado por el MINECO.

2. Victor de Boer et al. "Linking the Kingdom: Enriched Access to a Historiographical Text", En: *Proc. of the 7th International Conference on Knowledge Capture K-CAP '13*, Nueva York, June 23-26, 2013, 17-24.

Las fases en los proyectos de Historia Digital, a grandes rasgos, son:

1. Digitalización de documentos con valor histórico, lo cual implica una heurística para su localización y un análisis crítico sobre su valor.
2. Codificar la información de los documentos de forma que se favorezca su reuso y localización.
3. Publicación para que los investigadores accedan a los documentos

2. Métodos de investigación en historia

Como en otras metodologías, en historia los criterios de calidad y análisis en la recopilación de fuentes primarias y secundarias es crítico. Consideración que incluye a memorias, diarios, correspondencia, registros y censos o material iconográfico y oral. Dependiendo de las fuentes existirán diversos factores a analizar[3]. Así en las fuentes primarias se debería tener en consideración el cuándo, dónde, el qué, las entidades nombradas (personas, grupos u organizaciones), análisis de la procedencia del material del documento, su integridad como documento o su credibilidad y autenticidad[4]. Además de otros factores como: el cómo, el porqué, y su transcendencia y contribución para explicar determinado hecho. En otros tipos de fuentes, como las secundarias, es preciso considerar elementos adicionales, tales como el contexto del autor y las circunstancias y destinatarios a los que fue dirigido su estudio. Por último, hay que considerar que diarios y correspondencia pueden tener sesgos acusados debidos al contexto personal, social y cultural del autor. El discurso oral también puede ser fuente de grandes sesgos como argumenta Keightley[5].

La investigación comparativa examina sucesos históricos para extraer teorías que trasciendan el contexto específico en el que estos hechos tuvieron lugar. Básicamente estas teorías identifican la causalidad mediante la comparación con otras situaciones y hechos, bien sean pasados o actuales. Frecuentemente su aplicación se ha centrado en analizar como las sociedades tienden a evolucionar[6]. Kiser y Hechter[7] han puesto de relieve la necesidad de una metodología general, aceptada por la co-

3. Martha Howell y Walter Prevenier, *From Reliable Sources: An Introduction to Historical Methods*, Cornell University Press, Ithaca, 2001.

4. Gilbert J. Garraghan, *A Guide to Historical Method*, Fordham University Press: New York, 1946, 168.

5. E. Keightley, "Engaging with memory", in M. Pickering (ed.), *Research Methods for Cultural Studies*, Edinburgh: Edinburgh University Press, 2008, 175–192.

6. Dennis Smith, *The rise of Historical Sociology*. Philadelphia: Temple University Press, 1991.

7. Edgar Kiser y Michael Hechter, "The Debate on Historical Sociology: Rational Choice Theory and Its Critics", *American Journal of Sociology*, 104, 3, 1998, 785-816.

munidad, para realizar estos estudios. El objetivo de esta metodología general sería descubrir mecanismos de causalidad genéricos, aprovechar la acumulación de conocimiento entre diferentes dominios, revelar anomalías y responder a nuevos interrogantes. De acuerdo a estos autores la prevalencia de la inducción sobre la deducción dificulta la identificación correcta de los principios que subyacen a las relaciones y mecanismos causales. La correcta identificación de estas causas permitiría validar las conclusiones de las investigaciones. En la misma línea, De Boer[8] anima a crear metodologías genéricas para la investigación histórica, para así tratar el modelado, interconexión y acceso a los datos históricos en la web. Su propuesta recomienda específicamente evitar el desarrollo de herramientas que sean específicas de un conjunto de datos concretos, lo cual dificultaría su su reutilización futura.

Shutt[9] identificó varias etapas en la investigación comparativa: 1) definir una hipótesis de trabajo sobre determinado proceso observado; 2) seleccionar y acotar sucesos históricos que puedan mostrar cómo se desarrolla el proceso; 3) comparar y analizar similitudes y diferencias entre los sucesos históricos; 4) proponer una teoría de cómo el proceso observado se produce. La dificultad suele deberse a diferenciar entre causalidad y correlación de variables, por lo que el factor temporal en el análisis es relevante. Otros inconvenientes para realizar el análisis correcto pueden deberse a que los datos históricos son en esencia incompletos, y frecuentemente inconsistentes. Esto, junto al contexto del que los produjo y del que lo analiza produce frecuentemente sesgos.

La investigación comparativa precisa de métodos de consulta de múltiples fuentes y de su comparación. De Boer[10] ha establecido una serie de pasos necesarios para realizar estas tareas en el entorno web: el historiador debe estar presente en la formalización de los datos, para asegurar que los datos son correctos y que están correctamente modelados y enlazados. Si el paso anterior se ha realizado convenientemente, el historiador debe poder: acceder a repositorios de datos heterogéneos para comprender y encontrar anomalías en los datos; realizar consultas no predefinidas en la aplicación y que sean relevantes a su investigación; poder verificar los resultados; analizar los datos con sus propias herramientas; y por último, poder reusar y compartir los datos.

Gracias a la investigación comparativa podremos corroborar la hipótesis a aceptar. Es decir se pueden articular medios para corroborar que la afirmación es compatible con otros hechos ya corroborados, y se puede ver que la hipótesis complementa

8. Victor De Boer, 2015 "Linked data for digital history presentation for VU symposium «Connecting data for research»". Disponible en http://es.slideshare.net/vdeboer/. [Consulta: 16/01/2016].

9. R. K. Schutt, "Investigating the Social World: The Process and Practice of Research". 7ed. London: SAGE, 2011

10. De Boer, *loc. cit*.

y es congruente con otras previas. Pautas ya expuestas por McCullagh[11]. En este contexto la web y las herramientas informáticas pueden ser unos aliados competentes, ya que permiten el acceso y la interoperabilidad con un mayor rango de recursos. La utilización de técnicas estadísticas, tanto en el análisis de los datos como en la validación de la hipótesis, debe ser siempre considerada[12].

