Integración de procesos técnicos en la creación de objetos digitales de prensa histórica

Xavier Agenjo
Proyectos de la Fundación Ignacio Larramendi y de la Polymath Virtual Library
Madrid, Spain
E-mail: xavier.agenjo@larramendi.es

Jesús Domínguez
DIGIBÍS, Producciones Digitales
E-mail: jesus.dominguez@digibis.com

Francisca Hernández
DIGIBÍS, Producciones Digitales
E-mail: francisca.hernandez@digibis.com

Fernando Román
DIGIBÍS, Producciones Digitales
E-mail: fernando.roman@digibis.com

Meeting:

188 — *Newspapers in the Caribbean, Central and South America: production, distribution and conservation. Cultural heritage and news media in the digital age* — Newspapers Section

Resumen:

Los proyectos de digitalización y de creación de bibliotecas digitales o virtuales pueden ofrecer grandes ventajas para la integración de procesos técnicos de todo tipo, en especial del proceso de catalogación. La catalogación detallada de prensa histórica requiere la aportación de un gran esfuerzo por parte de los catalogadores y frecuentemente su descripción queda reducida a un registro global en el que se mencionan los fondos disponibles de una manera comprimida. Sin embargo, la organización de los procesos complementarios a la digitalización y la disposición de las herramientas informáticas necesarias permite que, en el mismo momento de la validación y control de calidad de las imágenes, puedan asignarse una serie de elementos descriptivos adecuados a las unidades componentes básicas de una publicación periódica. Así, en la validación hoja a hoja de una manifestación digitalizada, se puede llegar a disponer, mediante un sistema automático de retroalimentación, de una catalogación de estas unidades componentes, lo que supera ampliamente las habituales menciones de fondos y ejemplares comprimidos.

Se presentan en esta comunicación un subconjunto de etiquetas del formato MARC 21 adecuadas para crear registros de números, que podría entrar en la denominación de partes componentes, pero con un cariz nuevo ya que la descripción no se realiza de forma manual

sino que procede del proceso de digitalización y validación de un recurso continuo. El objetivo de estos registros obtenidos automáticamente no está sólo relacionado con la descripción bibliográfica que se mostrará en el OPAC, sino también con la estructuración física del objeto digital en sus unidades componentes básicas. Estos registros están destinados a las aplicaciones de indexación, búsqueda y presentación de los objetos digitales más que como información directa al usuario.

La comunicación muestra el uso de estos registros para la búsqueda de objetos digitales por fecha, para la delimitación de búsquedas a texto completo o sobre registros bibliográficos en grandes volúmenes de información a partir de la indexación al nivel más bajo de enumeración y cronología previsto en el modelo de una publicación concreta. Se detallan las ventajas que presentan estos registros para una búsqueda mucho más rápida de un número concreto, indexado individualmente, en el caso de que se haya llevado a cabo un reconocimiento óptico de caracteres o incluso una segmentación del texto, especialmente en aquellos casos en que el recurso continuo sea muy extenso como es frecuente en la prensa diaria que se extiende a lo largo de los años, de las décadas e incluso de los siglos.

Estas estructuras facilitan también completar virtualmente las manifestaciones digitales de un título concreto pues marca cada manifestación en el nivel más bajo posible de enumeración y cronología y, de igual modo, posibilita la preservación a largo plazo de manifestaciones virtuales procedentes de varias fuentes. Por todo ello, la descripción de los niveles más bajos de enumeración y cronología es especialmente adecuada para proyectos cooperativos de digitalización o de preservación digital. Incluso puede favorecer el desarrollo de proyectos de corrección cooperativa al modo de Trove o de asignación social de etiquetas de materias, lo cual contribuirá en un futuro próximo a la skosificación de éstas y a la transformación de datos bibliográficos como Linked Open Data.

Introducción

Esta comunicación está basada tanto en la reflexión teórica como en la enseñanza derivada de una práctica continuada de digitalización de prensa histórica que se remonta a más de una década con la digitalización, entre otros proyectos afines, de 123 cabeceras para la Junta de Castilla-La Mancha y que, posteriormente, se integraron en la Biblioteca Digital de Castilla-La Mancha. En realidad, fue el análisis biblioteconómico lo que ha modificado radicalmente estos procesos de digitalización; sus principios estaban enunciados en una serie continuada de concursos públicos promovidos fundamentalmente por la Subdirección General de Coordinación Bibliotecaria del Ministerio de Cultura de España y seguidos muy de cerca por diversas Comunidades Autónomas españolas, así como por otras instituciones de memoria repartidas por toda España.

Determinados estudios bibliográficos publicados tanto en revistas de biblioteconomía como en introducciones a catálogos de prensa histórica[1] fueron recogidas por esas instituciones que promovieron dichos procesos de digitalización. Sobre todo ello existe abundante bibliografía, pero para abreviar la consulta de la misma nos remitimos a un trabajo en concreto dónde puede encontrarse una reflexión y una recapitulación que viene a constituir un resumen de lo

[1] Véase la introducción al *Catálogo de publicaciones periódicas de Cantabria de la Biblioteca Municipal de Santander (1809-1996)*. - [Santander] : Ayuntamiento de Santander, Concejalía de Cultura y Deporte, 1997. pp. 11-15, redactado por uno de los autores de esta comunicación.

dicho[2]. Desde un primer momento, estuvo claro que una de las razones, de no poca importancia, para realizar un proyecto sistemático de digitalización de prensa se basaba en el hecho de que el papel en el que se halla impresa la prensa, haciendo exclusión de la anterior al primer tercio del siglo XIX -mucho menor en cuanto a su volumen-, se encontraba en grave riesgo de desaparecer a causa de su bajo pH. Estos problemas habían sido detectados ya mucho antes, como es lógico, pero lo más significativo era, como se señalaba ya a finales de los años 80, la imposibilidad de afrontar la desaparición de las colecciones hemerográficas con las técnicas de desacidificación masiva[3]. En primer lugar, la desacidificación masiva fue quedando relegada por sus efectos medioambientales; en segundo lugar, era imposible atajar el problema con métodos a menor escala y, en tercer lugar, resultaba necesario implantar procesos masivos de reforzamiento del papel irremediablemente dañado. La inviabilidad técnica y los costes económicos y de tiempo hicieron que, en pocos años, se hiciera hincapié en la preservación de la información por medio de la digitalización de textos, lo que caló, al menos, en un grupo muy significativo de bibliotecarios españoles situados en puestos clave para llevar a cabo los proyectos de digitalización a los que nos estamos refiriendo.

También la práctica biblioteconómica de buena parte de los bibliotecarios e informáticos que se vieron implicadas en la puesta en marcha de todos estos proyectos hacía evidente que el manejo de enormes cantidades de prensa encuadernada en grandes formatos era una de las tareas más penosas que podían llevarse a cabo en una biblioteca o hemeroteca y que repercutían enormemente en el trabajo diario, tanto desde el punto de vista de su proceso técnico como de su almacenamiento o en el servicio a los usuarios. Por último, la experiencia demostraba que una vez que el usuario había logrado dar con el resultado apetecido iba a solicitar una reproducción, tarea también dificultosa dadas las dimensiones de gran parte de la prensa, lo que, desde luego, no contribuía en nada a la preservación y conservación de la misma.

Por ello, en 2003, a instancias del Ministerio de Educación y Cultura (ese era su nombre en aquellos momentos), se decidió crear simultáneamente un *proyecto de digitalización* de prensa y una *biblioteca virtual* dónde esa prensa fuera consultada. El proyecto se concibió ya desde un principio con una doble vertiente, *digitalizar* prensa y *crear* una biblioteca virtual

---

[2] Agenjo Bullón, Xavier. *La cuarta salida de El monje digital y sus problemas hemerográficos : una recapitulación*. En: *Boletín de la ANABAD*. Tomo 54, nº 4 (2004): 119-138.

