El diseño del Archivo Digital Sebastià Juan Arbó: el empleo del modelo no relacional para la construcción de la base de datos bibliográfica

Joan Antoni Forcadell
Universitat Autònoma de Barcelona – GEXEL-CEDID
Joel Reverter Lainez
Universitat Autònoma de Barcelona – PROLOPE

Palabras clave: Sebastià Juan Arbó; archivo digital; base de datos; modelo no relacional; MongoDB
Keywords: Sebastià Juan Arbó; digital archive; database; non-relational model; MongoDB

Esta comunicación tiene como objetivo difundir los resultados de la primera fase de desarrollo del Archivo Digital Sebastià Juan Arbó, centrada en la elaboración de un prototipo de base de datos que reúne la bibliografía completa sobre el autor, con más de 8.500 referencias en su mayoría inéditas o escasamente conocidas. Aunque constituye solo una sección acotada dentro de lo que será la versión final de la web, esta fase ha permitido comprobar su integración y funcionamiento dentro de la estructura global del proyecto, además de hacer aflorar los principales problemas y retos que podrían surgir en las fases de desarrollo posteriores.

Sebastià Juan Arbó (la Ràpita, 1902 – Barcelona, 1984) fue una de las plumas más genuinas y representativas de la literatura española y catalana del siglo XX. Más allá de la relevancia del escritor en su época, una serie de condicionantes de base sociohistórica y política lo han arrinconado en el umbral de la marginalidad literaria, siendo todavía escasos los esfuerzos empleados para rescatar académicamente su figura y obra.

Uno de los condicionantes a los que se ha enfrentado la investigación al respecto ha sido la dificultad de localizar y recopilar evidencias documentales en fondos archivísticos, bibliotecas, centros de documentación y medios escritos publicados, una labor heurística que a menudo se ha afrontado de forma parcial —son intentos destacados, pero incompletos, las tesis doctorales de Ramis (2011) y Matas (2021)—. Esto ha contribuido a que haya quedado en la sombra su intensa actividad de creación literaria y periodística, así como su notable proyección pública como intelectual y figura destacada en el panorama cultural de los dos sistemas literarios de los que participó.

Rescatar del olvido a un escritor implica, junto con la edición, actualización y difusión de su obra, una tarea académica de cuidado sobre su legado para hacerlo accesible a lectores, docentes e investigadores. El proyecto de creación del Archivo Digital Sebastià Juan Arbó se plantea revertir definitivamente esta situación para poner a disposición pública en un portal web el conocimiento de toda evidencia material sobre cualquier faceta relacionada con el escritor, en la forma de una herramienta filológica útil para abastecer cualquier investigación relacionada con la vida y la obra del escritor. Facilitará el acceso en línea a un corpus gráfico, iconográfico, textual y bibliográfico integral reunido, clasificado, tratado informáticamente y disponible en abierto. Es el producto de una década de acondicionamiento, tratamiento documental, recopilación e investigación personales sobre el patrimonio documental del escritor en todas las instituciones que conservan vestigios archivísticos y hemerográficos del escritor, así como en cualquier medio de difusión pública en que estuvo presente.

Para abordar la primera fase de desarrollo, centrada en la presentación de una base de datos con la bibliografía completa de Sebastià Juan Arbó, se intentó en un primer momento replicar la metodología utilizada en diversos archivos digitales exitosos de características análogas, empleando tecnología SQL. Mediante PostgreSQL, se trazó un esquema de relaciones entre los diversos elementos que conformarían la base de datos.

Pronto, el esquema reveló una gran dificultad: clasificar un fondo tan vasto y variado como el que se ha recopilado resultaba problemático. Los registros eran tan diversos y heterogéneos que requerían características de catalogación distintas para poder almacenarse adecuadamente en una misma base de datos. Construir esta estructura en un modelo relacional implicaba diseñar un sistema complejo de tablas que pudiera representar las intrincadas relaciones entre los distintos elementos. Este problema, inicialmente referido al registro bibliográfico, se proyecta sobre los elementos de naturaleza heterogénea que eventualmente integrarán el archivo en su versión definitiva.