Entre los tipos de herramientas informáticas a utilizar AHRC[13] indica los siguientes grupos:

- Sistemas de escaneado y OCR
- Sistemas de estructuración de documentos
- Sistemas de consulta y recuperación
- Minería de datos
- Bases de datos
- Herramientas estadísticas: frecuentemente relacionadas con análisis de regresión, clasificadores y series temporales.
- Visualización
- Sistemas de Información Geográfica

En 2014, Daniel Alves[14] en la publicación "*Digital Methods and Tools for Historical Research: A Special Issue*" destaca la necesidad de sistemas de procesamiento del lenguaje, el modelado y la codificación reusable, y la validación mediante crowdsourcing.

3. La Web Semántica

Como ya se ha argumentado en la sección de métodos, la comparación de información procedente de distintas fuentes depende de ciertos compromisos por normalizar la publicación de estas fuentes. Así, unos requisitos mínimos son: que se comparta un mismo medio de comunicación; un vocabulario compartido que permita designar los recursos agrupados bajo un mismo concepto; compartir una misma sintaxis en la expresión permite buscar relaciones que asocien estos conceptos; por último, compartir un mismo sistema de organización del conocimiento que permita concretar y contextualizar el significado de los conceptos. Los sistemas más populares son tesauros y ontologías.

11. C. Behan McCullagh, *Justifying Historical Descriptions*, Cambridge University Press: New York, 1984, p.19.

12. AHRC ICT Methods Network (2007) "Tools and Methods for Historical Research". Disponible en: www.methodsnetwork.ac.uk. [Consulta:15/01/2016]

13. AHRC, *loc. cit.*

14. Daniel Alves. "Guest Editor's Introduction: Digital Methods and Tools for Historical Research", *International Journal of Humanities and Arts Computing,* 8, 1, 2014, 1-12.

Para lograr estos objetivos, Tim Berners-Lee *et al.*[15] propusieron una red en la que la información y los servicios estuvieran expresados semánticamente, de esta manera las peticiones de otros ordenadores y usuarios podrían ser entendidas y satisfechas. Los ordenadores así podrían analizar todos los datos de la Web: los contenidos, los enlaces, y las transacciones entre personas y las computadoras. Esta Web, que denominaremos semántica, permite la interconexión entre repositorios para comercio electrónico, la realización de consultas semánticas y la implementación de sistemas pregunta-respuesta. En resumen, la Web Semántica es una forma consensuada de especificar datos y relaciones. Esta Web provee de conceptos con una definición pública, un medio de compartir datos y mecanismos de mejora en la integración de datos. El objetivo es mejorar la combinación y navegación entre conceptos de distintas fuentes de datos.

La propuesta tiene una década de antigüedad y se han realizado grandes inversiones para implementarlas desde organismos públicos y privados. Actualmente sus logros son crecientes, aunque llegar a la situación actual ha necesitado más de una década de grandes inversiones, por otro lado su aprovechamiento es aún escaso en relación al esfuerzo invertido.

Tim Berners-Lee[16] definió los siguientes principios para asegurar que la publicación de datos en la web fuera accesible:

- Identidad: Cada concepto debe estar descrito en un recurso localizable que permita identificarlo unívocamente y consultarlo. Este localizador se denomina URI, es decir, las URIs indican la localización digital de un recurso en una red como Internet.
- Accesibilidad: Acceso a los recursos mediante el protocolo HTTP
- Estructura: la información tiene que tener una sintaxis estandarizada, de forma que permita una forma normalizada de expresar y recuperar la información, en la práctica en la web semántica se utilizan los estándares RDF y SPARQL
- Navegación: los recursos deben permitir la navegación conceptual, esto se logra mediante enlaces entre los recursos a través de sus URIs.

La gran ventaja de la adopción de este enfoque es el acceso masivo a recursos semánticos y la mejora de la interoperabilidad, es decir, la posibilidad de desarrollar y reusar aplicaciones para acceder a estos recursos semánticos.

15. T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web", *Scientific American*, 2001, 29-37

16. T. Berners-Lee, (2006) "Linked Data - Design Issues". Disponible en: https://www.w3.org/DesignIssues/LinkedData.html. [Consulta: 15/01/2016]

La Web Semántica es, en esencia, un sistema universal de intercambio de información[17]. Esto solo es posible si los documentos se estructuran y se codifican de una forma interoperable semánticamente. Este cambio de codificación puede convertir a la Web en una gran base de datos. La interoperabilidad entre los documentos puede conseguirse mediante un lenguaje de codificación del conocimiento, como RDF (Resource Description Framework)[18].

La sintaxis de RDF estructura el conocimiento en afirmaciones simples, formadas por una tripleta recurso-atributo-valor. Un ejemplo podría ser <El Quijote> <dc:creator> <Miguel de Cervantes>. Donde el prefijo *dc* designa la URI [http://dublincore.org/documents/dces/], donde está definido el significado de *creator*. Cualquier elemento de la tripleta puede ser una URI, pero en el caso del último elemento, valor, puede ser también texto, fechas, número, etc. Estas tripletas pueden estar enlazadas entre si y permiten una gran flexibilidad para expresar el conocimiento.

El segundo elemento de la tripleta, atributo, esta frecuentemente asociado a un tipo de recurso denominado *Element Set*[19] (Conjunto de Elementos), y es básicamente el nombre de los metadatos. Por otro lado, el tercer elemento suele estar relacionado con el valor que toman estos metadatos y se denomina *Value Vocabulary* (Vocabulario de Valor). En la siguiente tabla 1 se muestran algunos *Element Sets* y *Value Vocabularies* populares.

Element Sets	Value Vocabularies
Dublin Core	LCSH
FRBR	AAT
MARC21	VIAF
FOAF	GeoNames
SKOS	DBpedia
	Freebase

Tabla 1. Element Sets y Value Vocabularies populares.
Fuente: Elaboración propia

17. Y. Kalfoglou., "Knowledge society arguments revisited in the semantic technologies era". *Int. J. of Knowledge and Learning*, 3, 2/3, 2007, 225 – 244

18. D. Allemang, J. Hendler, "RDF –The basis of the Semantic Web", In *Semantic Web for the Working Ontologist,* 2nd Ed, 2011.

19. W3C (2011) "Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets de la W3C". Disponible en: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/. [Consulta: 15/01/2016].

Gracias a la interoperabilidad es factible reutilizar el conocimiento, navegar conceptualmente y enlazar recursos de un modo semántico.