[3] Hernández Carrascal, Francisca. *Panorama general de las técnicas de desacidificación masivas*. En: *Boletín de la ANABAD*, Tomo 42, Nº 2, 1992 , pp. 123-133. Véase también *La Biblioteca Digital Nacional: alcance y límites de un proyecto*. En: *Las bibliotecas virtuales y la digitalización*. - 3 casetes. - (Actos culturales en la Biblioteca Nacional) 27 de Febrero de 1996. - Contiene: *La memoria hispánica* / Carlos Ortega -- *La Biblioteca Digital Nacional : alcances y límites de un proyecto* / Xavier Agenjo--Internet y su futuro / Ángel Casado -- *Las publicaciones electrónicas y el acceso al documento primario : nueva fase del Catálogo Colectivo de Publicaciones Periódicas* / Francisca Hernández--*Sistemas de gestión de bases de datos bibliográficas y soluciones informáticas avanzadas de archivos digitales* / Javier Berlana -- *El formato SGML y la investigación asistida por ordenador* / Álvaro Klasse -- *Internet : la 3ª ola de información* / Isidro Cano.

Con motivo de la presentación recogida fonográficamente, se repartió al público una carpeta titulada *Memoria Hispánica*. Madrid, Biblioteca Nacional, 1996. 1 carpeta (9 h.). [s.i.t.], aunque impreso en Madrid el 28 de marzo de 1996, día de Santa Esperanza. Hoy constituye una rareza bibliográfica, pero en ella se hacía una especial referencia al problema que suponía la acidez de los libros impresos con papel elaborado industrialmente. Quizá el lector se extrañe de que no coincida el día de la grabación fonográfica con el de la fecha de impresión, informal, sin duda fruto del buen humor de alguno de los participantes que, obviamente, se refería a la conocida virtud teologal. Quizá ahora haya que verlo desde el punto de vista de las otras dos, o bien de la caridad, o bien de la fé, lo que se deja al buen juicio de quién esto leyere.

de prensa. Hay que tener en cuenta que cuando se iniciaron estos procesos ya se habían producido dos fenómenos muy importantes que pueden resumirse en la iniciativa *Digital Libraries* promovida por la *Agenda de Lisboa* y en los principios de Lund, es decir, que la Biblioteca Virtual de Prensa Histórica ya se concibió también con una perspectiva europea, lo que se ha ido acentuando progresivamente y que ha culminado con la publicación de la *Agenda Digital Europea* el 26 de agosto de 2010[4].

Desde un punto de vista rigurosamente histórico, la Biblioteca Virtual de Prensa Histórica estuvo precedida por la Biblioteca Virtual de Derecho Aragonés y, sobre todo, por la Biblioteca Virtual de Andalucía, donde ya empezaron a ensayarse los procedimientos que fueron cuajando con el paso del tiempo. Conviene señalar que aunque muchas de estas iniciativas son fruto de la licitación pública, también la empresa DIGIBÍS, donde se han realizado la mayoría de los proyectos a los que vamos a hacer referencia y cuya práctica se va a describir, colaboró a hacerlos posibles. No cabe duda de que se ha producido a lo largo de todos estos años un considerable mimetismo entre diferentes proyectos de digitalización de prensa histórica promovidos por distintas instituciones y que estos se han incorporado sucesivamente a esta solución estratégica.

Sin duda, el proyecto *Chronicling America*[5] ha tenido una considerable influencia en todas las iniciativas a las que se refiriere este texto, pero es fundamentalmente la propia práctica de la digitalización de prensa la que ha llevado a elaborar el modelo de catalogación que aquí se presenta. El número de páginas de prensa digitalizadas conforme a estos procedimientos supera la cifra de 7 millones. Todo ello se inició con la digitalización de las 123 cabeceras y 126.672 páginas de prensa histórica de Castilla-La Mancha a finales de la década de los 90. Cuando se abordó ese proyecto todavía no estaba en el ánimo de ninguna institución llegar a constituir una biblioteca virtual, pero sí que fue preciso incorporar unas estructuras de información mucho más amplias y precisas que las que se habían utilizado anteriormente.

En 2003 se inició la Biblioteca Virtual de Prensa Histórica -promovida por el Ministerio de Cultura a instancias de Carmen Caro y mantenida y aumentada por María Antonia Carrato, siempre bajo la detallada supervisión de María Luisa Martínez-Conde[6]- que en el momento de redactar esta comunicación, mayo de 2011, supera los 5 millones de páginas digitalizadas (5.420.480 y 962.572 números) correspondientes a 1.950 cabeceras de prensa histórica procedentes de 59 bibliotecas, lo que dobla el número de páginas digitalizadas en *Chronicling America*. También se ha digitalizado prensa histórica para la Biblioteca Virtual de Andalucía (293 cabeceras y 430.389 imágenes), para la Biblioteca Virtual de Aragón (20 cabeceras y 37.822 imágenes), para la Biblioteca Digital de Madrid (82 cabeceras y 324.225 imágenes), para el Centro de Documentación de MAPFRE (332 cabeceras y 8.667 imágenes), para la Biblioteca Virtual de Asturias (51 cabeceras y 43.027 imágenes), para la Biblioteca Virtual de Aranjuez (281 cabeceras y 57.405 imágenes) y de forma análoga, aunque mucho más dificultosa, pues fue necesario un laborioso mapeo de estructuras de información previo, para Galiciana: Biblioteca Digital de Galicia (279 cabeceras y 735.667 imágenes)[7].

[4] http://goo.gl/6Xsvb
[5] http://chroniclingamerica.loc.gov/

[6] Conviene recordar los nombres de quienes han iniciado y mantenido iniciativas ambiciosas, con amplitud de miras y perspectiva estratégica.

[7] Puede consultarse la relación de bibliotecas digitales o virtuales desarrolladas por DIGIBÍS en http://goo.gl/o4hGU

Los más de 7 millones de páginas de prensa digitalizadas disponen de diferentes esquemas de metadatos: descripción bibliográfica y de ejemplares según el formato MARC 21 para registros bibliográficos y para registros de fondos y localizaciones, menciones de fondos según la norma ISO 10324[8], así como el uso de Dublin Core[9], METS[10], PREMIS[11], METSRights[12], y últimamente DOI[13], lo que ha dado lugar a un conocimiento verdaderamente detallado de la problemática que puede surgir tanto en la digitalización de prensa como en la conversión de prensa digitalizada de forma no normalizada. Como se ha dicho, estas cifras superan las de *Chronicling America*, proyecto que, en cualquier caso, constituye el modelo que se ha tenido siempre en mente y cuya estela se ha querido seguir, máxime cuando lejos de permanecer ajeno a los cambios en las estructuras de información se ha convertido ya en un *dataset* de Linked Open Data[14] y dispone de una API[15] que lleva a cabo una serie de funciones muy útiles. Justamente cuando se redactan estas páginas se acaba de producir el 28 de mayo de 2011 una nueva actualización de la interfaz de este proyecto.

En resumen, se considera que la experiencia extraída del desarrollo de proyectos de digitalización de prensa de gran envergadura se ajusta al tema y objetivos que la Sección de Prensa se ha planteado para la 77ma. Conferencia General y Asamblea de la IFLA de Puerto Rico y se centra en *Los periódicos en el Caribe, América Central y del Sur: Producción, Distribución y Conservación: Patrimonio Cultural y Medios de Comunicación en la era digital*. La sección pretende compartir conocimientos sobre la edición de periódicos y los esfuerzos de los bibliotecarios por recoger, preservar y presentar los periódicos a su público como un patrimonio cultural. Además, como es obvio, del especial interés que las colecciones virtuales de prensa histórica española tienen como fuente de información para la historia de América, hay que considerar estas colecciones desde el punto de vista de sus módulos componentes: gestión de imágenes, búsqueda y recuperación, catalogación, pero también el control del *workflow* que retroalimente en detalle la descripción pormenorizada de un periódico[16].

