Élika Ortega, Juan Luis Suárez, David Brown / Redes textuales: bases de datos en grafo para estudios literarios

Introducción

La dimensión estética de los textos literarios con frecuencia se plantea en un aparente antagonismo con la información discreta y relativamente inequívoca de una base de datos. Los ejercicios interpretativos que han sido por mucho tiempo la base de los estudios literarios, con frecuencia, también se presentan como opuestos a los exámenes cuantitativos que, en primera instancia, ofrecen los análisis de datos. Además, la creencia popular es que las bases de datos son instrumentos mayoritariamente de almacenamiento, catalogación y acceso, y no entidades a partir de las cuales se realizan análisis y procesos de interpretación.

La investigación literaria en bases de datos tiene una larga historia en el contexto hispano. Desde 1984, Charles Faulhaber realizó esfuerzos pioneros en *PhiloBiblon*, una colección que ya recogía otras colecciones de textos antiguos españoles, catalanes, gallegos y portugueses que permitió al equipo involucrado realizar análisis tipográficos y formatear los textos tanto para versiones electrónicas como para impresas, entre muchos otros avances. No obstante, como el propio Faulhaber reconoce, «cada solución presenta un nuevo problema» (1991: 95; nuestra traducción) y gracias a las dificultades con las que el equipo de *PhiloBiblon* se ha ido encontrando desde hace treinta años, hemos aprendido mucho. Dos de los retos más importantes que se mencionan en su artículo son la dificultad de establecer relaciones semánticas entre los distintos datos y el nivel de complejidad necesario en una base de datos para hacer justicia a las muchas particularidades de los textos incluidos en este caso en *PhiloBiblon* (1991: 94-95). No obstante, se apreciará que por extensión lo mismo es aplicable a cualquier otra base de datos.

No hay duda, entonces, de que la elaboración de bases de datos para el estudio de fenómenos humanísticos requiere de una reflexión conceptual y una técnica cuidadosa, así como de que se han realizado muchos esfuerzos para desarrollar metodologías y modelos abstractos que se presten al estudio de textos literarios. Esfuerzos que conciernen tanto a informáticos como a críticos literarios. Cierto es también que las características propias de diversas estructuras computacionales como las bases de datos relacionales o las bases de datos en grafo pueden tener una influencia en cómo se modela un fenómeno humanístico y, luego, en cómo seremos capaces de interpretarlo, analizarlo y de qué forma reutilizarlo. De esta forma, el trabajo pionero de Faulhaber llevado a partir de bases de datos relacionales tiene cualidades particulares que permiten realizar de manera certera determinadas preguntas de investigación —consultas, en lenguaje más técnico— mientras que otro tipo de bases de datos admitirán otras.

Bases de datos en grafo y estudios literarios

La metodología de análisis que hemos adoptado fructíferamente en el CulturePlex Lab tiene su cimiento en las bases de datos en grafo. Una base de datos en grafo se constituye como una colección de nodos y las relaciones que los conectan; por eso es ideal para representar e interpretar datos conectados, es decir, aquellos que requieren ser entendidos en sus relaciones y no de forma individual (Robinson *et al.*, 2013: 1-2). Además, la estructura de las bases de datos en grafo permite añadir tanto a nodos como a relaciones un sinnúmero de propiedades, una funcionalidad que permite realizar caracterizaciones muy granulares de los fenómenos que se estudian.

Los casos de estudio que presentamos aquí tienen su origen en el modelado de datos —en sí, el diseño de una base de datos—. El modelado de datos constituye una aproximación que lleva de manera adjunta una aproximación teórica y esboza qué elementos se tienen en cuenta y cuáles no, partiendo del supuesto de que no tenemos la capacidad de modelar el mundo real en su totalidad. Por ello, un diseño bien informado y acotado es esencial. Además, la posibilidad de iterar varias veces el diseño del esquema de una base de datos determinada es necesaria para conseguir que refleje nuevos descubrimientos, nuevas formas de pensar y para aproximarse al fenómeno bajo examen. El diseño y análisis de redes textuales a partir de bases de datos en grafo que proponemos ofrece esta flexibilidad y una granularidad de análisis aplicable a fenómenos radicalmente distintos.