En la medida en que cada elemento, que puede aparecer en la tripleta, está descrito en una URI pública se hace más reusable e inteligible. Estos elementos pueden utilizarse y estructurarse de varias formas. Así, si se utilizan para dar una información adicional a un recurso o dato se denominan metadatos. Si se utilizan para navegar conceptualmente entre distintos conceptos estaremos en presencia de un sistema de organización del conocimiento (o KOS, por sus siglas en inglés). Un KOS muestra una representación de un área del conocimiento, y permiten una notable mejora en la navegación conceptual y en la recuperación de conceptos. Ejemplos son los tesauros y las ontologías. Existe un vocabulario de metadatos específico para expresar las relaciones entre conceptos en un sistema de organización del conocimiento denominado SKOS. Estas relaciones pueden ser desde conceptos equivalentes, más genéricos o simplemente relacionados. El vocabulario SKOS tiene una fuerte presencia en proyectos de humanidades digitales.

4. Linked Data y LOD

En los últimos años, la iniciativa de mayor interés en la Web Semántica es el proyecto *Linked Open Data* (LOD). El adjetivo open (abierto) se refiere a que son datos accesibles a cualquier usuario, permitiendo un trabajo colaborativo donde conceptos y recursos son enlazados de múltiples formas.

LOD se basa en la propuesta de Berners-Lee de publicar datos enlazados en la web mediante el uso de URIs. De esta manera se permite nombrar entidades y establecer enlaces entre recursos mediante sentencias RDF. Posteriormente estas tripletas RDF pueden ser recuperadas con un lenguaje de interrogación como SPARQL. Este lenguaje posee una gran potencia de recuperación de información gracias a la combinación de tripletas de diferentes recursos.

En 2011 el proyecto LOD ya contaba con 295 conjuntos de datos (datasets), con más de 31 millones de tripletas y diferentes recursos interrelacionados, caracterizándose por las enormes dimensiones de sus componentes. Por ejemplo, en 2016 FreeBase está formada por 58 millones de conceptos y tres mil millones de afirmaciones. Otro vocabulario bien reconocido es DBPedia, que contiene contiene 4,58 millones de conceptos, incluyendo 1,4 millones de personas o 735.000 lugares.

El elevado número de conceptos interrelacionados hace patente la necesidad e importancia de un método intuitivo para organizar, recuperar, filtrar, visualizar y navegar entre conceptos y acceder a sus recursos vinculados. En el estudio de Morato *et al.* se recogen y sintetizan diversas propuestas de la recuperación semántica.

5. La investigación histórica en proyectos de humanidades digitales

El incremento de conexiones entre datos permite una mayor compresión de los hechos estudiados. El proceso por el que los datos pasan a representar información y establecer pautas que nos den conocimiento se denomina Jerarquía DIKW (ver Figura 1). Es debido a la importancia de este número de conexiones entre los recursos del proyecto y otros datos donde reside la importancia de encontrar relaciones con otros recursos en los proyectos digitales.

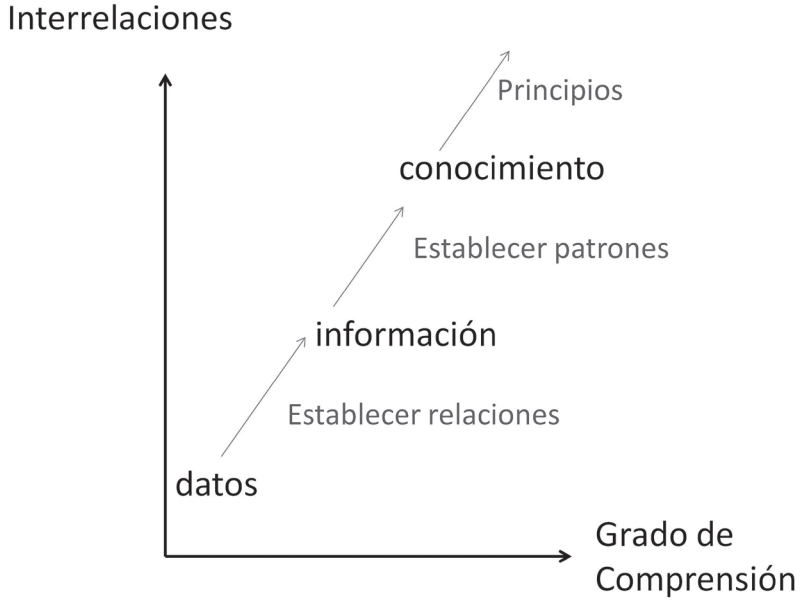


Figura 1. Jerarquía DIKW.
Fuente: http://www.systems-thinking.org/dikw/dikw.htm

Los pasos que hay que dar para que esta interrelación sea congruente pasan por por diferentes etapas (Figura 2):

- Definición del proyecto
- Adquisición de documentos externos, escaneado y OCR, limpiando los errores en el proceso
- Descripción con metadatos para su integración posterior con otros recursos o datos externos.
- Análisis e interpretación