---

[8] ISO 10324:1997 Information and documentation -- Holdings statements -- Summary level

[9] http://dublincore.org/

[10] http://www.loc.gov/standards/mets/

[11] La versión 2.1 elimina determinadas redundancias, ya detectadas, con METS [http://www.loc.gov/standards/premis/v2/premis-2-1.pdf] Se dispone de una traducción al español de la versión 2.0 [http://www.loc.gov/standards/premis/PREMIS_es.pdf].

[12] http://www.loc.gov/standards/rights/METSRights.xsd

[13] http://www.doi.org/

[14] http://ckan.net/package/chronicling-america

[15] http://chroniclingamerica.loc.gov/about/api/

[16] Larramendi, Tachi [et al.] *Datos y metadatos : la normalización dinámica de los elementos y de los procesos constituyentes de una biblioteca virtual*. En: *Interinformación: XI Jornadas Españolas de Documentación* : 20, 21 y 22 de mayo de 2009, Auditorio Palacio de Congresos de Zaragoza. FESABID, 2009. 109-116. http://www.fesabid.org/zaragoza2009/actas-fesabid-2009/108-116.pdf

Bibliografía material de la prensa histórica

Un aspecto fundamental a la hora de la descripción bibliográfica es la identificación precisa de la obra. No escasean, al menos en lo que se refiere a España, publicaciones de toda índole, desde extensas monografías a breves artículos en los que se describe (presuntamente) la prensa de un determinado lugar. Sin embargo, con mucha frecuencia, esas obras parecen, más que bibliografías o catálogos, estudios o ensayos de carácter general sobre la evolución de las publicaciones periódicas y rara vez entran en el detalle de la descripción bibliográfica precisa. En España ha sido Juan Delgado Casado quién, gracias a sus monografías *Las bibliografías regionales*[17] e *Introducción a la Bibliografía*[18], ha hecho un balance más preciso sobre el estado de la cuestión.

Desde luego, tras el proceso continuado, desde 2003, de digitalización masiva de prensa histórica es absolutamente necesario realizar una revisión sistemática del estado del conocimiento que se posee sobre la prensa histórica española. En primer lugar, porque se han digitalizado casi 60 colecciones de prensa española y no sólo las grandes y conocidísimas colecciones hemerográficas, lo que supone una ampliación muy significativa del conocimiento de estas en España. En segundo lugar, porque la digitalización ha sido sistemática, lo que ha sacado a la luz títulos de cuya existencia no se tenía constancia. Y en tercer lugar, porque la descripción bibliográfica y de ejemplares se ha realizado de forma detallada con el ejemplar a la vista.

La catalogación de la prensa histórica ha sido desde siempre, lamentablemente, muy rudimentaria, centrada la mayor parte de las veces en unas descripciones menos que someras, compuestas poco más que por el título y una relación compacta de las fechas extremas de los números. De hecho, ninguna de las colecciones digitalizadas a las que hemos hecho referencia disponía de mejores descripciones, y en la mayoría de los casos éstas eran inexistentes, lo que ha obligado a rehacerlas de nuevo, de lo que se han extraído provechosísimas consecuencias. Es decir, el acceso a las colecciones se había realizado hasta su digitalización en base a un proceso de ordenación en las estanterías basado en el título y la cronología o en unas descripciones que poco o nada añadían a la ordenación. Las bibliografías ayudaban sólo relativamente puesto que en su mayoría no hacían referencia a la biblioteca que conservaba las colecciones. En resumen, la digitalización se inició sobre una ausencia casi total de catálogos o registros bibliográficos.

En 1986, y ya iniciado por métodos automatizados el Catálogo Colectivo del Patrimonio Bibliográfico español, se decidió emprender el correspondiente al del siglo XIX en el que las características formales del libro, como es sabido, cambian de la imprenta y la fabricación del papel artesanal a la industrial e *ipso facto* aumenta enormemente tanto el número de publicaciones como el de lugares de impresión. Por ello, se decidió llevar a cabo un estudio comparativo entre tipobibliografías y catálogos de un conjunto importante de grandes bibliotecas patrimoniales españolas, lo que arrojó –con un margen de error- unas diferencias notabilísimas que pueden resumirse en que muchas obras referenciadas bibliográficamente no conocían ejemplar y, todavía con más frecuencia, que en muchos catálogos aparecían obras no representadas en las bibliografías locales. Este tipo de estudios debería llevarse a cabo

[17] Delgado Casado, Juan. *Las bibliografías regionales y locales españolas : (evolución histórica y situación actual)*. - Madrid : Ollero y Ramos, [2003]. - 370 p.

[18] Delgado Casado, Juan. *Introducción a la bibliografía* : *(los repertorios bibliográficos y su elaboración)*. - Madrid : Arco/Libros, [2005] 297 p.- (*Instrumenta bibliológica. Serie A*)

sistemáticamente antes de cualquier proceso de reconversión o de digitalización retrospectiva, con el objetivo de detectar lagunas en las bibliografías locales y carencias en los catálogos de las bibliotecas sobre los que se va a llevar a cabo un proceso de digitalización[19]. Como es lógico, queda claro que una verdadera bibliografía local o tipobibliografía habrá de ser esencialmente virtual, pues ninguna biblioteca dispondrá de todos los fondos impresos o editados en un determinado lugar. Y así mismo, cualquier catálogo o biblioteca habrá de ser virtual, pues ninguna poseerá unos fondos íntegros, en el sentido estricto de la palabra. Es claro, que este proceso deberá ser iterativo con el fin de, progresivamente, disponer de unas bibliografías retrospectivas y unos catálogos virtuales lo más completos posibles. Obviamente, al final de esas iteraciones se obtendrán las publicaciones que pueden considerarse desaparecidas.

Evidentemente, la dificultad de catalogar prensa histórica ha producido, de forma general, que se prestara una atención preferente a un aspecto fundamental, el de los sucesivos cambios de nombres de títulos, muy frecuentes en publicaciones periódicas de largo aliento, así como a la modificación de la periodicidad de esas mismas publicaciones. Pero, por el contrario, se ha ocupado muy poco de la descripción número a número, incluso página a página, de una publicación periódica. En principio, podría pensarse que todo ello es imposible y que semejante labor tampoco se realiza con las monografías, también en principio, mucho más fáciles de tratar, pero nada más lejos de la realidad.

El modelo de descripción bibliográfica del fondo histórico que ha dominado la catalogación de los materiales bibliográficos antiguos en España, entendiendo por tal el comprendido en los límites establecidos por la *Ley de Patrimonio Histórico Español* (1985), ha estado muy fuertemente influido por la publicación modelo de Jaime Moll, denominada *Problemas bibliográficos del libro del Siglo de Oro*[20], que venía a ser una acertada síntesis de la moderna escuela bibliográfica francesa, nacida a partir de la obra fundamental *La aparición del libro*[21] de Henri Jean Martin y Lucien Febvre y, sobre todo, de la escuela de bibliografía material inglesa que tiene su epítome en el libro de Philip Gaskell *Nueva introducción a la Bibliografía material*[22]. Todos estos trabajos han dado lugar tanto al Catálogo Colectivo del Patrimonio Bibliográfico, sobre el que existe amplia bibliografía[23], como a la

[19] Andújar Velasco, Ananda; Agenjo Bullón, Xavier; Palá Gasós, Pilar. *Estudio preliminar para la confección del catálogo colectivo de obras impresas en el siglo XIX*. En: Boletín de la ANABAD. - Madrid : ANABAD. XXXVI, 3 (jul.- sept. 1986), p. 461-471.