La adopción de una base de datos no relacional permitió superar estas dificultades y facilitó el trabajo a un equipo de investigación todavía reducido. Este modelo admite que cada registro tenga sus propios campos, ofreciendo una flexibilidad ideal para el proyecto. No obstante, esta misma flexibilidad puede derivar en errores por la ausencia de un esquema de datos fijo. El lenguaje empleado, MongoDB, permite integrar esquemas en las colecciones, exigiendo que cada registro contenga ciertos campos determinados. Si bien esto es útil para campos concretos —como el título del registro—, no resulta eficaz en nuestro caso, por la misma razón que descartamos el modelo relacional: la heterogeneidad de los datos dificulta la definición de un esquema sin sacrificar parte del rendimiento de la aplicación. Además, algunos registros acusan vacíos insalvables en ciertos campos, debido a la falta de información en origen.

Para preservar la integridad de los datos, se optó por introducir los registros inicialmente a través del gestor bibliográfico Mendeley. Esta herramienta permite trabajar con una interfaz intuitiva, operar en grupo, minimizar errores derivados

de la introducción manual —inevitable en este proyecto, dado que la mayor parte de la documentación es física y no puede importarse automáticamente—, clasificar las entradas en carpetas según once categorías generales y múltiples subcategorías, y exportar finalmente los registros generados.

Mendeley exporta los datos en formato XML, RIS y BIB, ninguno de los cuales es compatible con MongoDB. La interfaz gráfica Compass de MongoDB permite importar archivos JSON que genera automáticamente la base de datos con sus respectivos elementos. No existe, sin embargo, un método plenamente satisfactorio para convertir archivos XML, RIS o BIB a JSON sin errores significativos en el tratamiento de los datos. Esto se debe a que cada gestor bibliográfico utiliza estructuras distintas para describir campos equivalentes. No se ha encontrado ningún programa capaz de reconocer de forma fiable esas diferencias y equivalencias. Este problema afecta directamente a la estructura del documento y, por tanto, a la clasificación en la base de datos. Para resolverlo, se accedió a la API de Mendeley y se extrajeron desde sus servidores todos los registros en formato JSON. Esto devuelve un documento con varios registros. Cada registro, aunque varía según su tipología —libro, artículo de prensa, página web, etc.—, comparte una estructura general y varios campos comunes.

El JSON generado incluye algunos campos irrelevantes para nuestra base de datos, por lo que se diseñó un script en Python para limpiarlos y ajustarlos a las necesidades del proyecto. Se eliminaron los campos «folder_uuids», «hidden», «group_id», «confirmed», «private_publication», «profile_id», «read» y «starred». Además, Mendeley clasifica como «source» los libros que aparecen en los registros pertenecientes a un capítulo de libro. Esto genera ambigüedades, ya que este mismo campo se utiliza para identificar las publicaciones periódicas. Para evitar esta confusión, se reemplazó «source» por «book» en los capítulos. Mendeley tampoco distingue entre autor, editor, ilustrador, traductor, director y coordinador, que clasifica en el momento de la exportación indistintamente como «authors». También se desglosó el campo genérico «authors» en sus respectivos roles: «author», «editor», «illustrator», «coordinator», «translator» y «director». Asimismo, Mendeley exporta la UUID de la carpeta a la que pertenece el registro, pero no su nombre. Gracias al script, se recuperó no solo el nombre de la carpeta, sino también el de las carpetas superiores que constituyen la estructura jerárquica en que se clasifica el registro. Estos metadatos ayudarán a afinar las búsquedas en la base de datos.

Una vez obtenidos los datos depurados en formato JSON, se cargó el archivo en Atlas, el servicio en la nube de MongoDB. Se crearon índices para los campos numéricos —año, mes, día—, mientras que los campos de texto se optimizaron mediante el índice específico de Atlas Search. Este índice permite búsquedas textuales eficientes y ha sido configurado para tolerar mayúsculas y tildes, mejorando la precisión en consultas en español.