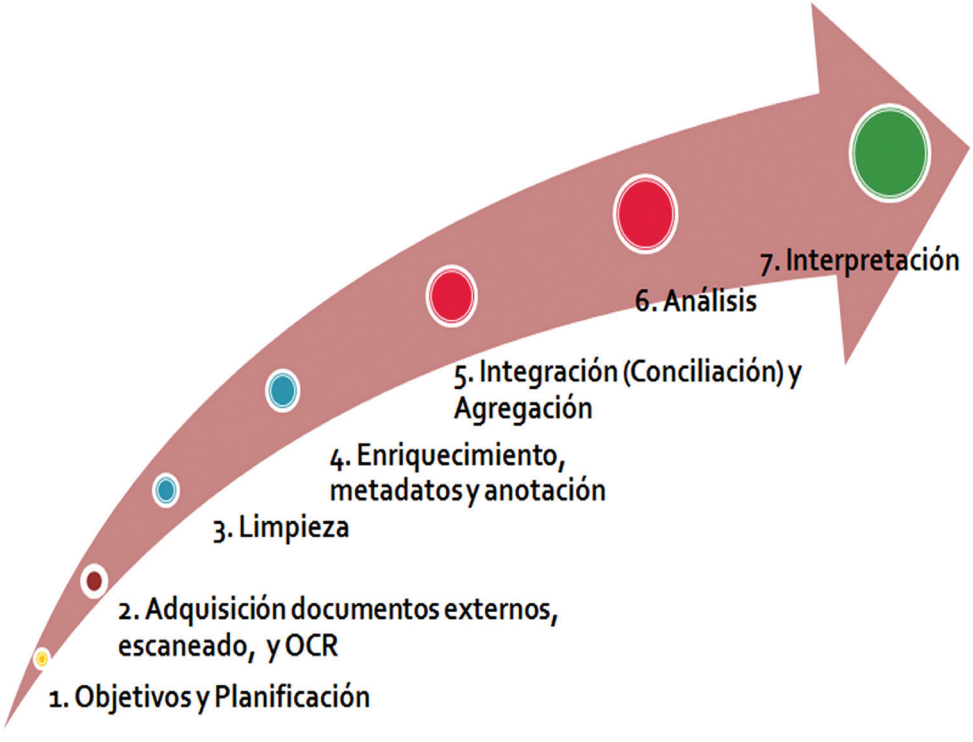


Figura 2. Pasos en el proceso de un proyecto en historia digital. Fuente:Elaboración propia

5.1. Objetivos y planificación

Un ejemplo sobre la importancia de la planificación adecuada se puede ver en los ejemplos a preguntas a contestar en el proyecto WW1LOD[20]. El proyecto LOD recopila documentos y hechos y entidades relacionadas con la Primera Guerra Mundial en varios países europeos. Ejemplos de estas consultas son:

- Trabajos que relacionen al General Friedrich Adolf Julius von Bernhardi con la Primera Guerra Mundial
- Documentos de Europeana, de la colección WW1 y del Proyecto Out of the Trenches que se relacionen con el Oeste de Flandes
- Unidades del tercer ejército alemán que cometieron crímenes de guerra en Bélgica
- Cambios demográficos en las provincias belgas durante los años de la guerra, en relación al número de atrocidades sufridos durante esos años.

La mayoría de estas consultas implican una planificación a priori que establezca qué otros recursos son necesarios incluir en el proyecto, y de qué forma se interrelacionaran los datos de ambos recursos.

5.2. Adquisición y limpieza

En el proceso de escaneado y reconocimiento de caracteres (OCR), en los de tabulación de datos y en los de importación de registros se producen errores. Estos errores deben de ser depurados automática o manualmente.

Existen herramientas que permiten eliminar los errores tipográficos de forma asistida como[21]:

- Similitud fonética con otro dato (p.e. algoritmo *Metaphone*)
- Semejanzas entre dos cadenas de texto (p.e. el algoritmo de *Levehenstein o n-grams*)

5.3. Enriquecimiento e integración

Una de las principales necesidades para la investigación comparativa es la interrelación semántica de nuestros datos con otros recursos para su descripción (metadatos), o de datos con otros conceptos equivalentes de otros proyectos (Figura 3). La

20. WW1LOD. Disponible en: http://www.ldf.fi/dataset/ww1lod/

21. Bertrand Lisbach, Victoria Meyer. "Linguistic Identity Matching. Germany: Springer, 2013.

consecuencia es que los datos del proyecto se habrán enriquecido con nuevos datos, permitiendo una recuperación más amplia.

El primer problema para unir nuestros datos a otros proyectos es identificar estos otros proyectos de interés. Existen grandes repositorios como el *datahub*[22] , catálogos de recursos como Almahisto[23] o buscadores semánticos como *Sindice*[24].

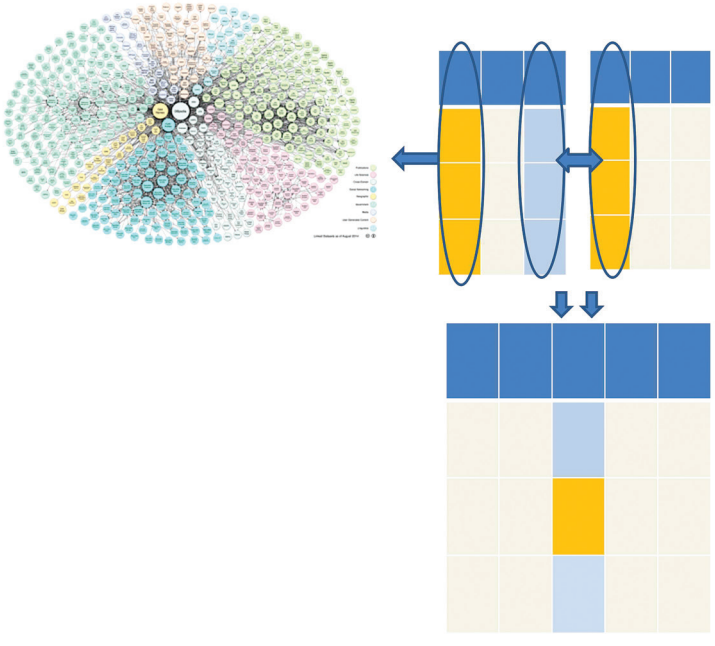
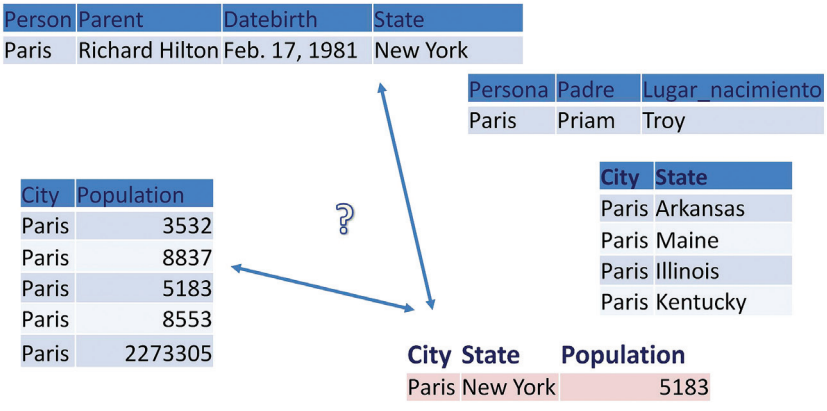


Figura 3. Los recursos deben haber sido descritos conceptualmente para su relación con conceptos externos afines

Los principales problemas para hacer esta fusión de una forma automatizada son la ambigüedad léxica, la homonimia y la polisemia. Pero existe un problema adicional, aún manualmente la falta de contexto puede dar lugar a relaciones erróneas. En la Figura 4 se muestra un ejemplo con el término París. La confusión viene dada

22. Datahub https://datahub.io/
23. Proyecto Almahisto. https://almahisto.wordpress.com/
24. Sindice http://www.sindice.com/

porque París pueden ser ciudades distintas y nombres de personas ficticias o reales. Adicionalmente, en la fusión hay que tener en mente que los recursos en Internet pueden estar en múltiples idiomas.