[20] *Problemas bibliográficos del libro del Siglo de Oro* / Jaime Moll // Boletín de la Real Academia Española. - Madrid : Real Academia Española, 1914-. - ISSN 0065-0455. – 59 (1979) pp. 49-107

[21] Febvre, Lucien; Martin, Henri-Jean. [*L'apparition du livre*. Español] *La aparición del libro. C*on el concurso de Anne Basanoff ... [et al.] ; traducción al español por el Dr. Agustín Millares Carlo. 1ª ed. en español. México : Unión Tipográfica Editorial Hispano Americana, [1962]. XXIV, 439 p., XXXII p. de lám., 2 p. de mapas. (*La evolución de la humanidad ; t. 70. Sección segunda, Orígenes del cristianismo y Edad Media*). *L'apparition du libre* se publico originalmente en 1958.

[22] *Nueva introducción a la bibliografía material* / Philip Gaskell ; [traducción, Consuelo Fernández Cuartas y Faustino Álvarez Álvarez]. - 1ª ed. - Gijón : Trea, 1999. - XXXI, 540 p. - (Biblioteconomía y administración cultural ; 23)

[23] La más significativa para esta comunicación es Agenjo, Xavier y Hernández, Francisca. *Del Catálogo Colectivo a la Biblioteca Virtual : La Biblioteca Virtual del Patrimonio Bibliográfico*. En: *I Jornadas sobre Patrimonio Bibliográfico en Castilla-La Mancha* : actas : 12, 13 y 14 de noviembre, Alcázar de Toledo. Toledo, 2003, pp. 391-418

Tipobibliografía española[24], lo que ponía de manifiesto que la catalogación cooperativa permitía, gracias a la validación sistemática que supone la incorporación a la descripción bibliográfica de nuevas localizaciones, el llevar a cabo un cotejo –pero un cotejo virtual- que iba enriqueciendo y precisando cada vez más la descripción específica de cada ejemplar[25].

El lector se verá tentado de pensar que todo ello es posible con impresos del siglo XV e incluso del siglo XVI, pues su número es relativamente pequeño y, sobre todo, su extensión rara vez es comparable al de las publicaciones periódicas. En efecto, es así, pero lo que supone la radical modernización del método de la descripción de los ejemplares de las publicaciones periódicas es el proceso que podríamos llamar de creación de los objetos digitales que tiene lugar mediante el escaneo, primero, de todas y cada una de las páginas; la validación, después, de todas y cada una de las páginas y mediante la generación de estructuras de intercambio de información, fundamentalmente METS. Si además se lleva a cabo un proceso de reconocimiento óptico de caracteres, utilizando, y así debería hacerse siempre en términos generales, un estándar como ALTO (*Analyzed Layout and Text Object*)[26], el análisis material va mucho más allá de lo que cualquier tipobibliógrafo hubiera podido soñar. Por medio de ALTO se establecen las coordenadas de cada una de las palabras, naturalmente referidas a la disposición tipográfica de cada una de las páginas, con lo que habremos obtenido a lo largo de todo el proceso, una información detallada a nivel de página. Es más, de cada una de las palabras (y/o caracteres) que conforman cada página. Evidentemente, ALTO permite la identificación de artículos, anuncios, índices, fotografías, etc., pues especifica la segmentación de estas partes de una página y su referencia a la localización en la disposición del texto.

Modelo de datos

Evidentemente, y así ha sido requerido en muchos de los concursos públicos que se han llevado a cabo en España en los últimos años, el modelo obvio es el Formato MARC 21 para registros de fondos y localizaciones, progresivamente enriquecido a lo largo de esta década y al que se han incorporado ya, al menos parcialmente, una buena parte de las RDA, aunque habrá que esperar a la actualización número 13 y sucesivas para ver hasta qué punto se recoge la información esencial del nuevo modelo. Como es sabido, los campos 863 a 865 están dedicados a la enumeración y cronología, los campos 866 a 868 a las menciones textuales de fondos y los campos 876 y 878 a los campos de información sobre la unidad física.

Si analizamos los campos 866/868 veremos que resultan especialmente apropiados para recoger de forma automática la información acerca del estado de la cuestión de cada uno de los números, en los se señala de forma análoga a los catálogos de librerías de viejo las especiales características detectadas en un número, desde la propia falta de una página hasta

---

[24] Martín Abad, Julián. *La imprenta en Alcalá de Henares (1502-1600)* ; introducción a la "Tipobibliografía española", José Simón Díaz. - Madrid : Arco Libros, 1991. - 3 v. - (*Tipobibliografía española).* Hay un buen resumen de este proyecto, que sería necesario actualizar, en Reyes, Fermín de los. *El proyecto Tipobibliografía española.* En: *Boletín de la Biblioteca de Menéndez Pelayo*. - Santander : Sociedad Menéndez Pelayo, 1919-. - ISSN 0006-1646. - LXXVIII (EneroDiciembre 2002) pp. 171-197.

[25] En ese sentido, constituye un verdadero modelo la extraordinaria introducción de Julián Martín Abad a la reciente edición del *Catálogo bibliográfico de la colección de incunables de la Biblioteca Nacional de España.* Elaborado por Julián Martín Abad.- Madrid : Biblioteca Nacional de España, 2010.- 2 v., dónde se estudian las características de cada ejemplar de un modo admirable.

[26] http://www.loc.gov/standards/alto/

el mal estado en el que este se encuentra, la fragmentación, roturas, cortes, manchas de humedad y, lo que es más frecuente, aunque no siempre se distingue, la presencia de todo tipo de bibliófagos. A partir de estos campos pueden construirse las menciones textuales de fondos (campos 866/868)

Sin embargo, hay que mencionar que las menciones textuales de fondos cuando se refieren a un ejemplar de prensa suelen ser muy farragosas y difícilmente inteligibles por los usuarios. Sobre todo, en el caso de cabeceras que se extienden a lo largo de varios años,; la mención de fondos puede ser muy detallada, pero muy poco usable. La norma ISO 10324:1997 que se corresponde con la norma Z39.71, tiene no sólo un valor descriptivo, sino también vinculado a las transacciones de circulación entre distintos sistemas de información bibliográfica y bibliotecas, pero en aras de la usabilidad debe ser complementada con una presentación de la relación de números en forma de listado cronológico o calendario.

Es evidente, aunque no suele señalarse, que ambas normas poseen también un importante valor de orden económico, pues aunque las colecciones de publicaciones periódicas o incluso los números sueltos o determinados periodos no tienen un comercio de anticuariado tan importante como el característico de las monografías, en absoluto puede decirse que carezcan de él. Es más, tiende a crecer y cada vez es más frecuente ver ofertas de publicaciones periódicas, tanto para coleccionistas de determinadas materias como para las propias bibliotecas que, bien por razones de especialización, bien por razones de control bibliográfico retrospectivo, incluyen la búsqueda de ejemplares que permitan conformar y completar una colección, en realidad, un ejemplar. De ahí el enorme sentido de las bibliotecas virtuales que pueden transformar, como ya se dijo antes, un conjunto de catálogos en un catálogo único virtual, agregando digitalmente ejemplares o ítems depositados en diversas bibliotecas, pero que conforman un único ejemplar virtual, aunque este pueda estar formado por diversos ítems de una misma manifestación.