La puesta en marcha de esta metodología nos ha llevado a desarrollar herramientas que aprovechan las potencialidades de las bases de datos en grafo con la investigación humanística. SylvaDB, desarrollada en nuestro laboratorio, potencia la flexibilidad de estas en tanto que facilita el diseño de los esquemas, su manipulación, y desarrollo a lo largo de las varias iteraciones. Además, SylvaDB permite añadir detalles sutiles sobre los datos recogidos y que caractericen tanto a los nodos como a sus relaciones; esto añade un nivel de complejidad semántica que muchas bases de datos relacionales no soportan. Así, el resultado es un modelo computacional que representa de la forma más precisa posible el objeto de estudio y, por tanto, se presta a consultas y miradas interpretativas diversas.

En los estudios que siguen, las redes textuales no se configuran a partir del cuerpo de un texto, sino de una serie de informaciones que entretejen la producción, distribución y recepción literarias tales como el género y el tema o el medio y el lugar de publicación, entre muchos otros. Estos datos, que siempre han sido parte de los estudios literarios, lejos de ser marginales y superficiales, nos permiten pensar en los textos de una forma innovadora, a una escala distinta y como parte de un ecosistema editorial, social y mediático, que tiene el potencial de iluminar aspectos de la producción literaria complementarios a la labor filológica. De esta forma, una metodología basada en el análisis de una base de datos en grafo permite la observación de relaciones intricadas. Estas, además, tienen su fundamento en un proceso de investigación, reflexión e interpretación tanto preliminar a la construcción de la

base de datos como posterior a los análisis cuantitativos. Los dos casos analizados siguieron el mismo proceso de conceptualización del fenómeno, identificación de sus elementos constitutivos y de las relaciones entre ellos, seguido del diseño y la construcción de una base de datos en grafo particular y su posterior análisis tanto cuantitativo como cualitativo.

Preliminares

El proyecto *Preliminares* recoge la gran riqueza de información contenida en las páginas preliminares de obras literarias (novelas, colecciones de poesía y teatro) publicadas durante los Siglos de Oro. Su objetivo primordial es entender los patrones de publicación y las redes de personas, lugares e instituciones involucradas en la producción de la literatura áurea. En la investigación bibliográfica hemos identificado documentos paratextuales como son la aprobación, carta, dedicatoria, erratas, poema, privilegio, licencia y tasa. La mayoría de estos documentos están firmados y datados por al menos un oficiante eclesiástico o civil e incluyen además de fecha y lugar de emisión, las afiliaciones institucionales de los signatarios. En el caso de las cartas, dedicatorias y poemas, los documentos comúnmente están dirigidos a algún noble, miembro del clero o autor, y están firmadas por sus autores. El objetivo inicial de *Preliminares* era construir una red de publicación que pudiera explicar, por ejemplo, la extensísima producción de Lope de Vega en comparación con la de otros autores, o bien los intercambios de objetos culturales y literarios a lo largo y ancho del Imperio español.

Este proyecto, además, presentó la dificultad de identificar ediciones específicas, ya que en la mayoría de los casos cada una denota un ecosistema de publicación propio. Un examen inicial del corpus estudiado nos permitió modelar un esquema de datos como el que se presenta en la figura 1.