A pesar de ser un desafío complejo, una base de datos como la que aquí proponemos es interoperable con una relacional como las antemencionadas. De entre las diversas estrategias que existen, abogamos por la creación de APIs

que permitan a los diferentes archivos digitales exponer sus datos de manera estructurada y consistente. Esto habilitaría a proyectos externos consultar y federar la información de múltiples archivos sin necesidad de conversiones complejas. El acceso a la API de Mendeley es un ejemplo de este flujo de trabajo. Con todo, lo deseable sería trabajar con bases de datos no relacionales, dadas las ventajas expuestas respecto a las no relacionales.

En definitiva, el desarrollo de la base de datos que nutre el Archivo Digital Sebastià Juan Arbó, mediante el uso de un modelo no relacional, ha permitido resolver dificultades que con un modelo relacional y los recursos disponibles habrían resultado insalvables. Este enfoque ha brindado mayor flexibilidad en la entrada de datos, contribuyendo a una mejor optimización de la aplicación. El uso de Mendeley ha garantizado la integridad de los registros, y todo el proceso ha enriquecido los datos con metadatos que permiten búsquedas más precisas y abren nuevas posibilidades de escalabilidad. Los resultados obtenidos evidencian la eficacia de las bases de datos no relacionales en proyectos de desarrollo de archivos literarios digitales.

Bibliografía

BIA PLATAS, A., & RODRÍGUEZ SALA, J. (2017), «Construcción de una base de datos y un repositorio de documentos de investigación para el proyecto TRACE», *Revista de Humanidades Digitales*, 1, pp. 287–295.

COHEN, S. (2017), «Apertura, difusión y accesibilidad: el uso teórico y práctico de una base de datos», en R. LÁZARO NISO (ed.), *Corpus y bases de datos para la investigación en literatura*, Fundación San Millán de la Cogolla, pp. 35–48.

GONZÁLEZ SORIANO, J. M. (2021), «Mnemosine: Biblioteca digital de la otra Edad de Plata (orígenes, contenidos, perspectivas)», *Signa: Revista de la Asociación Española de Semiótica*, 30, pp. 31–58.

MATAS ROCA, M. (2021), *Sebastià Juan Arbó: De la realitat viscuda a la ficció narrativa. Anàlisi d'un desarrelament en la literatura catalana*, Universitat Rovira i Virgili.

MELO FLÓREZ, J. A. (s. f.), *De las fichas de archivo a las bases de datos*. Recuperado el 24/04/2024 de https://taller-abierto-de-humanidades-digitales.github.io/bases_de_datos/

RAMIS, J. M. (2011), *Autotraducció i Sebastià Juan Arbó*, Universitat Pompeu Fabra.

ROMERO LÓPEZ, D. (2014), «Hacia la Smartlibrary: Mnemosine, una biblioteca digital de textos literarios raros y olvidados de la Edad de Plata (1868–1936). Fase I», en S. LÓPEZ POZA & N. PENA SUEIRO (eds.), *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, Universidade da Coruña y SIELAE, pp. 411–422.

SANTOS ZAS, M., MARTÍNEZ RODRÍGUEZ, F., & MASCATO REY, R. (2014), «La creación y gestión del archivo digital valleinclaniano: corpus manuscrito

e impreso», en S. LÓPEZ POZA & N. PENA SUEIRO (eds.), *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, Universidade da Coruña y SIELAE, pp. 435–457.

SIMONE KIRBY, J. (2019), «How NOT to create a digital media scholarship platform: The history of the Sophie 2.0 project», *IASSIST Quarterly*, 42(1), pp. 1–16.

TOMASI, G. (2020), «Las humanidades digitales y la base de datos MeMoRam para un enfoque sistemático hacia los motivos en los libros de caballerías», *Historias Fingidas*, 8, pp. 129–156.