Es obligado hacer aquí un comentario sobre la definición de ítem que proporcionan los FRBR que afecta particularmente a la prensa. En la primera traducción al español[27] se utilizó el término ítem [con tilde], admitido por el Diccionario de la Real Academia Española. Para la segunda traducción [en prensa], con adiciones, se ha utilizado *ejemplar* en lugar de *ítem*, para mantener la unidad con el resto de las traducciones de requisitos funcionales [FRAD y FRSAD], aunque uno de los coautores de la traducción y también coautor de esta comunicación no esté completamente de acuerdo. En efecto, como ocurre en todos los ejemplares compuestos de más de una parte, son sustanciales las diferencias entre el ejemplar y las unidades físicas disponibles. De hecho, la relación Todo/Parte puede establecerse entre manifestaciones y los propios FRBR ponen como ejemplo de este tipo de relación la de un volumen con una manifestación multivolumen. Quiere esto decir que lo que habitualmente entendemos como ejemplar de una publicación periódica, los números o partes que las componen, pueden ser descritas como manifestaciones y, por tanto, el ítem sería la unidad física correspondiente a una de las partes de una cabecera, o de una publicación en varios volúmenes.

[27] *Requisitos funcionales de los registros bibliográficos : informe final*. Grupo de estudio de la IFLA sobre los Requisitos funcionales de los registros bibliográficos ; traducción de Xavier Agenjo y María Luisa Martínez-Conde. — [Madrid] : Ministerio de Cultura, 2004. — ISBN 84-8181- 213-7. http://travesia.mcu.es/portalnb/jspui/bitstream/10421/422/1/frbr.pdf (En *Travesía*, Ministerio de Cultura de España). http://www.ifla.org/files/cataloguing/frbr/frbr-es.pdf (Servidor de IFLA)

Desde el punto de vista del modelo FRBR (Work-Expression-Manifestation-Item, conocido habitualmente por WEMI), no cabe duda que la diferenciación entre expresión y manifestación ha resuelto o puede llegar a resolver algunos importantes problemas catalográficos. Sin embargo, la definición de ejemplar (que consideramos que se corresponde con la de manifestación) o de ítem (que consideramos que se corresponde con la unidad física) ya no es tan distinta, lo cual puede notarse mucho más en la prensa diaria. Se ha anunciado para la próxima actualización del formato MARC 21 la forma de codificar, y distinguir, Obras/Expresiones/Manifestaciones/Ítems (ejemplar en la segunda edición de FRBR en español), lo que en el caso de la prensa resultará de gran ayuda.

Por otro lado, la información propia del patrón de publicación (853/855), y de enumeración y cronología (863/865) de un determinado título formarían parte de la descripción de Expresión/Manifestación, mientras que en el formato MARC 21 son campos hasta ahora típicamente de *holdings*. Estos campos están también disponibles en los registros bibliográficos, pero no es aconsejable utilizarlos cuando se están describiendo ítems pertenecientes a distintas bibliotecas, es decir, cuando se está elaborando un catálogo colectivo o una biblioteca virtual. Es decir, consideramos que en el caso de un ejemplar virtual, que puede ser el más completo existente, no hace referencia a un ejemplar físico, sino más bien a las partes de una manifestación. El ejemplar virtual al ser el más completo es el que más se acerca a la manifestación. Por su lado, los ítems, las unidades físicas, están ubicados en diferentes bibliotecas y tienen distintas peculiaridades físicas o procedencias cuyo registro es necesario.

Este es un aspecto importante que las RDA afrontan, pero que de momento no tiene una resolución definitiva, puesto que es perfectamente concebible la creación de un ejemplar ideal o virtual (manifestación) formado por un conjunto de ítems muy diferentes. La definición de manifestación o de ítem deja margen a que la mención textual de fondos pueda realizarse según mejor se acomode, lo cual puede verse por los ejemplos[28] que siguen, tomados de la Biblioteca Nacional de España y de la Library of Congress: *Historia de Menéndez Pidal* o el *Summa Artis* o en el ámbito anglosajón con *A Study of History* de Toynbee. Como puede verse por estos casos se está construyendo ejemplares formados por ítems pertenecientes a diferentes manifestaciones. Mucho más se da en la descripción de la prensa diaria en la que, con frecuencia, pueden aparecer hasta 3 o más ediciones distintas un mismo día, si se ha producido, por ejemplo, algún acontecimiento de especial relevancia. No debe pensarse nunca que este es un detalle menor, sino todo lo contrario; es seguro que el lector, aún más el investigador se encontrará particularmente interesado por esas tres o cuatro

[28] *Historia de España Menéndez Pidal* [Texto impreso] / dirigida por José María Jover Zamora .- Madrid : Espasa-Calpe, 1989-<2007> .- v. <1, (3 v.), 2 (2 v.), 3 (2 v.), 4-6, 7 (2 v.), 8 (3 v.), 9, 10 (2 v.), 11-12, 13 (2 v.), 14-16, 17 (2 v.), 18-21, 22 (2 v.), 23-25, 26 (2 v.), 27-28, 29 (2 v.) 30, 31 (2 v.), 32 (2 v.), 33-34, 35 (2 v.), 36 (2 v.), 37, 38 (2 v.), 39 (2 v.), 40, 41, pte. 1, 42> : il, col. y n. ; 28 cm

Pijoán, José (1881-1963) *Summa Artis* [Texto impreso] : *historia general del arte* / José Pijoán .- Madrid : Espasa-Calpe, 1931-2001 .- 51 v. : il. col. y n. ; 28 cm
Nota general: Volúmenes correspondientes a distintas ediciones

Toynbee, Arnold, (1889-1975) *A study of history*. London, New York, Oxford University Press [1948]-61. 12 v. maps (part fold., part col.) tables. 23 cm. (v. 11:26 cm.) Partial contents: v. 11. Historical atlas and gazetteer, by A. J. Toynbee and E. D. Meyers.--v. 12. Reconsiderations.
Notes: Vols. 1-3: 2d ed.; v. 4-6: 1st ed.
"Issued under the auspices of the Royal Institute of International Affairs." Includes bibliographies.
Obsérvese que de los volúmenes 7 al 12, la Library of Congress no dice nada.

tiradas que tuvo un título un día determinado, puesto que eso ocurre cuando se da justamente una situación de relevancia especial.

Pongamos otro ejemplo imaginario de combinaciones digitales de manifestaciones o ítems para formar un ejemplar virtual: una publicación periódica formada por 100 números, 20 de los cuales están digitalizados en TIFF, otros 20 en JPEG, otros 20 en TIFF en escala de grises, otros 20 en JPEG2000 y, por último, un quinto más en JPEG con 16,7 millones de colores. El grado de resolución puede así mismo variar y puede ser que las páginas hayan experimentado un proceso de OCR, unas en formato ALTO y otras de forma menos estandarizada. La experiencia real que poseemos, profesionalmente y como proveedores de servicios, nos lo recuerda constantemente. Véase, por ejemplo, los casos de *El Diario Mercantil de Cádiz*[29] o *El Guadalete*[30] de Jerez de la Frontera, formados por diferentes procesos de digitalización realizados con los fondos de distintas bibliotecas.

Volviendo a la cuestión del formato MARC 21 de fondos, consideramos irrelevante analizar aquí el formato IBERMARC[31] para fondos y localizaciones pues está basado en la primera versión del *MARC 21 Format for Holdings records*[32] aparecida en el año 2000 y evidentemente no sirve para el propósito de este trabajo. Por desgracia, aún son abundantes las aplicaciones informáticas que en España, y aún en Hispanoamérica, siguen utilizando el formato IBERMARC y que prescinden del de fondos, utilizando simplemente un objeto para la circulación, lo que impide cumplimentar los diferentes campos y subcampos que se están analizando aquí y que resultan esenciales por su valor descriptivo, desde el patrón de publicación a las condiciones y estado del ejemplar.