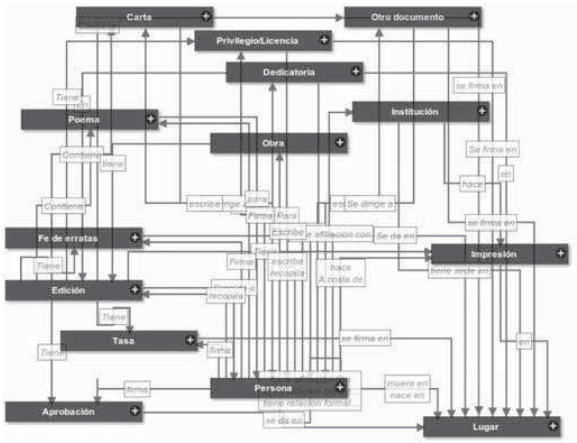


Figura 1. Captura de pantalla en SylvaDB del esquema del proyecto *Preliminares.*

Una vez diseñado el esquema, fue necesario poblar la base de datos con la información de cada obra y/o edición de forma manual revisando cada sección de los preliminares. Fue, además, durante esta etapa del proceso cuando comenzaron a emerger excepciones y anomalías, ya que los preliminares tienen, con frecuencia, inconsistencias, lo que requirió ajustes sutiles pero significativos en el esquema de la base de datos. En última instancia, desarrollamos un sistema de anotaciones adjunto a la base de datos que ilumina las ambigüedades que no pudieron ser resueltas. Actualmente, la base de datos de *Preliminares* continúa creciendo y sigue en proceso de desarrollo conforme ganamos acceso a nuevas ediciones y hacemos nuevos descubrimientos en nuestra investigación.

No obstante, el tipo de red textual desarrollada nos ha permitido hasta ahora realizar exámenes parciales de periodos delimitados por factores tanto prácticos como históricos, es decir, la recolección de un corpus lo suficientemente robusto para ofrecer conclusiones significativas sobre la administración del duque de Lerma entre los años de 1598 y 1618, por ejemplo. La búsqueda de mecenazgo, exposición e influencia, así como los rigurosos procesos de censura y licencia que dependían de miembros del clero y la nobleza son constantes y se observan a lo largo de los preliminares. Por lo tanto, gran parte de los resultados obtenidos de los primeros análisis sugieren que además del talento creativo de autores como Lope de Vega y Miguel de Cervantes, su posicionamiento social y político, así como sus habilidades estratégicas para dedicar e involucrar a ciertas figuras de la nobleza y el mundo literario —en sí, su lugar en la red de publicación (figura 2)— bien pudieron haber jugado un papel muy importante en su prominencia (Brown y Suárez, 2013).

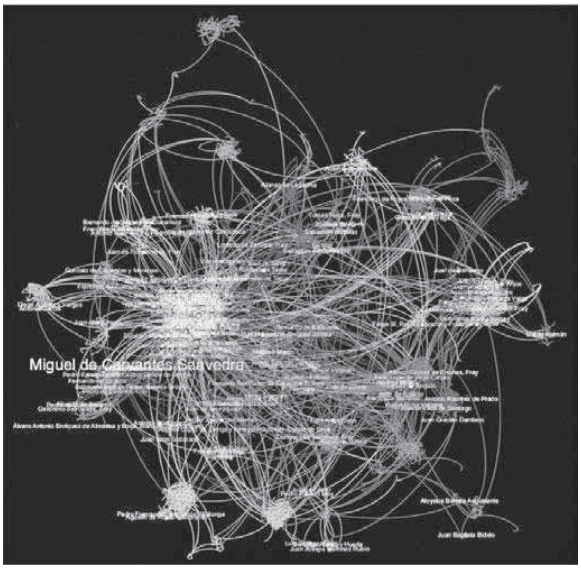


Figura 2. Red de publicación de Miguel de Cervantes.

Orsai

En contraste con el estudio de *Preliminares*, que busca explorar una red muy amplia a lo largo de un periodo que se extiende entre dos siglos, este segundo caso estuvo centrado exclusivamente en *Orsai*, un proyecto editorial que produjo desde España y Argentina una revista literaria-periodística durante tres años, mantuvo varios blogs y conformó una comunidad de lectores fieles en el mundo hispanohablante. El objetivo principal de esta investigación consistía en observar la construcción transmedia de *Orsai* y las relaciones textuales que mantenían la cohesión del proyecto a través de ☞

sus encarnaciones mediáticas y las distintas plumas de editores, colaboradores y autores invitados. Un segundo objetivo, basado en el éxito comercial de *Orsai*, fue la exploración de las prácticas lectoras en las distintas plataformas, impresas y electrónicas.

