LA LITERATURA MEDIEVAL HISPÁNICA EN LA IMPRENTA (1475-1600)

COLECCIÓN PARNASEO
28

LA LITERATURA MEDIEVAL HISPÁNICA EN LA IMPRENTA (1475-1600)

Edición de
María Jesús Lacarra

Con la colaboración de
Nuria Aranda García

VNIVERSITAT DE VALÈNCIA
2016

Octubre de 2016
I.S.B.N.: 978-84-370-9998-9
Depósito Legal: V-2151-2016

Diseño de la cubierta:
Celso Hernández de la Figuera y José Luis Canet

Imagen de la portada:
Juan de Iciar, *Ortografía práctica* (Zaragoza: Bartolomé Nájera, 1548),
imagen manipulada, que sirve de logo del Catálogo Comedic.

Maquetación:
Héctor H. Gassó

Publicacions de la Universitat de València
http://puv.uv.es
publicacions@uv.es

Parnaseo
http://parnaseo.uv.es

Este volumen se incluye dentro del Proyecto de Investigación del
Ministerio de Economía y Competitividad, referencia FFI2014-51781-P,
y del grupo investigador *Clarisel*, financiado por el Gobierno de Aragón.

La literatura medieval hispánica en la imprenta (1475-1600) / edición de María Jesús Lacarra, con la colaboración de Nuria Aranda García

Valencia : Publicacions de la Universitat de València, 2016
272 p. ; 17 × 23,5 cm — (Parnaseo; 28)
ISBN: 978-84-370-9998-9

1. Literatura castellana- -1475-1600. 2. Impremta--Història--Origen i antecedents. I. Lacarra, María Jesús. II. Publicacions de la Universitat de València

821.134.2.09"14/15"
655.11"14/15"

ÍNDICE GENERAL

PhiloBiblon and the Semantic Web
Notes for a Future History[1]

Charles B. Faulhaber
University of California, Berkeley

Before I get into the future history of PhiloBiblon, it will be useful to define it and place it within the context of what used to be called «humanities computing».

PhiloBiblon is nothing less than the attempt to create a union catalog of all texts, manuscripts, and printed books of interest for the study of the vernacular Romance cultures of medieval Spain, the individuals and institutions that created and used those texts, and the secondary literature surrounding them. Currently it contains more than 340,000 records and has become an indispensable resource for the study of the cultures of the medieval Iberian Peninsula.

«Humanities computing» began in the 1940s with the efforts of Father Roberto Busa to persuade Thomas J. Watson of IBM to provide him with data processing equipment to automate his *Index Thomisticus*, a concordance to the works of St. Thomas Aquinas.[2] Until the 1990s humanities computing remained essentially the province of a fringe of scholars scattered throughout the humanities both in the United States and Europe. The progress and interests of the field can be traced in the journals *Computers and the Humanities* (1966) and *Literary & Linguistic Computing* (1986).[3]

1. Written pursuant to National Endowment for the Humanities grant PW-51633. All websites were consulted between March 22, 2016, and April 11, 2016.

2. Roberto Busa (comp.), *Index Thomisticus: Sancti Thomae Aquinatis operum omnium indices et concordantiae in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur,* Stuttgart-Bad Cannstatt, Frommann-Holzboog, 1974-1980; Roberto Busa, «The Annals of Humanities Computing: The Index Thomisticus», *Computers and the Humanities*, 14.2 (1980), pp. 83-90; Robert W. Schmidt, «An Historic Research Instrument: The Index Thomisticus», *The New Scholasticism*, 50.2 (1976), pp. 237-49.

3. For a silver anniversary retrospective of humanities computing see Joseph Raben, «Humanities Computing 25 Years Later», *Computers and the Humanities*, 25 (1991), pp. 341-50; and for a review from the perspective of the new Digital Humanities, Susan Hockey, «The History of Humanities Computing», in *A Companion to Digital Humanities*, eds. Susan Schreibman, Ray Siemens and John Unsworth, Malden [UK], Blackwell, 2004, pp. 3-19.

Despite interest among some Hispanists as early as 1967, humanities computing had little impact in the Spanish-speaking world during this early period.[4] The first sustained effort, which took place in the U.S., thanks to support from the National Endowment for the Humanities in 1973, was the Dictionary of the Old Spanish Language Project (DOSL) at the University of Wisconsin, Madison; and its first product was the *Bibliography of Old Spanish Texts* (BOOST) in 1975, conceived originally as an in-house database of incunabula and pre-1500 manuscripts to catalog potential source materials for DOSL. It was re-published with that same purpose in 1977 and expanded to include post-1500 manuscripts in 1984.[5]

The DOSL project came to completion with the publication of the *Diccionario de la prosa castellana del Rey Alfonso X* (2002), but its influence and many of its materials live on with the transfer of the Hispanic Seminary of Medieval Studies to the Hispanic Society of America and the on-going web publication of the hundreds of transcriptions prepared for the project during the course of its existence.[6] PhiloBiblon, the direct lineal descendent of BOOST, can justly lay claim to being the oldest surviving Digital Humanities project in the Hispanic world. Its history has some cautionary wisdom to offer to its more recent epigones.

There is no point in recapitulating that history in detail, since John Nitti, Ángel Gómez Moreno, Francisco Marcos-Marín, and I have reviewed it periodically since 1983.[7] The goals of the project have remained remarkably

4. One may cite only Elsa T. de Pucciarelli, «Aspectos técnicos y literarios de la traducción», *Boletín de Estudios Germánicos*, 5 (1964), pp. 137-55; Alice M. Pollin *et al.*, *Concordancias de la obra poética de Eugenio Florit*, New York, New York Institute for Computing in the Humanities, 1967; and several articles concerning the use of computers for lexicography, all from the same project: Manuel Ariza Viguera *et al.*, «Atlas lingüísticos plurilingües con ordenadores electrónicos», *Boletín del Centro de Cálculo de la Universidad de Madrid*, 23 (1973), pp.12-15; and Ignacio del Campo Martín *et al.*, «El análisis sintáctico automático como una ayuda para la elaboración del diccionario», *Boletín del Centro de Cálculo de la Universidad de Madrid*, 23 (1973), pp. 16-31.