Probablemente, en estos momentos no exista ninguna estructura de metadatos lo bastante amplia como para poder recoger todos los datos que pueden obtenerse de los distintos procesos que dan lugar a la creación de un objeto digital, tal vez con la excepción del diccionario de datos PREMIS, donde efectivamente existe una enorme cantidad de elementos perfectamente cualificados para recoger en ellos las incidencias de las manifestaciones o ítems digitales. La combinación de METS y PREMIS permite que la descripción de incidencias a nivel de página, tanto en las unidades físicas originales como en las digitales, pueda ser tan detallada como se desee. Indudablemente, una posibilidad es combinar la utilización de descripciones MODS para la cabecera, los números o las páginas, como recomienda el *METS Profile for Historical Newspapers*[33] con los propios elementos *PREMIS*.

[El 23 de mayo de 2011, ya redactadas estas líneas, Sally McCallum ha informado en la lista MARC de la existencia de un nuevo proyecto que se denomina *Transforming our Bibliographic Framework*[34], así como de la creación del sitio Web *Bibliographic Framework Transition Initiative*[35] para ir dando cuenta de los progresos de esta iniciativa.]

---

[29] http://prensahistorica.mcu.es/es/consulta/registro.cmd?id=3625

[30] http://prensahistorica.mcu.es/es/consulta/registro.cmd?id=3633

[31] *Formato IBERMARC para registros bibliográficos*. Madrid : Biblioteca Nacional, 2004 [http://goo.gl/HZSbd]

[32] http://www.loc.gov/marc/holdings/echdhome.html

[33] *METS Profile for Historical Newspapers* http://www.loc.gov/standards/mets/test/ndnp/00000010.html

[34] http://www.loc.gov/marc/transition/news/framework-051311.html

[35] http://www.loc.gov/marc/transition/index.html

Además de la necesidad de una estructura de datos más amplia que recoja el análisis FRBR o RDA y permita describir una publicación en profundidad hasta el nivel de página, es necesario también que los proceso inherentes a la digitalización y al OCR (Optical Character Recognition) resulten significativos bibliográficamente hablando. Es decir, que se produzca un proceso de *retroalimentación* de los datos específicos obtenidos los procesos de digitalización, validación, OCR e indexación a las estructuras bibliográficas existentes. Sin duda, se trataría de un proceso especialmente laborioso, aunque imprescindible, si se llevara a cabo de forma completamente manual y por ello y, sobre todo, dado el gran volumen de digitalización de los proyectos emprendidos por la empresa DIGIBÍS se ha desarrollado un programa denominado DIGIPRO, que combinada con DIGIBIB[36], realiza automáticamente la translación de datos de procedimiento al objeto digital, lo que tiene, claro es, una enorme importancia también para la preservación a largo plazo de la manifestación digital[37].

Reconocimiento óptico de caracteres

La disponibilidad de imágenes digitalizadas de una publicación periódica abre un enorme abanico de posibilidades para la consulta y la difusión de los materiales bibliográficos. Pero, además, permite realizar nuevos análisis, automatizados en mayor o menor grado, que amplíen la información disponible sobre una obra. El más común de estos análisis es el proceso de Reconocimiento Automático de Caracteres (OCR). Ya desde 1960 existe software que permite realizar este proceso, pero, ha sido en los últimos años cuando la calidad del reconocimiento[38], es decir, el número de caracteres que se reconocen correctamente, ha aumentado hasta niveles que permiten utilizar los resultados sin necesidad de un costoso procedimiento de corrección manual posterior[39]. Además, los motores de OCR recientes identifican no sólo el texto sino su disposición en párrafos y columnas y sus características tipográficas, y disponen de funciones que permiten recuperar la posición de cada carácter y término reconocido sobre la imagen digitalizada.

Esta información permite incorporar nuevas funcionalidades a los sistemas de gestión y consulta bibliográfica. Más allá de la ya común búsqueda a texto libre (incluyendo la posibilidad de presentar los términos resaltados en su posición sobre la imagen digitalizada), es posible marcar un fragmento de la imagen y exportar el texto identificado incluido sólo en esa zona, traducirlo o corregirlo de forma cooperativa vía web[40]. Y es posible, también, utilizar la información tipográfica y de composición del texto para extraer automáticamente

---

[36] DIGIBIB es la aplicación informática integrada desarrollada por DIGIBÍS para la gestión de bibliotecas digitales [http://goo.gl/FgWTu]

[37] Ob. cit., nota 16

[38] http://www.impact-project.eu/

[39] Simon Tanner, Trevor Muñoz, Pich Henry Ros, "*Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive*". En: "D-Lib Magazine", ISSN 1082-9873, Volumen 15, Número 7/8, Julio/Agosto 2009. Versión online en: http://www.dlib.org/dlib/july09/munoz/07munoz.html

[40] Rose Holley, "*Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*", National Library of Australia Staff Papers, 2009. Versión online en: http://www.nationaltreasures.nla.gov.au/openpublish/index.php/nlasp/article/view/1406/1688.

tablas de contenido, índices o la numeración original de las páginas, y en el ámbito específico de las publicaciones periódicas, identificar las partes componentes de una publicación: artículos con sus títulos, fotografías, anuncios, etc. Normalmente, estos datos extraídos automáticamente necesitarán una revisión manual posterior, pero puede plantearse de forma cooperativa gracias a la ubicuidad de las redes de comunicación. Otra posibilidad de gran interés es la generación de libros electrónicos con el texto contenido en una obra en formato ePub[41] que permite redimensionar el texto al tamaño del lector empleado.

En el desarrollo de herramientas específicas para la gestión de los procesos de OCR pronto se hizo patente la conveniencia de utilizar una estructura de datos específica, y a ser posible normalizada, que actuara como contenedor o representación de los objetos digitales resultantes de los procesos de digitalización y de OCR. Trabajar con formatos normalizados simplifica, por ejemplo, la carga en diferentes plataformas de consulta o posibilita el desarrollo de funciones posteriores de análisis de grandes volúmenes de textos, por ejemplo, el análisis de la evolución temporal de las frecuencias de uso de diferentes términos. En la actualidad hay dos estándares de amplia difusión, ALTO (*Analysed Layout and Text Objects*), mantenido por la Library of Congress, y *hOCR*, que almacena la información detallada devuelta por el OCR dentro de un fichero XHTML[42], lo que permite consultar el texto extraído sin formato en cualquier navegador HTML y acceder a información adicional si se precisa.

La obtención a partir de las imágenes de nuevos juegos de ficheros con el texto completo de la obra, y posibles derivados adicionales (ePub y otros formatos alternativos para lectores, PDF, etc.), presenta una dificultad a la hora de intercambiar el recurso digital completo (por ejemplo, al enviarlo a un repositorio centralizado). Es necesario disponer de un mecanismo que nos permita integrar tanto los metadatos de la obra como los de los ficheros generados, información de derechos y/o de preservación, etc.

El estándar METS (*Metadata Encoding and Transmission Standard*) soluciona todos estos problemas. Lo que para algunos usos supone una debilidad (la enorme flexibilidad de la norma que dificulta el desarrollo de aplicaciones genéricas que trabajen con cualquier fichero METS, y que ha conducido a la definición de múltiples perfiles diferentes) es, en este caso, una fortaleza, pues permite definir un perfil *ex profeso* para la funcionalidad requerida, incluyendo todos los tipos de ficheros y relaciones que se deseen. Sin embargo, para que todo ello sea verdaderamente efectivo debe de pivotar en torno a la descripción de las unidades de contenido básicas, que coinciden con las unidades físicas, es decir con un registro específico para los números de una publicación periódica que permita incluir tanto los datos bibliográficos como datos de ejemplar. O en términos FRBR, tanto los atributos de las manifestaciones como los de los ítems.