País	Ciudad	Estado	Poblacion
Estados Unidos	Paris	Arkansas	3532
Estados Unidos	Paris	Maine	5183
Estados Unidos	Paris	Illinois	8837
Estados Unidos	Paris	Kentucky	8553
France	Paris	Paris	2273305

FIGURA 4. RELACIONES ERRÓNEAS ENTRE CONCEPTOS

Un segundo problema, existe la problemática asociada a la lentitud de relacionar dos vocabularios pertenecientes a recursos distintos. De nuevo, este proceso se puede hacer de forma asistida por herramientas informáticas[25]. Una de las estrategias usuales es identificar nombres propios y otros sustantivos que logren identificar el contexto. La fusión se hace comparando cadenas de texto comunes (frases o palabras)[26]. Dado que existen términos homógrafos es frecuentemente necesario utilizar Sistemas de Organización del Conocimiento que acoten el dominio semán-

25. S. van Hooland, R. Verborgh, M. De Wilde, J. Hercher, E. Mannens, and R. Van de Walle. "Evaluating the success of vocabulary reconciliation for cultural heritage collections". *Journal of the American Society for Information Science and Technology (JASIST),* 64, 3, 2013, 464–479.

26. G. Weikum, J. Hoffart, N. Nakashole, M. Spaniol, F. Suchanek, and M. Yosef. Big data methods for computational linguistics. IEEE Data Eng. Bulletin, 35(3), 2012.

tico mencionado en el documento. Otra aproximación es mediante crowdsourcing. En esencia el crowdsourcing trata de recabar la ayuda de usuarios de Internet, retribuida o no, para realizar tareas sencillas, como pueden ser la corrección de errores tipográficos y la interrelación de recursos.

5.4. Análisis y Evaluación

Cómo se ha comentado el objetivo del enriquecimiento e integración es poder responder a preguntar y hacer comparaciones con otros recursos. Estas preguntas y comparaciones son consultables gracias al proceso de enriquecimiento semántico que se ha llevado a cabo. Las consultas deben de poderse traducir a un lenguaje de interrogación como SPARQL.

La división de la información en tripletas permite una comparación entre hechos que difícilmente se podría realizar a gran escala sin la asistencia de proyectos de historia en la Web Semántica.

Para realizar esta comparación, la respuesta a la consulta debe de poder analizarse, bien por medios estadísticos (estadística descriptiva, clasificadores, etc), bien mediante el análisis crítico sintético de los datos.

6. Análisis de algunos proyectos de Historia Digital

En esta sección se analiza con que recursos suelen estar enlazados los proyectos de Historia Digital. Un problema que surge a la hora de interrelacionar recursos es elegir el vocabulario de metadatos adecuado. De entre los miles de vocabularios disponibles en la LOD existe una dificultad en escoger el que mejor cubra las necesidades del proyecto, siendo al tiempo popular entre la comunidad. Existen dos estrategias que se pueden seguir: 1) estudiar qué metadatos se utilizan en proyectos similares, o si no existieran o fueran deficientes, expandir el que más se adapte con los conceptos que falten (creando un perfil de metadatos), 2) buscar en un repositorio de metadatos, como p.e. con el buscador LOV[27].

Existen esquemas más o menos estandarizados. Los esquemas son, básicamente, conjuntos de elementos a describir de manera predefinida y vocabularios de metadatos que se deben emplear para completar estos elementos. Algunos como EDM de Europaena tienen cientos de campos, existen otros como MIDAS, CIDOC CRM o LIDO[28]. Cada uno de estos esquemas son utilizados por múltiples proyectos, por

27. Linked Open Vocabularies (LOV). Disponible en: http://lov.okfn.org/dataset/lov/. [Consulta: 15/01/2016]

28. Paola Ronzino, Nicola Amico, Franco Niccolucci, "Assessment and comparison of metadata schemas for architectural heritage", in *XXIII International CIPA Symposium,* Praga, República Checa, 12-16 Septiembre, 2011.

ejemplo EDM lo utiliza Carare o Euroscreen. En la práctica muchos proyectos optan por prescindir de toda la complejidad de estos esquemas, aunque, por norma general, tratan de mantener la compatibilidad con algunos campos de los esquemas más importantes, como EDM.

Los recursos han sido descritos según la documentación disponible en las páginas públicas de Internet sobre estos proyectos (Tabla 2).

PROYECTO	URI	OBJETIVOS	LOD
Carare	http://www.carare.eu/	Europeana, mostrar monumentos historicos	Geonames
Euscreen	http://www.euscreen.eu/	Europeana, acceso a programas para estudiar la historia de la TV	Dbpedia, Eurostat, Freebase, NY Times
BiographyNet	www.biographynet.nl/	Biografias enlazadas con lugares y hechos relevantes	Geonames, Dbpedia
WW1LOD	http://www.ldf.fi/dataset/ww1lod/	documentos de la WW1, con sucesos, lugares, protagonistas, periodos	Dbpedia, Geonames, Canadian Names Authorities
Pelagios Project	pelagios-project.blogspot.com	Ubicación de objetos (textos, imágenes.) del mundo antiguo	Geonames
Nomisma	http://nomisma.org/	Numismática en LoD	Getty, Dbpedia, Freebase, geonames, VIAF, pleiades, British museum
Smithsonian Amer. Art Museum	http://americanart.si.edu/collections /search/lod/about/	Enlazar objetos de arte a LoD	Dbpedia, Getty, Geonames, NY Times

Tabla 2. Vocabularios de metadatos utilizados en proyectos de historia

Resulta significativo que se utilicen con una frecuencia muy elevada vocabularios genéricos, no centrados en historia. También se pude observar la alta frecuencia de uso de *geonames*. Este vocabulario es frecuentemente utilizado para la localización geográfica, traduciéndose en los proyectos en una visualización mediante mapas. Por último, se observa un uso alto de listados de nombres propios como VIAF. Como consecuencia se observa una falta de recursos específicos de la investigación en historia, factor que puede afectar a su uso en investigación.