5. *Bibliography of Old Spanish Texts*, ed. Anthony Cárdenas, John Nitti, and Jean Gilkison, Madison, Hispanic Seminary of Medieval Studies, 1975; 2nd ed. Anthony Cárdenas *et al.*, Madison, Hispanic Seminary of Medieval Studies, 1977; 3rd ed. Charles B. Faulhaber *et al.*, Madison, Hispanic Seminary of Medieval Studies, l984.

6. Lloyd August Kasten and John Nitti (comps.), *Diccionario de la prosa castellana del Rey Alfonso XI*, New York, Hispanic Seminary of Medieval Studies, 2002; Francisco Gago-Jover (ed.), *Digital Library of Old Spanish Texts*, New York, Hispanic Seminary of Medieval Studies, 2011 (http://www.hispanicseminary.org/textconc-en.htm).

7. Charles B. Faulhaber and John Nitti, «Boost$_3$, Debits and Credits», *La Corónica*, 11(1983), pp. 286-92; Charles B. Faulhaber and Ángel Gómez Moreno, «BOOST$_4$», *La Corónica*, 14 (1986), pp. 291-92; Charles B. Faulhaber and Francisco Marcos-Marín, «ADMYTE: Archivo Digital de Manuscritos y Textos Españoles», *La Corónica*, 18 (1990), pp. 131-45; Charles B. Faulhaber, «Hispanismo e informática», *Incipit, 6* (1986 [1987]), pp. 157-84; Charles B. Faulhaber, «Bibliography of Old

consistent since the printed edition of 1984 —a union catalogue, in the words of PhiloBiblon's home page— of «the romance vernacular sources of medieval and early modern Iberian culture».[8]

What *has* changed is the database management system (dbms) used to catalogue those sources and the delivery systems used to make them available to interested scholars and students. In the early days of humanities computing the computer was conceived of primarily as a tool to produce a printed book; the book itself was the desired outcome, as it had been for over five hundred years. In the case of BOOST, the computer was a UNIVAC 1110 at the University of Wisconsin's Madison Academic Computing Center running FAMULUS, a flat-file dbms. The first edition of BOOST was simply a photo-offset copy of FAMULUS print-out, but the editors, John Nitti and Lloyd Kasten, did something rather clever. At the end of the printed volume they bound in a dozen «Entry Update Forms» so that scholars could send in additions and corrections —one of the earliest known examples of crowdsourcing for humanities computing projects. The technology for the second and third editions was essentially unchanged, although Wisconsin programmers expanded the limitations of FAMULUS with regard to the number of data fields and their size and formatted the printed edition to make it more legible, with upper and lower case letters, accented characters, and italics and boldface. This system was used to produce the third edition of BOOST (1984) and the first (and only) edition of the *Bibliography of Old Catalan Texts* (BOOCT*)* (1985).[9]

In 1986 the whole operation was moved from Madison to Berkeley's SPIRES (Stanford Public Information Retrieval System) main frame flat-file dbms. Instead of the batch file processing of FAMULUS, SPIRES was interactive;

Spanish Texts: Evolution of a Data Base», in *Databases in the Humanities and Social SciencesB4. Proceedings of the International Conference on Databases in the Humanities and Social Sciences held at Auburn University at Montgomery, July, 1987*, ed. Lawrence J. McCrank, Medford, NJ, Learned Information, 1989, pp. 213-21; Charles B. Faulhaber, «Desiderata para el estudio de las literaturas hispánicas medievales», in *Medioevo y literatura. Actas del V Congreso de la Asociación Hispánica de Literatura Medieval (Granada, 27 septiembre — 1 octubre 1993)*, ed. Juan Paredes, Granada, Universidad de Granada, 1995, I, pp. 93-107; Charles B. Faulhaber and Ángel Gómez Moreno, «De BOOST a BETA: de Madison a Berkeley», in *Los códices literarios de la Edad Media. Interpretación, historia, técnicas y catalogación*, dir. Pedro M. Cátedra, ed. Eva Belén Carro Carvajal and Javier Durán Barceló, San Millán de la Cogolla, Cilengua; Instituto de Historia del Libro y de la Lectura, 2009, pp. 283-92; Charles B. Faulhaber, «PhiloBiblon: Pasado y futuro», *Incipit*, 29 (2009), pp. 191-200; Charles B. Faulhaber, «PhiloBiblon, Information Technology, and Medieval Spanish Literature: A Balance Sheet», in *Humanitats a la xarxa: Món medieval. Humanities on the Web: The Medieval World*, ed. Lourdes Soriano *et al*., Bern, Peter Lang, 2014, pp. 15-43. See also the review of the 2011 web version of PhiloBiblon by Francisco Gago-Jover, «PhiloBiblon», *Digital Philology*, 1.2 (2012), pp. 323-27.

8. PhiloBiblon, <http://vm136.lib.berkeley.edu/BANC/philobiblon/>.

9. For an overview of PhiloBiblon's technical history to 2001, see PhiloBiblon, «Technical History», <http://vm136.lib.berkeley.edu/BANC/philobiblon/history_en.html>.

instead of the fixed-length fields of FAMULUS, SPIRES allowed variable-length fields, much more useful for the capture of textual data. The only problem: no funding to program SPIRES output for printing.

More change came within a year. In 1987, thanks to a grant from IBM we mapped both BOOST and BOOCT into the proprietary Revelation relational dbms running on a standard PC. We chose Revelation since it was the only PC-based relational dbms that allowed for variable-length fields and that, with a 64K limit on the size of a record or any given field in a record, had no effective limitations on our data. This vastly increased the ease of data entry, since in a relational dbms changes only needed to be made once, for example, in the master record for a manuscript rather than in the records of all the texts it contains. That master record could then be linked to the records for each text contained in the manuscript, which in turn could be linked to the master record for each work, which records all of the invariant information about a given text, such as its author, its title, and its date of composition.

With the change to the new system, however, we were faced with the problem of distributing the data. By then we had realized that print was not the answer, not only because a printed database becomes outdated as soon as it is printed but also, and primarily, because print neutralizes most of a database's major advantages: highly customizable searches and the ability to understand complex relationships among various data elements.