Búsqueda a texto completo e interfaz Web

Una de las mayores ventajas que ofrece el sistema de gestión bibliográfico DIGIBIB a partir de la versión 6.0, utilizada en las bibliotecas virtuales y digitales citadas anteriormente, es la posibilidad de realizar búsquedas a texto completo en el contenido de los números de prensa.

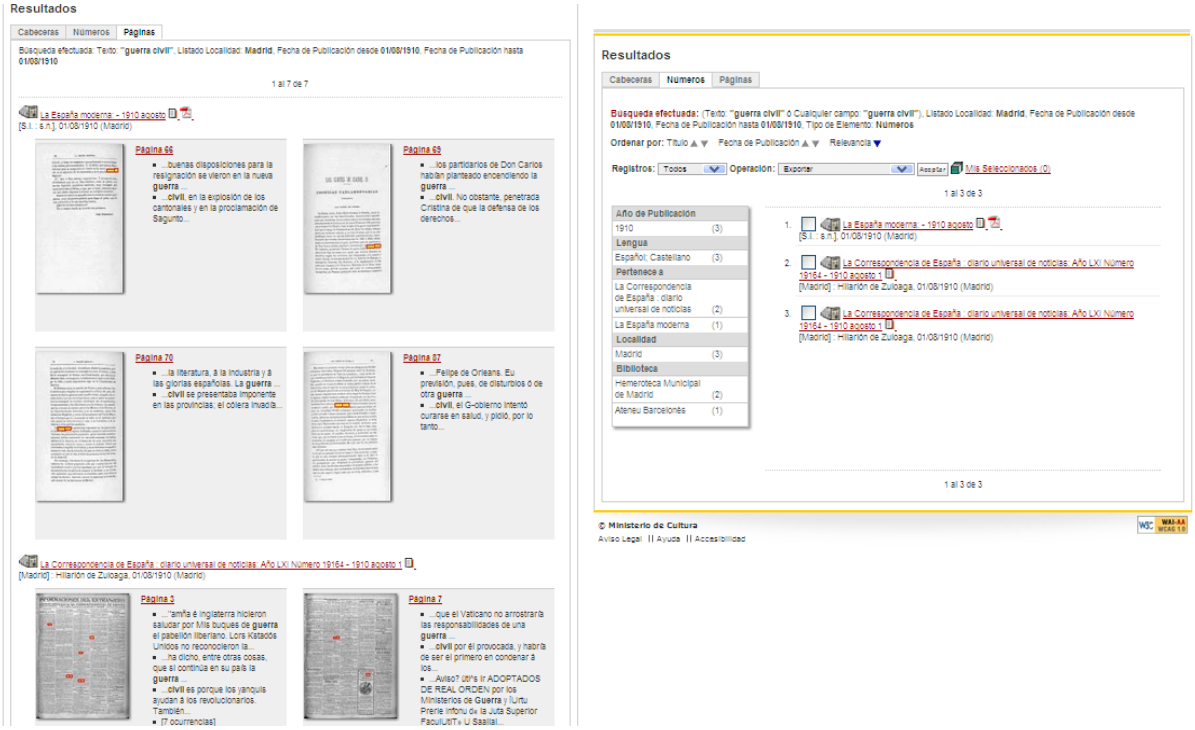
[41] http://idpf.org/

[42] T. Breuel, "*The hOCR Microformat for OCR Workflow and Results*", icdar, vol. 2, pp.1063-1067, Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2, 2007

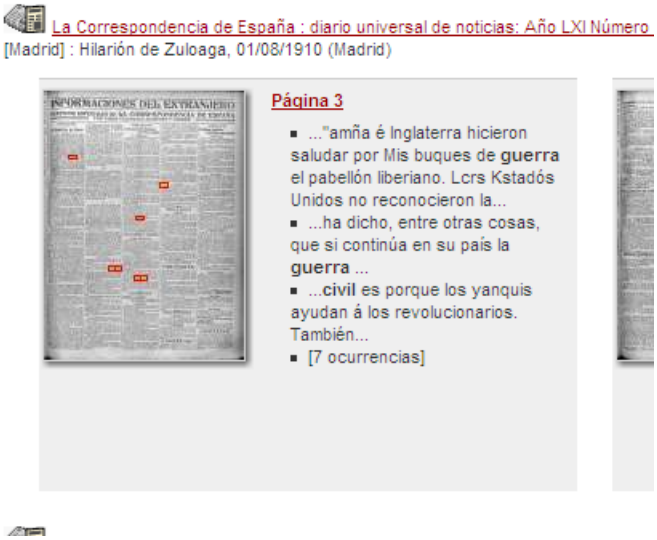
Para llegar a esta solución ha sido necesario introducir una *nueva* estructura de datos para los números de prensa, información que se gestionaba anteriormente por medio de datos externos al sistema. Como se ha dicho, el problema que se presentaba radicaba en que las imágenes digitales y sus textos pertenecían a los números de publicaciones, y éstos no se consideraban unidades bibliográficas, y por tanto no se podían utilizar a la hora de la búsqueda y/o la presentación y ordenación de resultados. Sin embargo, sí se representaban los números de una publicación por medio de "registros virtuales" con su propia entidad, se creaban unidades de organización e indexación, permitiendo búsquedas de números tanto por sus campos bibliográficos, como por los textos asociados. Se ha mencionado que MODS sí ofrece la posibilidad de estructurar jerárquicamente las descripciones de una cabecera, de sus números, e incluso de las páginas o de las partes intelectuales constitutivas, pero DIGIBIB utiliza la estructura MARC 21, que ofrece mucha mayor interoperabilidad y granularidad, como formato común y como origen de conversiones a otros esquemas de metadatos (p.e., a Dublin Core y MODS).

Junto con la indexación de la información bibliográfica de las cabeceras, de los números y de los artículos (si se hubieran identificado éstos), DIGIBIB también indexa cada una de las páginas de los números como unidad de indexación independiente, integrando los metadatos del número y de la obra a la que pertenece, junto con el texto de cada una de ellas. Bien es cierto que de esta forma aumenta en buena medida el tamaño de los textos indexados, pero, a su vez, como muy interesante contrapartida, permite realizar búsquedas tanto por el texto como por los metadatos de las obras, mostrando al usuario en los resultados sólo las páginas que incluyen los términos buscados, y permitiendo ordenar y filtrar los mismos por múltiples criterios. Ahora bien, aunque el usuario normalmente prefiere que en la lista de resultados aparezcan páginas individuales, estas páginas deben aparecer agrupadas por números para permitir una mejor contextualización de las mismas. Y, alternativamente, en ocasiones el usuario lo que querrá es encontrar los números de la publicación sin necesidad de acceder a las páginas.

Por todo ello, y tras evaluar las pruebas de usabilidad, la decisión final adoptada consistió en presentar una interfaz de usuario, en la que los resultados se separan por diferentes solapas para Cabeceras, Números y Páginas, agrupadas estas últimas por la unidad bibliográfica de la cabecera, los números de las publicaciones, y permitiendo al usuario, de esta forma, ver los resultados de una búsqueda de forma independiente por los tres tipos de elementos mencionados. Además, los resultados presentados en cada solapa podrán ordenarse y filtrarse, facilitando al usuario la obtención de resultados más precisos.



En el caso concreto de la solapa de Páginas, los usuarios se han mostrado particularmente satisfechos con la opción de mostrar a la vez pequeños fragmentos de texto con los términos buscados resaltados, y la posición de esos textos sobre la miniatura de la página, pues así pueden comprobar simultáneamente si la semántica del término en el contexto es la que ellos buscaban, y la importancia del mismo (si aparece al principio de un artículo o en un titular, o en medio de un bloque de texto).