Por otro lado es patente la utilidad de los recursos resultado de la interrelación de estos vocabularios. Un ejemplo significativo son las bibliotecas virtuales de la Fundación Larramendi, con colecciones tales como la "Ciencia y Técnica en la Empresa Americana"[29] o la lista de autoridades denominada "Colección de Polígrafos Españoles". Obtenida mediante una agregación automática, mucha de la información en estas colección es novedosa.

7. Un caso práctico, el repositorio Almahisto

El proyecto Almahisto[30] trata de diseñar, desarrollar y alojar una biblioteca virtual que facilite la investigación historiográfica. El periodo histórico cubierto por el proyecto comprende desde la llegada de los Borbones a España hasta el final de la Guerra Civil (1700-1936). Se trata de un recurso virtual que da un valor añadido a documentos de bibliotecas digitales[31]. Por tanto, los documentos pueden ser tanto coetáneos o como posteriores, pero tratando del periodo en cuestión.

7.1. Objetivos y Planificación

1) Objetivos:

El proyecto trata de dar un valor añadido a ciertos documentos históricos de valor, y que por tanto, frecuentemente, no pertenecen a la institución. El objetivo es facilitar la investigación historiográfica futura. Los requisitos de diseño son:

- Asegurar en lo posible la pervivencia futura del recurso, más allá del periodo de desarrollo del proyecto.

29. Ciencia y Técnica en la empresa americana http://www.larramendi.es/cytamerica/i18n/micrositios/inicio.cmd; Colección de Polígrafos Españoles http://www.larramendi.es/i18n/consulta_aut/indice_campo.cmd?campo=idpoliespa.

30. Proyecto Almahisto. https://almahisto.wordpress.com/.

31. Aunque frecuentemente biblioteca virtual y digital se han utilizado como sinónimos, actualmente el segundo se reserva para los repositorios que almacenan documentos, mientras que el primero puede dar un valor añadido a los recursos del segundo, sin poseer directamente el documento digital.

- Los recursos deben de incorporar información adicional que no esté presente en el documento ni en su descripción en el repositorio de origen. Esta información debe facilitar la investigación historiográfica sobre el periodo estudiado
- Respetar los objetivos de diseño propuestos por Berners-Lee[32], estos son: utiliza formatos no propietarios, publicación bajo una licencia abierta de los datos en la Web, publicar los datos de forma estructurada y vinculados con otros datos.
- Tratar de utilizar un esquema fácil de mantener y sencillo, para que la inserción y actualización futura de nuevos registros sea más ágil.
- Utilizar metadatos populares y evitar crear nuevos valores para esos metadatos. Por ejemplo, antes de crear un autor nuevo asegurarse que no existe en los *value vocabularies*.

2) Selección del software de alojamiento

Como se afirma en Moreiro et al.[33], el software libre permite mayores posibilidades de configuración para un proyecto concreto, una longevidad no sujeta a la supervivencia de una empresa concreta, y un mayor volumen de usuarios y programadores. Estos aportan una mayor experiencia de uso, mejor documentación y un mantenimiento y actualización del software adecuado. Es por esta razón, y no el coste, lo que recomienda utilizar un software libre en proyectos relacionados con el dominio de la historia. Se debe tener en cuenta que el valor de un repositorio digital en historia aumenta con el número de recursos y la longevidad del proyecto. Por otra parte, frecuentemente la instalación y configuración de este tipo de software suele ser más compleja que en el caso del software propietario, requiriendo recursos humanos y financieros para su instalación.

A nivel mundial el software gratuito de código abierto más utilizado para gestionar colecciones digitales es DSpace (ver Figura 5). Como ilustración en la imagen se muestra su uso según dos registros públicos: el Registro de Repositorios de Acceso Abierto[34] (ROAR) y Repository66[35]. Solo la gran comunidad de usuarios de Dspace ya recomienda su uso en proyectos de historia. Pero en el caso del proyecto Almahisto coadyuvó otro factor. La mayoría de las grandes colecciones digitales está en bibliotecas y universidades. El proyecto Almahisto se desarrolló dentro de la Universidad Carlos III, y el repositorio institucional de esta es DSpace. Dado que un factor

32. Berners-Lee, *loc. cit*.

33. J.A. Moreiro, Sonia Sánchez-Cuadrado, Vicente Palacios, Eduardo Barra, "Evaluación de software libre para la gestión de archivos administrativos". *El profesional de la información*, 2011, 20, 2, 2011, 206-213.

34. Registry of Open Access Repositories, http://roar.eprints.org/view/software/

35. Repository66 http://maps.repository66.org/

crítico para un proyecto en historia es su longevidad se trató de asegurar su supervivencia, insertando el proyecto en el repositorio institucional[36].

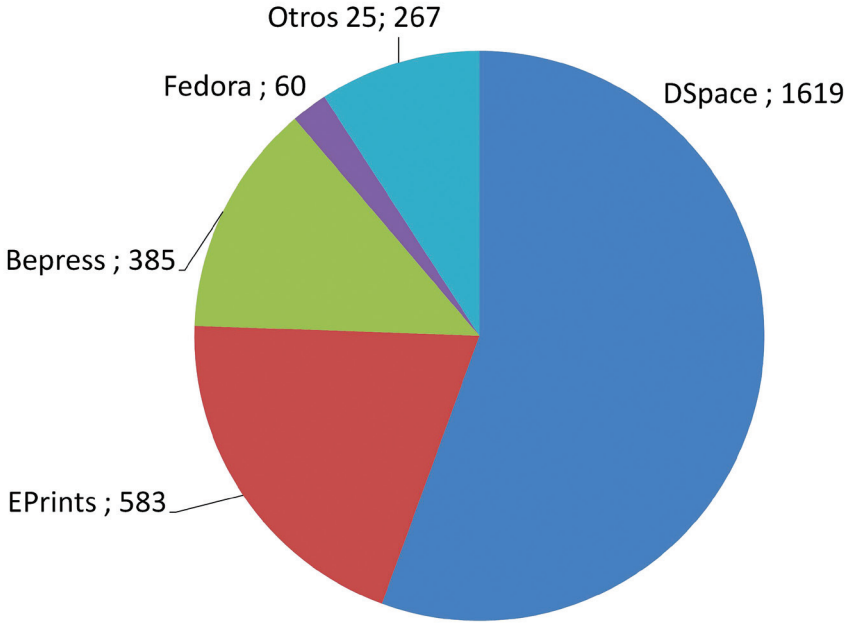


Figura 5. Principales Repositorios de Acceso Abierto ordenados según número de instalaciones.