Dirigido por Hernán Casciari y Christian Basilis, el desarrollo paulatino de *Orsai* constituyó un laboratorio de pruebas en el que el esquema diseñado desde el inicio fue fructíferamente adaptado para reflejar su evolución. Desde su primera transición, de blog a revista en el año 2010, el paisaje de *Orsai* cambió de manera consistente.

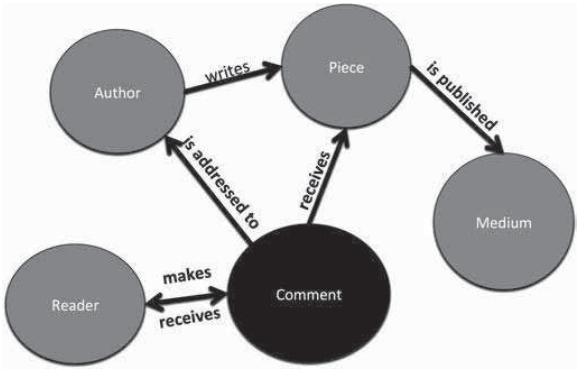


Figura 3. Esquema conceptual de la base de datos en grafo de *Orsai.*

El esquema inicial (fig. 3) nos permitió observar cómo *Orsai* se expandía y complejizaba (fig. 4). Este hecho constituye por sí mismo una prueba de la importancia que tiene crear un esquema de datos que refleje de forma certera y rigurosa las características del objeto de estudio y de aquí, por tanto, la necesidad de redimensionar la base de datos para incluir nuevos desarrollos. *Orsai* se examinó en sus componentes emergentes y fue posteriormente conceptualizado e interpretado como una red textual intermedial cuyo desarrollo se localiza en plataformas como blogs, aplicaciones, revistas impresa y electrónica e, incluso en lugares, como sucede en el caso de Orsai Bar de Buenos Aires.

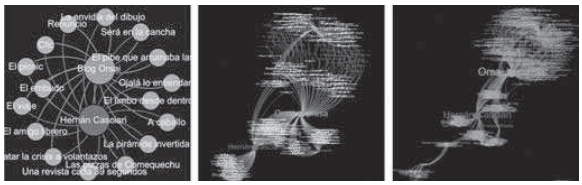


Figura 4. Desarrollo temporal y mediático de *Orsai.*

Los datos recabados nos permitieron investigar qué tipos de textos (qué géneros y qué contenidos) producían un mayor interés en los lectores y compradores de *Orsai* (fig. 5). Paso a paso, los resultados obtenidos dieron un giro a nuestra percepción inicial del proyecto, en el que las plataformas electrónicas parecían subsidiarias de la revista impresa y sus contenidos, y demostraron que el conjunto de plataformas y la variedad de piezas constituían un proyecto narrativo general: la mitología de la propia *Orsai*. El diseño de la base de datos también permitió la incorporación de un elemento que rara vez se incluye en los análisis literarios, los lectores de carne y hueso y sus marcas sobre la narrativa con la que han tenido contacto. Al hacer esta incorporación también fuimos capaces de apuntar que los intereses lectores, como se manifiestan en los comentarios publicados en las entradas de blog y las versiones web de la revista, estuvieron sobre todo centrados en la mitología y el desarrollo del propio proyecto. A partir de esta preferencia, concluimos que los lectores respondían con mucho entusiasmo a las piezas que los ubicaban como co-creadores y partícipes del proyecto, no solamente en su papel lector. El lugar renovado del lector en la ecología de medios de *Orsai* bien puede ser el aspecto que determinó su éxito y eventualmente su declive comercial a partir del tercer año (Ortega *et al.*, 2014).