Greatly daring, and with the incorporation of the new database of medieval Portuguese materials in 1988 (originally to be called *Bibliography of Old Portuguese Texts* (BOOPT), we decided to forego further print editions in favor of electronic distribution on CD-ROM disks, first introduced for data in that same year of 1988.[10] We thought that, with luck and sufficient funding, we *might* be able to release the set of three bibliographies along with the Revelation dbms itself by the year 2000. Both the luck and the funding arrived in the run-up to the 1992 celebration of the 500th anniversary of the *Encuentro de dos mundos*, as the Sociedad Estatal Quinto Centenario delicately put it. Our friend and colleague Francisco Marcos-Marín (Universidad Autónoma de Madrid) was appointed as the Director of the Industrias de la Lengua of the Sociedad Estatal.[11] One of its major projects was the *Archivo Digital de Manuscritos y Textos Españoles* (*ADMYTE*), whose first CD-ROM, published in July of 1992, contained searchable transcriptions of fifty texts printed between 1481 and 1520, along with facing digitized facsimiles —to my knowledge the first such

10. «CD-ROM», *Wikipedia,* <https://en.wikipedia.org/wiki/CD-ROM>.

11. Francisco Marcos-Marín, «Industrias de la lengua», *Idiomas. Todo sobre los idiomas,* 13 (1992), pp. 6-11.

edition of text and image.[12] The second disk, called «Vol. 0» but released in 1993,[13] included:

- another 64 texts from both manuscripts and printed editions, but without the facsimiles.
- TACT, the Textual Analysis Computing Tools suite developed by Ian Lancashire at the University of Toronto.
- Marcos-Marín's UNITE program for computer-assisted textual criticism and the newly-baptized PhiloBiblon, with its three component bibliographies:
  - BETA (= BOOST)
  - BITECA (= BOOCT)
  - BITAP (= BOOPT), later BITAGAP

The CD-ROM version of PhiloBiblon offered both the data and the dbms needed to manipulate those data, along with a detailed manual to explain how to find, extract, and print information concerning primary sources, texts, related individuals —in short, everything in the three bibliographies of conceivable interest to scholars. This product was not for the casual user, however; the learning curve to use it effectively was rather steep.

Even with ample development support from the Sociedad Estatal Quinto Centenario, it was necessary to price the CD-ROM at $800 in order to allow Micronet, S.A., its producer and distributor, to recover costs and make a small profit. This necessarily limited access to major research libraries and consequently to scholars affiliated with those libraries.

Nevertheless, with CD-ROM publication we had moved from information distribution via the 1500-year-old codex format, which had not varied in its essentials in the change from manuscript to print, to something that was truly different. Instead of fixed text accessed through multiple indexes or simply read sequentially, we now had the capability of entering the corpus at any point, interactively, and then following the complex web of relationships that would lead us from one record to another: from the name of an author to the texts written by that author to the manuscripts containing those texts to the libraries holding those manuscripts to the previous owners of those manuscripts... Moreover, users now had the capability of printing just the portions of the data of interest to them, not the entire corpus.

At this stage PhiloBiblon was still at the cutting edge of humanities computing in general and of dbms software for the humanities in particular. The single

12. Francisco Marcos-Marín *et al., ADMYTE: Archivo Digital de Manuscritos y Textos Españoles*, Madrid, Quinto Centenario - Biblioteca Nacional — Micronet, I, 1992 [CD-ROM disk]; Francisco Marcos-Marín and Charles B. Faulhaber, *art. cit.*

13. Francisco Marcos-Marín *et al., ADMYTE: Archivo Digital de Manuscritos y Textos Españoles*, Madrid, Quinto Centenario — Biblioteca Nacional — Micronet, 1993, 0. [CD-ROM disk].

flat-file dbms of the BOOST days, after successive iterations by our Revelation programmer, John May, had become a relational dbms with ten complexly related tables with over 650 data elements. Early on we had realized that it was not sufficient to catalogue manuscripts, early editions, and the texts they contain; it was also necessary to capture prosopographical information about the people and the institutions that produced and used those texts as well as the places where production and use took place. PhiloBiblon was becoming a window into the cultural universe of medieval Iberia.

In 1990 the information world changed, radically, with Tim Berners-Lee's invention of the World Wide Web at the Conseil Européen pour la Recherche Nucléaire (CERN) in Geneva. Its impact was not realized, however, until CERN made its code base available for free in 1993.[14] In the spring of 1994 I was teaching a graduate course on humanities computing for Hispanists at Berkeley. During the last class one of the students, María del Mar Fernández Vega —whose sister, it turned out, was working at CERN— asked me if I planned to say anything about the WWW. My response showed that despite more than fifteen years of experience with information technology I still could not forecast the future: «Why? It's not going anywhere».

Nevertheless, by 1997 we had managed to put up a simple web page that allowed searches on all three bibliographies simultaneously. The development process was straightforward. Consultant Jack Kessler designed a PhiloBiblon web page and mapped the output from a data dump of the MS_Ed table (records of manuscripts and printed books) into HTML (HyperText Markup Language, one of Berners-Lee's key standards for the WWW). Kirk Hastings (UC Berkeley Library) then wrote a Perl script to convert the records of each manuscript or printed edition in BETA, BITECA, and BITAGAP into HTML files, one for each record. These files were in turn indexed using a public domain program (SWISH),[15] and a simple CGI (Common Gateway Interface) query form allowed searches by author and title or on free text. The web site contained a home page, home pages for each of the three bibliographies, and a link to the search engine.[16]

14. World Wide Web Foundation, *History of the Web*: <http://webfoundation.org/about/vision/history-of-the-web/>. See also Tim Berners-Lee and Mark Fischetti, *Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web*, New York, HarperCollins, 2000.

15. SWISH is an acronym for «Simple Web Indexing System for Humans», later re-written as Swish-e, <http://swish-e.org/>.

16. Charles B. Faulhaber (ed.), *PhiloBiblon*, Berkeley, University of California, 1997-2010. The earliest archived version of the original PhiloBiblon web site on the Internet Archive dates from April 29, 1999, two years after the web site was implemented. It shows the static web pages, but the search engine is disabled: <http://web.archive.org/web/19990428185054/http://sunsite.berkeley.edu/PhiloBiblon/phhm.html>.