Un problema frecuente en la implementación de proyectos tecnológicos es la evolución de la tecnología, y las medidas a tomar para minimizar el esfuerzo para adaptarse a las nuevas versiones.

En el caso del proyecto Almahisto la principal dificultad residió en la escasa madurez del software para gestionar datos enlazados en la Web Semántica. Al inicio del proyecto DSpace se encontraba en su versión 3.0. Por defecto en las versiones originales, DSpace solo tenía por defecto dos vocabularios con una cobertura limitada: *Norwegian Science Index* y Swedish *Research Subject*. Si se quería incluir otro vocabulario era preciso incluir un documento XML con todos los elementos. Lamentable-

36. Repositorio Almahisto http://e-archivo.uc3m.es/handle/10016/19460

mente este vocabulario no está enlazado con la LOD. No ha sido hasta recientemente que se han incluido tripletas y terminales de consulta mediante el lenguaje SPARQL (ver Tabla 3).

Versión		**Enlazado**
3,0	Desde la versión 3.0 se pueden poner vocabularios propios o externos. Esta información debe codificarse en XML.	No hay acceso a tripletas ni un terminal de consulta Para recolectar los datos utiliza OAI-PMH.
4,0	Se mejora 3.0 con ficheros de autoridades como ORCID, VIAF o LC (Biblioteca del Congreso).	En la práctica da errores y su puesta a punto es compleja, se trata de un XML propio, no se utilizan tripletas RDF
5,0	Generación de URIs automática, exporta tripletas	En 2014 Dspace publica la primera versión con LoD. Esta versión incluía un terminal de consulta

Tabla 3. Evolución de DSpace en relación a la LOD

La adaptación a los nuevos recursos pasa por dos opciones, bien:

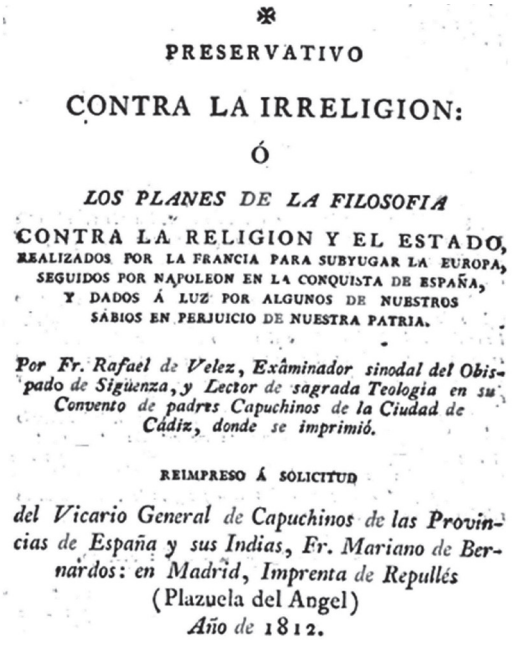
- Extraer tripletas que puedan ser integrados en LOD con otro software y navegar conceptualmente con él.
  - · Ventajas: El coste en tiempo de implementación será reducido.
  - · Desventajas: tener dos repositorios implica duplicar el esfuerzo en el futuro.
- Esperar a la actualización a Dspace 5, y una vez que de acceso a LoD integrarlo.
  - · Ventajas: el mantenimiento reside en un único recurso.
  - · Desventaja: el trabajo que implica no es conocido, ya que reutilizar el trabajo previo precisa de una solución ad-hoc.

En el caso de Almahisto la opción ha sido la segunda, ya que como se ha comentado, los repositorios digitales en historia deben permitir un mantenimiento sencillo a largo plazo.

7.2. Adquisición de documentos

Dado que se trata de una biblioteca virtual que agrega recursos de bibliotecas digitales se firmaron acuerdos con estas bibliotecas. Dspace[37] puede recolectar registros de estas bibliotecas mediante el protocolo denominado OAI-PHM. El problema que surgió fueron los problemas en la importación, principalmente importación a campos equivocados. Se decidió realizar la limpieza de forma manual dada la casuística tan diversa que presentaban los errores.

Como se comentó en la sección de métodos en historia, la correcta valoración de la contribución de una fuente debe tener en cuenta aspectos tales como la época, el contexto del autor, el público al que va dirigido o la credibilidad y autoridad para explicar ciertos hechos, entre otros aspectos. En la colección Almahisto podemos ver frecuentes ejemplos de la relevancia de dichos factores y los sesgos patentes (como puede verse en las figuras 6 y 7). El objetivo del proyecto Almahisto es poner de relieve estos aspectos para facilitar a los historiadores una investigación historiográfica adecuada.



✠

PRESERVATIVO

CONTRA LA IRRELIGION:

Ó

LOS PLANES DE LA FILOSOFIA

CONTRA LA RELIGION Y EL ESTADO,
REALIZADOS POR LA FRANCIA PARA SUBYUGAR LA EUROPA,
SEGUIDOS POR NAPOLEON EN LA CONQUISTA DE ESPAÑA,
Y DADOS Á LUZ POR ALGUNOS DE NUESTROS
SÁBIOS EN PERJUICIO DE NUESTRA PATRIA.