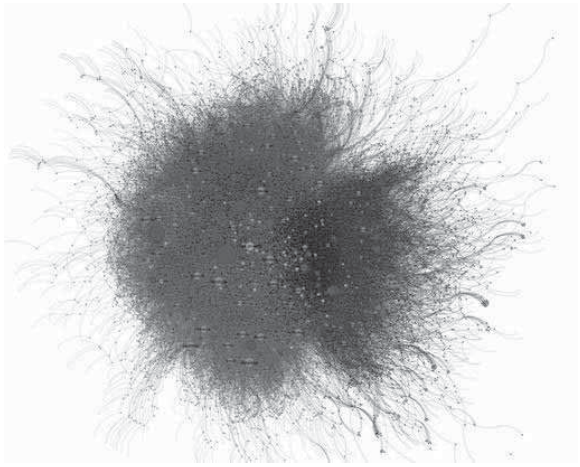


Figura 5. Red de *Orsai* que incluye la extensa participación de sus lectores y las conexiones entre los comentarios y las piezas publicadas.

Conclusiones

De los dos casos presentados aquí, se derivan varias cuestiones que merecen un análisis diferenciado. No solamente los fenómenos son distintos, sino también la complejidad de los esquemas diseñados para cada proyecto y la cantidad de datos analizados. El intricado esquema de *Preliminares* denota el concienzudo trabajo de investigación bibliográfica e interpretativa hecho previo al esquema; en cambio, en el esquema de la base de datos de *Orsai*, los elementos son menos y sus relaciones más simples. Además, el conjunto de datos es también específico a cada uno: en *Preliminares* aparecen unos dos mil nodos y en *Orsai* cerca de cuarenta y cinco mil. La diferencia de esquemas ilustra la adaptabilidad de la metodología a la aproximación particular de cada caso. Debe quedar claro, no obstante, que un esquema con más elementos no es indicativo de un proyecto más grande o más complejo conceptualmente, sino de las características del fenómeno en el mundo real.

En *The Language of New Media*, Lev Manovich establece una oposición entre las formas narrativas, como la novela y gran parte del cine, y las bases de datos; la prevalencia de estas ha llevado al teórico de medios ruso a proponer que las colecciones de datos estructurados, las bases de datos, son la forma clave de expresión cultural de la era computacional (2001: 194). Si bien el enfoque del trabajo de Manovich está centrado en la producción artística y no en los estudios académicos, es

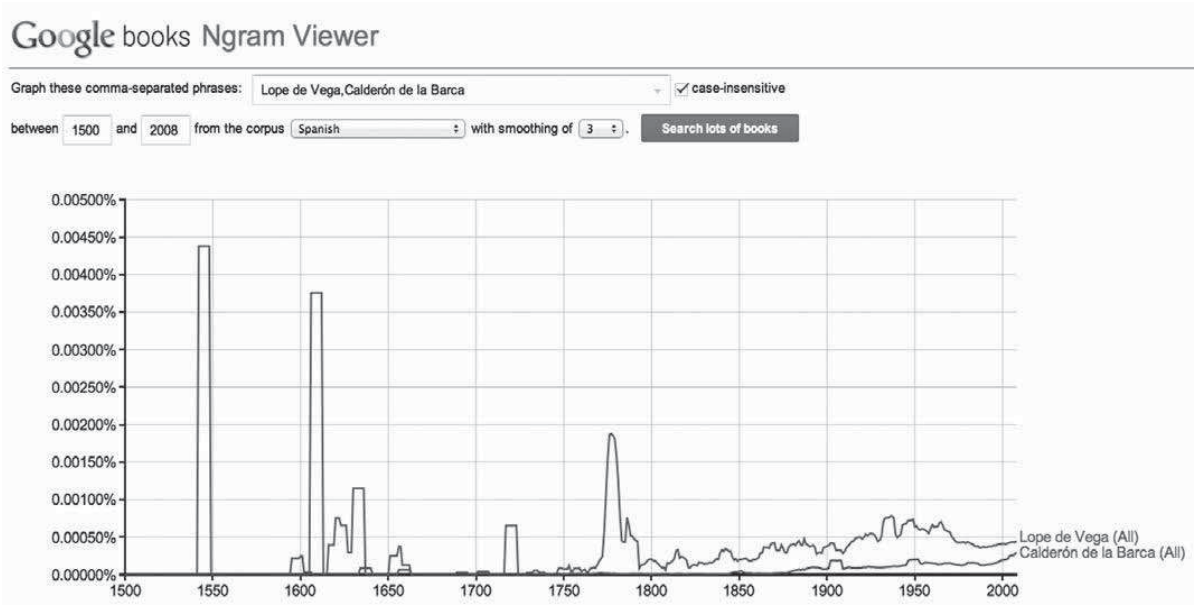
fácil observar que en nuestros días las bases de datos también se están volviendo cada vez más ubicuas en la investigación no solo literaria, sino también en los negocios y hasta la salud. La complejidad y la abundancia de datos que se crean en nuestros días por segundo —los *Big Data*— nos obligan a desarrollar y a pensar a través de nuevos modelos conceptuales y computacionales. De ahí que nuestra propuesta de adoptar las bases de datos en grafo para los estudios literarios sea apenas la punta del iceberg de la influencia que las investigaciones humanísticas como las que hemos descrito aquí podrían tener en otros ámbitos.