The results page returned a list of every record that contained the search term —sometimes hundreds of records— but the only way to discover which of those records was of interest was to open each one and search inside it using the browser's search function. A nice feature was that searches functioned simultaneously across all three bibliographies, making it possible to see at a glance the pan-Iberian fate of authors and texts.[17]

Throughout this period work continued on PhiloBiblon's dbms. A major step forward was the conversion from DOS to Windows in 2001 and the continued enhancement of its input, search, and report capabilities. While the ten tables remained constant, the number of data fields almost doubled, to the current 1,246 data fields, with 98 controlled vocabulary lists, 30 data-entry screens, and 110 indexes.

Digital Humanities

The term «digital humanities», used at least as early as 1987, was canonized in 2004 as a replacement for «humanities computing» because of a seismic development that had rapidly changed the scholarly landscape for the humanities.[18] By the mid-1990s it had become both technically and economically feasible to digitize large quantities of manuscript and printed material.[19] In the U.S., after a five-year pilot program called «American Memory», the Library of Congress launched the National Digital Library Program with the collaboration of research libraries around the country.[20]

17. Because of the limitations of the web version, The Bancroft Library published PhiloBiblon and the BETA, BITECA, and BITAGAP databases on a stand-alone CD-ROM in 1999: Charles B. Faulhaber, Arthur L-F. Askins, and Harvey L. Sharrer (comps.), *PhiloBiblon. Electronic Bibliographies of Medieval Catalan, Castilian, Galician, and Portuguese Literature*, Berkeley, The Bancroft Library, 1999. This provided access to the full range of data as well as to the full search and report capabilities of the dbms. The entire production run of 200 copies sold out (at $100/copy) in just over two years, and the net income was used to enhance the newly created Windows version of PhiloBiblon.

18. P. M. Fraser and E. Matthews (eds.), *A Lexicon of Greek Personal Names*, Oxford, Oxford University Press, 1987, p. v; Susan Schreibman, Ray Siemens and John Unsworth (eds.), *A companion to digital Humanities*, Malden [UK], Blackwell, 2004; revised as *A New Companion to Digital Humanities*, Chichester, West Sussex, UK, John Wiley & Sons, 2016. For a discussion of the evolution from «humanities computing» to «digital humanities», see Patrik Svensson, «Humanities Computing as Digital Humanities», *Digital Humanities Quarterly*, 3.3 (2009) (http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html). See also Anne Burdick *et al*. (eds.), *Digital Humanities*, Cambridge MA, The MIT Press, 2012.

19. Gil Press, «A Very Short History of Digitization», *Forbes* (http://www.forbes.com/sites/gilpress/2015/12/27/a-very-short-history-of-digitization/#276deac225f1).

20. Library of Congress, *National Digital Library Program*, <https://memory.loc.gov/ammem/dli2/html/lcndlp.html>.

Berkeley's Bancroft Library, for example, as its contribution, on the history of the Chinese in California, digitized some

> 8000 photographs, original art, cartoons and other illustrations, letters, excerpts from diaries, business records, and legal documents; as well as pamphlets, broadsides, speeches, sheet music, and other printed matter from its Western Americana collections as its contribution.[21]

The digitization of medieval manuscripts was not far behind. Again, The Bancroft Library, in collaboration with the Rare Books and Special Collections Library of Columbia University, took the lead with the Digital Scriptorium Project. With funding from the Andrew W. Mellon Foundation, both libraries digitized selected images from all of the medieval MSS and documents in their collections. Currently the project, now run by a consortium, offers over 75,000 images from 8000 manuscripts in thirty-five institutions, primarily in the U.S.[22]

The Digital Scriptorium, however, has been far outstripped by the major European digital library projects, although they started much later. Europeana, the European digital library portal, first proposed in 2005 and in existence since 2008, currently contains more than 52 million digital objects from over 2,000 contributing institutions.[23] Spain's digital library, Hispana, which began in 2006 as the Directorio y Recolector de Recursos Digitales, currently houses 6 million digital objects from over 200 institutions,[24] while the Biblioteca Digital Hispánica of the Biblioteca Nacional de España, founded in 2008, now contains more than 134,000 digital objects, including all of its incunables and almost a thousand manuscripts in Spanish and Catalonian dated between 1200 and 1500.[25]

The Semantic Web

Meanwhile, the web continued to evolve. In 1999 the term «web 2.0» was coined, although it did not come into wide usage until 2004.[26] Web 2.0, the

21. *The Chinese in California, 1850-1925*, <http://bancroft.berkeley.edu/collections/chineseinca/index.html>.

22. Digital Scriptorium, <http://bancroft.berkeley.edu/digitalscriptorium/>.

23. Europeana, <http://www.europeana.eu/portal/>; «Europeana», *Wikipedia*: <https://en.wikipedia.org/wiki/Europeana>.

24. Ministerio de Educación, Cultura y Deporte, *Hispana*, <http://hispana.mcu.es/es/estaticos/contenido.cmd?pagina=estaticos/presentacion>; «Hispana (recolector OAI)», *Wikipedia*: <https://es.wikipedia.org/wiki/Hispana_(recolector_OAI)>.

25. Biblioteca Nacional de España, *Biblioteca Digital Hispánica,* <http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/>.

26. «Web 2.0», *Wikipedia*, <https://en.wikipedia.org/wiki/Web_2.0>.

development of interactive web pages, has been instantiated especially in social network applications but has had little impact on humanities computing beyond the adoption of blogs, Facebook, and Twitter on institutional web sites. PhiloBiblon was a relatively late adopter; our first blog did not appear until February of 2011, while we joined Facebook and Twitter in March of 2012.