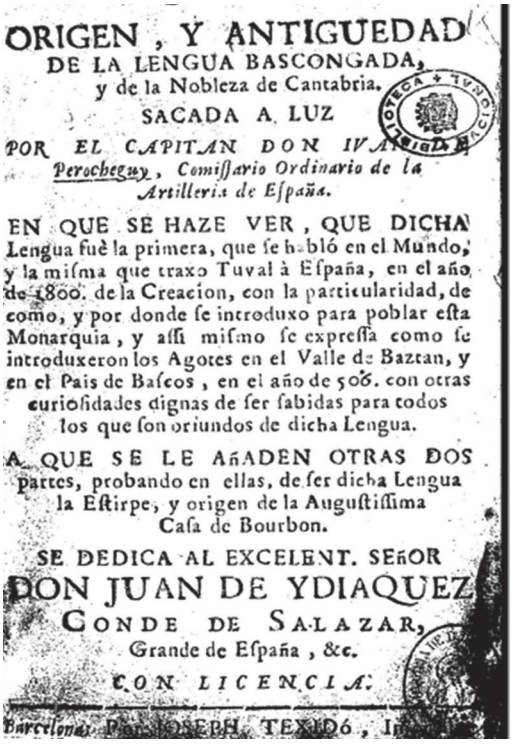
Por Fr. Rafael de Velez, Exâminador sinodal del Obispado de Sigüenza, y Lector de sagrada Teologia en su Convento de padres Capuchinos de la Ciudad de Cádiz, donde se imprimió.

REIMPRESO Á SOLICITUD

del Vicario General de Capuchinos de las Provincias de España y sus Indias, Fr. Mariano de Bernardos: en Madrid, Imprenta de Repullés
(Plazuela del Angel)
Año de 1812.

Fig. 6 Preservativo contra la Irreligión
(Fuente: colección Almahisto, y este de Cervantes Virtual)

37. DSpace http://www.dspace.org/



ORIGEN, Y ANTIGUEDAD
DE LA LENGUA BASCONGADA,
y de la Nobleza de Cantabria.
SACADA A LUZ
POR EL CAPITAN DON IVA
Perochegui, Comiſſario Ordinario de la
Artilleria de Eſpaña.
EN QUE SE HAZE VER, QUE DICHA
Lengua fuè la primera, que ſe habló en el Mundo,
y la miſma que traxo Tuval à Eſpaña, en el año
de 1800. de la Creacion, con la particularidad, de
como, y por donde ſe introduxo para poblar eſta
Monarquia, y aſſi miſmo ſe expreſſa como ſe
introduxeron los Agotes en el Valle de Baztan, y
en el Pais de Baſcos, en el año de 506. con otras
curioſidades dignas de ſer ſabidas para todos
los que ſon oriundos de dicha Lengua.
A QUE SE LE AñADEN OTRAS DOS
partes, probando en ellas, de ſer dicha Lengua
la Eſtirpe, y origen de la Auguſtiſſima
Caſa de Bourbon.
SE DEDICA AL EXCELENT. SEñOR
DON JUAN DE YDIAQUEZ
CONDE DE SALAZAR,
Grande de Eſpaña, &c.
CON LICENCIA.
Barcelona: Por JOSEPH TEXIDO, I

Figura 7. Origen y antigüedad de la lengua vascongada.
Fuente: colección Almahisto,
y éste de la Biblioteca Digital Hispánica

7.3. Limpieza

Otro elemento crítico es la limpieza de datos. En el caso del proyecto Almahisto se utilizaron recolectores como el de la Biblioteca Digital Hispánica. Se trata de recursos de acceso libre, con 29 millones de objetos y 2300 instituciones en noviembre de 2015. Utiliza el esquema EDM, ya que tiene una relación estrecha con Europeana. Uno de los problemas que se producen en la recolección es que los registros no estén bien estructurados o contienen errores en la importación. En estos casos el conjunto de registros debe ser limpiado para impedir errores a largo plazo. Ha de tenerse en cuenta que cualquier integración automática posterior con recursos externos es imposible si los registros contienen errores.

7.4. Enriquecimiento mediante metadatos e integración con recursos externos

Dentro del proyecto se pueden incorporar nuevos metadatos. En LOD en Almahisto se han incorporado, además de otros propios para la investigación historiográfica, los mencionados en la tabla 4. Estos metadatos permiten la comparación con recursos externos, de acuerdo a la metodología de comparación en historia.

Metadato	**LoD**
autor	https://viaf.org/
editor	https://viaf.org/
prologuista	https://viaf.org/
ilustrador	https://viaf.org/
impresor	https://viaf.org/
cobertura geográfica	thes.cindoc.csic.es o www.vocabulary server.com/toponimos/index.php
cobertura geonames	http://www.geonames.org

Tabla 4. Metadatos relacionados con recursos en la LOD en Almahisto. Fuente: Elaboración propia

8. Conclusiones

En este documento se debaten iniciativas para explotar estos recursos para la investigación histórica, y en concreto su empleo para la investigación historiológica. Se ha tratado de mostrar que las tecnologías digitales no solo son consistentes con las metodologías empleadas en historia, sino que presentan una oportunidad para estos estudios. Estas metodologías permiten agilizar las tareas en investigación, al tiempo que amplían el número de documentos con el que sustentar una hipótesis. Por último, la estructura de representación mediante tripletas permite una comparación muy complicada de realizar por metodologías más clásicas.

Cientos de proyectos en humanidades digitales han recibido una financiación elevada durante los últimos años[38], al mismo tiempo miles de documentos con rele-

38. Matthew K. Gold, *Debates in the digital humanities*. U. Minnesota Press, 2012

vancia histórica son escaneados y subidos a la web cada año[39] . La pregunta que subyace es hasta qué punto toda esta inversión es aprovechada por los investigadores. La respuesta es que estos recursos son frecuentemente infrautilizados, bien por carecer del conocimiento para explotarlos, bien por considerarse ajenos a las metodologías propias de la historia.

Como se ha mostrado a lo largo del trabajo, las tecnologías presentes en los proyectos de humanidades digitales son diversas. Existen muchos diseños alternativos para los proyectos de Historia Digital, y decidir el diseño idóneo precisa de conocer las ventajas e inconvenientes de cada solución. Por otro lado, los vocabularios usualmente empleados y la observación de algunos proyectos parecen indicar que en su diseño no se ha considerado las necesidades de la investigación en historia. Se exponen las estrategias y técnicas básicas para aportar al historiador las herramientas y conocimientos para una explotación adecuada de estos recursos.

39. Feigel, Kirsten, *The Digitization and Accessibility of Documents: A Case Study at the Rochester Public Library*. Tesis. Rochester Institute of Technology, 2015