E. O., J. L. S. Y D. B.—CULTUREPLEX LAB, WESTERN UNIVERSITY, CANADÁ

Nàdia Revenga / La visualización de la literatura

Si alguna vez se ha preguntado cuántos libros diferentes se han publicado en el mundo a lo largo de la historia, Google le puede dar una respuesta: 129.864.880. Una cifra de 2010 que crece cada día, puesto que solo en España, en 2011, se publicaron unos 44.000 títulos diferentes, según la Federación de Editores Europeos. Google ha realizado este cálculo gracias a un algoritmo que combina información extraída de bibliotecas, WorldCat, catálogos colectivos nacionales, etc. Esta cifra despierta, irremediablemente, una sensación de desasosiego al saber el reducido porcentaje de libros que podremos leer a lo largo de nuestra vida: «*So many books, so little time*», reza una frase famosa atribuida al compositor Frank Zappa. Sin embargo, no es momento para el desánimo. El estudio de la literatura está cambiando gracias, en parte, a la tecnología, puesto que a través de ella es posible llevar a cabo nuevos métodos de análisis y hacer uso de sencillas herramientas para presentar los resultados de la investigación de una manera innovadora y efectiva a partir de mapas, gráficas, líneas de tiempo, etc. El número de textos que podremos leer de manera atenta y con detalle siempre será limitado, pero gracias a procedimientos tales como el de la «lectura distante», término acuñado por Franco Moretti (2007), y a las herramientas de análisis estadístico y visualización de grandes cantidades de datos, podremos, al menos, procesar la información y descubrir patrones y significados escondidos entre una multitud de datos.

Antes de entrar en materia sobre la visualización, veamos un ejemplo de cómo es posible encontrar detalles interesantes a partir de un extensísimo corpus textual. El visor N-Gram de Google permite buscar palabras o secuencias de palabras (hasta un máximo de cinco) en una parte del conjunto de libros del proyecto *Google books*, formado por más de veinte millones de títulos (en español, inglés, francés, alemán, hebreo, italiano, ruso o chino) y visualizar la frecuencia de aparición de una palabra o varias a través del tiempo (desde 1500 hasta 2008). Esta aplicación nos permite analizar las tendencias culturales en diferentes lenguas y descubrir qué palabras han tenido una mayor presencia en las publicaciones a lo largo de los siglos: por ejemplo, ¿quién es más popular en las publicaciones en español, Lope de Vega o Calderón de la Barca? ¿Shakespeare o Cervantes? ¿En algún momento de la historia Cervantes tuvo más popularidad que Shakespeare en los libros escritos en inglés?



Captura de pantalla de *Google books* Ngram Viewer.