A truly different version of the web, the Semantic Web, sometimes called web 3.0, sprang from the fertile minds of Tim Berners-Lee and his collaborators at the World Wide Web Consortium (CW3) in 2001. Building on the web's founding principles, CW3 has developed a series of standards to make the web truly universal:

> For anyone to be able to publish anything on the Web, all the computers involved have to speak the same languages to each other, no matter what different hardware people are using; where they live; or what cultural and political beliefs they have. In this way, the Web breaks down silos while still allowing diversity to flourish.[27]

Essentially, the semantic web seeks to define all relationships among data objects in machine-readable terms in an attempt to open up the mutually incompatible «silos» that wall data off.[28] The introduction to *W3C Semantic Web Activity* explains this with a couple of homely examples:

> The Semantic Web is a web of data. There is lots of data we all use every day, and it is not part of the web. I can see my bank statements on the web, and my photographs, and I can see my appointments in a calendar. But can I see my photos in a calendar to see what I was doing when I took them? Can I see bank statement lines in a calendar?
>
> Why not? Because we don't have a web of data. Because data is controlled by applications, and each application keeps it to itself.
>
> The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.[29]

27. World Wide Web Foundation, *ob. cit.* See also World Wide Web Consortium, <https://www.w3.org/>.

28. «Semantic Web», *Wikipedia,* <https://en.wikipedia.org/wiki/Semantic_Web>.

29. W3C, *W3C Semantic Web Activity*, <https://www.w3.org/2001/sw/>; see also W3C, *W3c Data Activity: Building the Web of Data,* <https://www.w3.org/2013/data/>.

In terms of the humanities, the different kinds of data objects we use at the most basic level are manuscripts, printed books, works of art, pieces of music... Books and manuscripts in turn are vehicles for texts and frequently for illustrations. The disciplines that work with these data objects have developed different and frequently idiosyncratic means of dealing with and describing them. In order for machines to deal with these objects, however, they must all be described in some standard way.

In practice this has meant the development of a series of internationally accepted standards.[30] The most important of these currently is RDF, the Resource Description Framework for the definition of classes and properties of resources on the web in very abstract terms.[31] It is based on principles instantiated most succinctly in the fifteen broad descriptors of objects, or «resources» (a book, a painting, a piece of music...), in the Dublin Core standard:[32]

1. Contributor: entity responsible for the resource.
2. Coverage: spatial or temporal topic of the resource.
3. Creator: entity responsible for making the resource.
4. Date: point in time related to the resource.
5. Description: an account of the resource.
6. Format: physical medium of the resource.
7. Identifier: unambiguous reference to the resource, e.g., a URI (Uniform Resource Identifier).
8. Language: natural or artificial language used in resource, e.g., Spanish, C++.
9. Publisher: entity that makes resource available.
10. Relation: a related resource.
11. Rights: information pertaining to the control of the resource.
12. Source: a related resource from which the named resource is derived.
13. Subject: topic of the resource.
14. Title: name of the resource.
15. Type: genre of the resource.

RDF takes a richer set of such objects and properties and focuses on their description or on the relationships among them through a Subject – Predicate

30. Pascal Hitzler, Markus Krötzsch and Sebastian Rudolph, *Foundations of Semantic Web Technologies*, Boca Raton, FL, CRC Press, 2010.

31. See, for example, W3C, *RDF Current Status*, https://www.w3.org/standards/techs/rdf#w3c_all: «All things described by RDF are called resources, and are instances of the class rdfs:Resource. This is the class of everything. All other classes are subclasses of this class. rdfs:Resource is an instance of rdfs:Class».

32. The Dublin Core standard takes its name from the 1995 meeting convened in Dublin, Ohio (U.S.), home of the OCLC library consortium, to articulate this set of descriptors: *Dublin Core Metadata Element Set, Version 1.1,* <http://dublincore.org/documents/dces/>.

– Object syntax known as a «triplestore» or simply «triple».[33] Each subject, predicate, or object —a concept, a person, a thing— must have a unique Uniform Resource Identifier that refers to it unambiguously. PhiloBiblon, for example already uses such identifiers internally, «record keys», to identify people, manuscripts, printed editions, texts. Thus Íñigo López de Mendoza, 1. marqués de Santillana, is bioid 1031; his son, Ínigo López de Mendoza y Figueroa, 1. conde de Tendilla, is bioid 1547. Individuals exist in a web of relationships that can thus be explicitly and unambiguously stated:

> Íñigo López de Mendoza (bioid 1031) = father = Ínigo López de Mendoza y Figueroa (bioid 1547)

Note, however, that the identifier for Santillana, bioid 1031, is unique to BETA. In BITAGAP Santillana is bioid 1543 and in BITECA, bioid 1454. Thus the importance of the URI established by the Virtual International Authority File, where Santillana is VIAF 39642259 (Personal).[34]

These objects and properties have been rigorously described and documented by W3C using a set of interrelated standards:

> • Linked Data (LD) «refers to an approach to publishing data that puts linking at the heart of the notion of data, and uses the linking technologies provided by the Web to enable the weaving of a global distributed database. By naming real world entities —be they web resources, physical objects such as the Eiffel Tower, or even more abstract things such as relations or concepts— with http(s) URLs, whose meaning can be determined by dereferencing the document at that URL, and by using the relational framework provided by RDF, data can be published and linked in the same way web pages can. The Linked Data Protocol specifies how web applications can, using the HTTP protocol, find resources and follow links, publish new resources, edit and delete existing ones».[35]
> • When Linked Data is made publicly available, it becomes Linked Open Data (LOD).

33. W3C, *Resource Description Framework (RDF)*, <https://www.w3.org/RDF/> (February 25, 2014 version).

34. The Virtual International Authority File (VIAF) combines multiple name authority files maintained by 44 national libraries and other institutions into a single searchable data base, with each variant of a given name linked to a single VIAF number. See «Virtual International Authority File», *Wikipedia*, <https://en.wikipedia.org/wiki/Virtual_International_Authority_File>. «Santillana, Íñigo López de Mendoza, Marqués de, 1398-1458» is the form used by the BNE and nine other national libraries, while the Biblioteca Nacional de Portugal uses «Santillana, Marquês de, 1398-1458», but the VIAF URI links them all together: http://viaf.org/viaf/39642259/.

35. W3C, *Linked Data*, <https://www.w3.org/standards/semanticweb/data>.

- Web Ontology Language (OWL) is a formal knowledge representation language.[36]
- Simple Knowledge Organization System (SKOS) defines explicitly the properties of «thesauri, taxonomies, classification schemes and subject heading systems».[37]

A detailed explanation of these three standards go far beyond the purposes of this article. Suffice it to say that they work together in the semantic web. Thus «The SKOS data model is formally defined in this specification as an OWL Full ontology [OWL-SEMANTICS]. SKOS data are expressed as RDF triples [RDF-CONCEPTS]».[38] In turn specific ontologies such as VIAF or the Getty Thesaurus of Geographic Names (TGN) are expressed as SKOS ontologies.[39]

PhiloBiblon on the Web, 2007-2016

However, the semantic web was still in its infancy in the first decade of the 21st century. Critical standards like RDF, OWL, and SKOS were still being defined. Nevertheless, it was clear that complete web access to PhiloBiblon's data was essential, so we had no alternative other than to come up with an *ad hoc* solution, which we were able to do thanks to two successive NEH grants, in 2007 and 2014. Under the first Rick Beaubien (UC Berkeley Library) wrote an XML schema to map Philobiblon's data fields into XML tagsets.[40] This enabled PhiloBiblon programmer John May to write export routines to map the contents of every record to the corresponding data elements in the schema.

Each of PhiloBiblon's tables is associated with a subset of the schema, and each of PhiloBiblon's data fields or structures is mapped to a set of XML elements. For example, the subset of the schema dealing with Persons, corresponding to PhiloBiblon's Biography table, begins:

```
<xs:element name="Person">
    <xs:complexType>
        <xs:sequence>
            <xs:element name=»BioID" type="IDType"/>
```

36. W3C, *Web Ontology Language Current Status*, <https://www.w3.org/standards/techs/owl#w3c_all>. The use of OWL «búho» instead of the expected acronym WOL is a classic example of the puckish humor of computer programmers.

37. W3C, *SKOS Current Status*, <http://www.w3.org/standards/techs/skos#w3c_all>.

38. *Ibídem*.

39. Patricia Harpring (ed.), *The Getty Vocabularies and Linked Open Data. Introduction and Editorial Perspective*, Paul Getty Trust, 2014 (http://www.getty.edu/research/tools/vocabularies/Linked_Data_Getty_Vocabularies.pdf).

40. «Philobiblon xsd», <http://sunsite.berkeley.edu/NewPhilobiblon/schema/Philobiblon.xsd>.

```
<xs:element name=»Name» type=»PersonNameType"/>
<xs:element name=»Sex» type=»xs:string" minOccurs="0"/>
<xs:element name=»Titles» type=»PersonTitleGroupType» min
Occurs=»0»/>[41]
```

Thus the 221 data fields in BETA's record for Santillana, map to 2098 lines in the Persons.Data.XML file, beginning as follows:

```
<Person>
    <BioID>
        <Bibliography>BETA</Bibliography>
        <Type>bioid</Type>
        <IDNo>1031</IDNo>
    </BioID>
    <Name>
        <Uniform>Íñigo López de Mendoza, 1. marqués de Santillana (1445-08-08 - 1458-03-25)</Uniform>
        <FirstAndLast>Íñigo López de Mendoza, 1. marqués de Santillana</FirstAndLast>
        <FirstAndLast>Íñigo López de Mendoza</FirstAndLast>
    </Name>
    <Sex>hombre</Sex>
    <Titles>
        <TitleInfo>
            <Title>1. marqués de Santillana (1445-08-08 - 1458-03-25)</Title>
        </TitleInfo>
        <TitleInfo>
            <Title>1. conde de Manzanares el Real (1445-08-08 - 1458-03-25)</Title>
        </TitleInfo>
    </Titles>
```

These tags define Santillana as a person and give him his unique identifier (bioid 1031), his name and its variants, his sex, and his titles.

For appropriate web access software we turned to the eXtensible Text Framework (XTF).[42] Developed by the University of California's California Digital Library (CDL), XTF is a flexible indexing and query architecture that supports searching across collections of heterogeneous textual data (XML, PDF, Word) and the presentation of results and documents in a highly configurable manner. It is an open source system divided into four components: crossQuery, the front-end to the collection search system; dynaXML, the interface to

41. The «xs» prefix identifies the element as part of an XML schema.

42. For a description see *XTF EXtensible Text Framework,* <http://xtf.cdlib.org>.

individual documents; Text Engine, used by crossQuery and dynaXML to perform text searches; and Indexer, the full-text indexer.

Under the 2014 grant Giulia Hill (UC Berkeley Library) re-wrote the schema in order to be able to index and display data that had been omitted from the original schema (e.g., Institutions). The grant also allowed us to establish a mirror website at the Universitat Pompeu Fabra in Barcelona.[43] and organize two seminars at the Centro Internacional de Investigación de la Lengua Española (Cilengua) in San Millán de la Cogolla in 2015 and 2016.[44] The updated web site went live in Berkeley and Barcelona in the spring of 2016.

PhiloBiblon and the Semantic Web

After 40 years of pioneer work, frequently on the cutting edge —or as we sometimes say in despair, the bleeding edge— of technical progress, PhiloBiblon finds itself, for the first time, behind the times. Digital Humanities work in Spain has now surpassed it.[45] Even the latest versions of PhiloBiblon, on Windows and on the web, do not comply with RDF international standards. Moreover, the current system is not sustainable in the long term, for several reasons. In the first place, the senior scholars who have led PhiloBiblon since the early 1980s, Charles Faulhaber, Arthur Askins, and Vicenç Beltran, are at or over retirement age. In the second place, the existing system, in the picturesque terminology of computer programmers, is a «kluge», accurately defined as «an ill-assorted collection of parts assembled to fulfill a particular purpose». The Windows dbms and the web access software have no organic relationship. The process of updating PhiloBiblon's data on the web depends entirely too much on human intervention:

43. We are grateful to Mercè Cabo Rigol, María Morrás, Joan Trench Rubal, and Marc Esteve, for their essential support in mounting the PhiloBiblon mirror site, <http://philobiblon.upf.edu/philobiblon/>.

44. These seminars have been organized under the auspices of Cilengua's Instituto de Literatura y Traducción, thanks to the support of its director, Carlos Alvar Ezquerra: I Seminario Internacional PhiloBiblon, <http://www.cilengua.es/convocatorias/i-seminario-internacional-philobiblon>; II Seminario Internacional PhiloBiblon, <http://www.cilengua.es/convocatorias/ii-seminario-internacional-philobiblon>.

45. The earliest survey of Digital Humanities in Spain, to my knowledge, is Antonio Rojas Castro, «El mapa y el territorio. Una aproximación histórico-bibliográfica a la emergencia de las Humanidades Digitales en España», *Caracteres*, 3 (2013), pp. 10-53; see now the monographic number of *Ínsula* edited by María Morrás and Antonio Rojas Castro, *Humanidades Digitales y Literaturas Hispánicas*, 70.822 (2015), pp. 1-40, with an up-to-date bibliography.

- Eight of PhiloBiblon's ten tables must be exported manually from BETA, BITECA, and BITAGAP.[46]
- The 48 exported files (eight data files and eight index files from each bibliography) must be validated manually against the PhiloBiblon schema using <oXygen/>.[47]
- Files that do not validate properly must either be corrected manually or, preferably, re-exported after corrections have been made in the Windows version.
- The validated files must then be uploaded, manually, to servers at Berkeley and the Universitat Pompeu Fabra.

At this point automation finally takes over: a chron file runs a Perl script against the XML data and index files at each institution, exploding them into individual records which are converted to HTML files on the fly as a result of a search by an end user on either web site.

This highly inelegant process has simply evolved over the past 40 years as an attempt to keep up with hardware and, especially, software changes in an environment of constrained resources. Human labor has taken on tasks that in a well-designed system would be performed by software. But it can't last. Serious illness or death of any of the senior individuals involved would very likely mean the end of PhiloBiblon as a viable scholarly resource.

What to do? The time has come to place PhiloBiblon into younger hands, but before doing that we must step back and re-examine the entire system in the light of the development of the semantic web. It is time to move PhiloBiblon, input and access, entirely to the web, taking full advantage of new international standards and state-of-the-art technologies. Thus, while building on 40 years of experience with PhiloBiblon, we propose to move PhiloBiblon from its proprietary database software to MySQL, optimized for compatibility with semantic web standards, i.e., RDF, LOD, and associated ontologies.

For the purposes of PhiloBiblon the most relevant version of RDF is the Europeana Data Model (EDM), since it sets the standards for the description of digitized Cultural Heritage Objects, like medieval manuscripts, to be included in Europeana, Hispana, and the BNE's Biblioteca Digital Hispánica.[48] In turn, of

46. The Subject Heading and Geography tables are not exported. Their data are incorporated as necessary in records in the other tables.

47. <oXygen/> is the software of choice for authoring, editing, and validating XML files and schema: <https://www.oxygenxml.com>.

48. See *Europeana Data Model Documentation* <http://pro.europeana.eu/page/edm-documentation>; *Europeana Data Model Primer 14/07/2013* <http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf>; *Definition of the Europeana Data Model v5.2.6 17/12/2014* <http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM%20Definition%20v5.2.6_01032015.pdf>.

even greater relevance are the specifications in *Digitised Manuscripts to Europeana* (DM2E), derived from EDM, which focus in greater detail on the kinds of data included in the descriptions of manuscripts, both medieval and post-medieval. The technical specifications created by the DM2E project between 2012 and 2015 are available on the web.[49]

PhiloBiblon's ten complexly related tables with over 1200 data elements will be mapped to DM2E, with such adjustments as may be necessary to accommodate data elements omitted from DM2E, and implemented in the open-source MySQL Community Server dbms.[50]

The project is neither conceptually nor technically difficult. It will include four phases:

1. Comparison of the PhiloBiblon data model with the EDM and DM2E;
2. Implementation of the new web version of PhiloBiblon in MySQL Community Server.
    a. Mapping the data fields in PhiloBiblon to the data elements in DM2E
    b. Replicating existing PhiloBiblon functionality, e.g., being able to identify automatically the reciprocal relationships among individuals.[51] Thus if we state that the 1. marqués de Santillana is the father of the 1. conde de Tendilla, it should not be necessary to specify manually that the 1. conde de Tendilla is the son of the 1. marqués de Santillana. The relationship will be implicit in the semantic triplestore.
3. Export of the data from PhiloBiblon's component bibliographies into the new dbms.
4. Permanent establishment of the new PhiloBiblon website at a Spanish institution.

While it is not possible to specify here how PhiloBiblon's data structure will be mapped to DM2E, the PhiloBiblon example given above, of the biographical record for Íñigo López de Mendoza, 1. marqués de Santillana , might look like this in DM2E format:

49. *Digitised Manuscripts to Europeana – Final Public Report (1st February 2012 – 31st January 2015)*: <http://dm2e.eu/files/Public_Final_Report_150630.pdf>; *The DM2E Model Specification: The textual model description* <http://dm2e.eu/files/DM2E_Model_V1.2.pdf)>; *The owl ontology* <https://github.com/DM2E/dm2e-ontologies/blob/master/src/main/resources/dm2e-model/DM2Ev1.2.owl>; *The web Representation* <http://onto.dm2e.eu/>.

50. Oracle Corporation, *MySQL Community Edition,* <https://www.mysql.com/products/community/>. For a layman's overview see «MySQL», *Wikipedia,* <https://en.wikipedia.org/wiki/MySQL>.

51. Discussed specifically in Charles B. Faulhaber, «Philobiblon: Problems and Solutions in a Relational Data Base of Medieval Texts», *Linguistic & Literary Computing,* 6 (1991), pp. 89-96.

dm2edata:agent/https://viaf.org/viaf/39642259/
dm2edata:agent/http://pb.lib.berkeley.edu/saxon/SaxonServlet?source=BETA/Display/1031Person.xml
skos:prefLabel "Santillana,_Íñigo_López_de_Mendoza,_Marqués_de,_1398-1458"@es
skos:altLabel "Íñigo López de Mendoza, 1. marqués de Santillana"@es
skos:altLabel "Íñigo López de Mendoza"@es
rdaGr2:gender "hombre"@es
rdaGr2:titleOfThePerson "1. marqués de Santillana"@es
edm:begin "1445-08-08"
edm:end "1458-03-25"
rdaGr2:titleOfThePerson "1. conde de Manzanares el Real"@es
edm:begin "1445-08-08"
edm:end "1458-03-25"

The prefixes in each element refer to W3C «namespaces» where these prefixes are defined:

edm: <http://www.europeana.eu/schemas/edm/> [European Data Model]
skos: <http://www.w3.org/2004/02/skos/core#> [Simple Knowledge Organization System]
dc: <http://purl.org/dc/elements/1.1/> [Dublin Core]
dm2edata: <http://data.dm2e.eu/data/> [Digitized Manuscripts to Europeana]
rdaGr2: <http://RDVocab.info/ElementsGr2/> [Open Metadata Registry]
@es [the language is Spanish]

PhiloBiblon has its own internal controlled vocabulary for such objects as watermarks, graphics, and codicological features. That controlled vocabulary will have to be enhanced and modified as needed by access to thesauri and taxonomies from external sources, e.g.:

1. *Enluminures* for the description of illuminated manuscripts: <http://www.enluminures.culture.fr/documentation/enlumine/fr/>
2. *Vocabulaire codicologique* for codicological terms: <http://www.palaeographia.org/cipl/gloss.htm>
3. *Getty Art &Architecture Thesaurus* for terminology concerning styles, design motifs, graphics representations: <http://www.getty.edu/research/tools/vocabularies/aat/index.html>
4. *Bernstein / The Memory of Paper*, for watermarks: <http://www.memoryofpaper.eu/BernsteinPortal/appl_start.disp>

There are other desiderata that go beyond the mapping of data, the incorporation of external controlled vocabularies, and the replication of existing

PhiloBiblon functionality. To be truly successful PhiloBiblon must once again place itself at the cutting edge of database technology. For researchers this means a virtual and collaborative research environment in the cloud, with:

1. Powerful tools for record creation in a crowd-sourcing environment.[52]
2. Faceted and semantic search facilities.[53]
3. Tools to work with texts, add metadata, and link them to the semantic cloud.[54]
4. A report generator capable of exporting selected or complete data in multiple formats, e.g.: TEI MsDescr, MARC, T_EX, XML.[55]
5. Visualization modules based on Big Data models.[56] While the number of records in PhiloBiblon is not large by Big Data standards (megabytes rather than terabytes), it is still too large to discern patterns easily, e.g., prosopographical projection in chronological (timelines) and geographical perspective (maps).

52. The most famous example of crowdsourcing is Wikipedia, but the technique is currently used, e.g., to allow untrained users to transcribe Greek papyri in the Ancient Lives Project: <http://www.papyrology.ox.ac.uk/Ancient_Lives/>.

53. Faceted search is the process of filtering the results of an initial search, on the web or in an OPAC, through a series of data elements such as author, language, date, or format. See «Faceted Search», *Wikipedia*, <https://en.wikipedia.org/wiki/Faceted_search>.

54. The earliest set of tools for textual analysis was Textual Analysis Computing Tools (TACT), created by Ian Lancashire at the University of Toronto for use under the DOS operating system: <http://projects.chass.utoronto.ca/tact/>. His work has been extended by Geoffrey Rockwell (University of Alberta) and his colleagues, first in TAPoR (Text Analysis Portal for Research), which now serves as a repository for almost 500 tools for textual analysis and data mining: http://tapor-test.artsrn.ualberta.ca/home, and most recently in Voyant Tools: <http://docs.voyant-tools.org/>.

55. TEI MsDescr is the Text Encoding Initiative's DTD for the description of manuscripts: *P5: Guidelines for Text Encoding and Interchange. 10 Manuscript Description* <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>. MARC, maintained by the Library of Congress, is the international standard for the description of bibliographical resources: <https://www.loc.gov/marc/>. T_EX, invented by Donald Knuth (Stanford University), is a universal typesetting system for computers. See «TeX», *Wikipedia*, <https://en.wikipedia.org/wiki/TeX>. XML, eXtended Markup Language, «is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable», «XML», *Wikipedia*, <https://en.wikipedia.org/wiki/XML>.

56. The foundational text for visualization techniques is Edward R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed., Cheshire, CT, Graphics Press, 2001. For a discussion of the difference between «big data» and «smart data» see Christof Schöch, «Big? Smart? Clean? Messy? Data in the Humanities», *Journal of Digital Humanities*, 2.3 (2013) (http://journalofdigitalhumanities.org/2-3/big-smart-cleanmessy-data-in-the-humanities/). For geospatial mapping see Todd Presner and David Shepard, «Mapping the Geospatial Turn», in *A New Companion to Digital Humanities*, eds. Susan Schreibman, Ray Siemens and John Unsworth, Chichester, West Sussex, UK, John Wiley & Sons, 2016, pp. 247-59. For visualization issues see Johanna Drucker, «Graphical Approaches to the Digital Humanities», in *A New Companion to Digital Humanities*, eds. Susan Schreibman, Ray Siemens and John Unsworth, Chichester, West Sussex, UK, John Wiley & Sons, 2016, pp. 290-302.

To summarize: Four decades of radical change in the computing environment have compelled PhiloBiblon to evolve as well in order that its component databases —significantly increased in size and density and detail— might remain a viable and vital resource for the study of medieval and early modern Iberia. PhiloBiblon's evolution serves as an object lesson of what is now a normal feature of all digitally based projects: the necessity of migration from one hardware and software platform to another as those platforms change and successively more powerful search and display tools become available. Currently, the distributed and decentralized model of the semantic web is the only cost-effective method of data acquisition and dissemination. Will that change? We do not know, but project developers should assume that it will and plan accordingly.

It is important to point out that the single most important restriction on development of PhiloBiblon has been the availability of funding for software development. The progress that has been made over the last forty years has been due almost entirely to grants from the U.S. federal government's National Endowment for the Humanities, although NEH support has been supplemented by institutional support from the University of California's Berkeley and Santa Barbara campuses, and by smaller grants from the Mellon, Pine Tree, Skaggs, and Delmas foundations in the U.S. and the MAPFRE Tavera and Ignacio Larramendi foundations in Spain.

In order to maintain PhiloBiblon as a free web resource open to scholars, students, and the general public around the world without regard to ability to pay, it will first be necessary to bring the medieval world of PhiloBiblon into the modern world of the semantic web.