En cuanto a los diferentes tipos de búsquedas permitidas por la interfaz de usuario, es importante destacar la posibilidad de buscar no sólo en el texto completo de las obras, sino también en los campos de los registros bibliográficos. Así, un usuario podría limitar una búsqueda no sólo por un cierto término de texto, sino también, por ejemplo, por un rango de fechas de publicación o por una localidad geográfica, lo que pone de manifiesto el gran potencial que este tipo de búsquedas ofrece a los usuarios.

A todo lo descrito hasta ahora, habría que añadir la posibilidad de consultar en la prensa histórica simultáneamente la búsqueda a texto libre con una presentación en forma de calendario, que permite explorar rápidamente los contenidos del repositorio en el tiempo, una dimensión que en el caso de prensa es fundamental, especialmente en el caso de la prensa de carácter histórico, en el que muy a menudo el primer criterio de búsqueda es un acontecimiento localizado de forma muy precisa en el tiempo (y a veces también en el espacio).

Procesos técnicos para la creación de un objeto digital

La integración de los procesos de digitalización, OCR, control de calidad, descripción, e indexación junto con las funcionalidades de búsqueda y recuperación de una biblioteca virtual, conforman un flujo de trabajo (o *workflow*, para utilizar innecesariamente el término inglés, tan extendido) que debe ser *proactivo* y no meramente *descriptivo*. No se trata únicamente de obtener información precisa del flujo de trabajo de los procesos de creación de un objeto digital, desde su escaneo hasta la validación y asignación de metadatos, sino de lograr que exista una retroalimentación de todo el proceso que se refleje tanto en los datos como en los metadatos del referido objeto digital.

De forma resumida, los procesos técnicos que hemos señalado, organizados conforme a un flujo de trabajo determinado a lo largo del cual se proporcionan datos significativos y suficientes para alimentar los sucesivos subprocesos, finalizan en la creación de un objeto digital, que no es sino la suma estructurada de datos, metadatos y ficheros que representan, en definitiva, mucho más que el objeto original. Sin entrar en cuestiones sobre las características técnicas de la digitalización iremos desgranando de las sucesivas fases y procesos los tratamientos que se realizan y los datos que se obtienen y que son reutilizados y que realimentan los consiguientes procesos.

Procesos de control de calidad

Desde el inicio de la digitalización los ficheros que componen la imagen de un original son organizados en un sistema de ficheros al que se asigna un nombre que refleja la cronología de un determinado título. Lamentablemente, algunos proyectos de digitalización se quedan en este paso, siendo la estructura de ficheros la única referencia al original. Nunca se insistirá demasiado en que este procedimiento, tan extendido, es completamente insuficiente para realizar las funcionalidades propias de una biblioteca virtual y que a menudo supone un escollo, o un gasto adicional posterior, para superar esas limitaciones.

En el proceso de validación y control de calidad se revisan cada una de las imágenes que se han obtenido en la digitalización. En principio el objetivo de este proceso es asegurar la calidad técnica de las imágenes, su legibilidad y su correcta ordenación. Sin embargo, en este proceso rápidamente se detectan problemas típicos de la prensa que podemos incluir dentro de los datos descriptivos de la *manifestación* puesto que se trata de errores de numeración, de cronología y/o de paginación. Del mismo modo, en muchas ocasiones se detecta que el ejemplar encuadernado que se conserva está compuesto por números que pertenecen a diferentes manifestaciones de un mismo título como ocurre cuando se mezclan diferentes ediciones de un mismo día o cuando se han encuadernado números pertenecientes a ediciones relacionadas con diferentes localidades.

Rápidamente, se observó que se debían subsanar este conjunto de problemas puesto que, de no ser tratados adecuadamente, o de ser tratados únicamente como una organización de ficheros, daba lugar a aparentes repeticiones de números y a la imposibilidad de ofrecer una lista de números correctamente relacionados por su cronología y numeración. Es importante recalcar aquí la influencia del principio ya señalado de formación cooperativa de una biblioteca virtual y de, lo que no es nada desdeñable, una biblioteca virtual que se forma a lo largo de sucesivos años. Era obligado tener en cuenta la existencia de ediciones, puesto que podían aparecer en años sucesivos o bien podían aparecer en otras bibliotecas. Por otro lado, es obvio que la calidad y detalle de una descripción depende en gran medida de lo completo que sea un determinado ejemplar y que virtualmente se podía obtener la manifestación más completa, el ejemplar, virtual, más completo.

Como es lógico, la reorganización de la estructura de ficheros para acomodarla a la manifestación real debía corresponderse con una descripción clara del modelo de publicación, detallando al máximo los campos de numeración y cronología de modo que pudiera preverse la posterior aparición de números que completaran el ejemplar virtual.

Igualmente, en el proceso de validación y control de calidad se consignaban otros elementos descriptivos que facilitaran la descripción bibliográfica de un título y el momento de la revisión página a página era el más adecuado para señalar y marcar todos aquellos cambios que debían ser consignados en el registro bibliográfico. De esta forma, se marcaron por medio de notas asociadas a un determinado número o página cambios de título, subtítulo, periodicidad, editor o impresor, así como los errores de numeración y cronología. Del mismo modo, se señalaba la presencia de índices, suplementos o monografías incluidas en el texto.

Puede observarse, por tanto, que del proceso de control de calidad se obtienen datos propios de la manifestación, pero también datos propios del ejemplar, puesto que se anota la presencia de daños en las páginas, especialmente todos aquellos que dificultan o impiden la legibilidad total o parcial del texto. Desde luego, la falta de números es fácilmente deducible, aunque no pueda establecerse con seguridad si se trata de un número que falta o de un número que no llegó a publicarse nunca.

Debe tenerse en cuenta que la organización de ficheros, para que sea un reflejo exacto de la manifestación (que no del ejemplar), se realiza sobre la copia digital máster, de la que se obtendrán posteriormente otras versiones digitales para diferentes usos. Es decir, en el proceso de creación del ejemplar virtual no sólo se reconstruye la manifestación del original, sino que se están construyendo al mismo tiempo otras manifestaciones digitales, una de las cuales lleva asociado el texto obtenido por medio del proceso de OCR.

Además, la necesidad de vincular los ficheros obtenidos de los procesos de digitalización o de OCR a las unidades correspondientes para construir las manifestaciones digitales y la construcción de las funcionalidades de búsqueda de las bibliotecas virtuales de prensa requerían de un tipo de registro que aunara todas las características de las partes de una publicación periódica. Este registro formado por una parte de datos bibliográficos comunes a la manifestación general y otros específicos de la manifestación de la parte, así como de los datos propios de la unidad física llevaron a la creación de los denominados por nosotros registros de número.

Registros de número

La catalogación detallada de prensa histórica requiere la aportación de un gran esfuerzo por parte de los catalogadores, debido a su volumen y presenta algunas dificultades a la hora de encajar adecuadamente la información propia de cada uno de los números de un recurso continuo en un registro MARC 21. La descripción bibliográfica de los números de un periódico puede llegar a ser una tarea imposible (obsérvese que en DIGIBIS se han tratado en los últimos años 1.292.195 números) realizada manualmente. De hecho, este ha sido tradicionalmente uno de los cuellos de botella en el control bibliográfico de las colecciones de prensa e incluso de publicaciones periódicas.

Además, ya se ha señalado la necesidad de contar con una descripción a nivel de número para representar correctamente las distintas manifestaciones de una cabecera y de sus partes componentes, un registro de este tipo era absolutamente necesario para construir sobre él las funcionalidades de búsqueda y recuperación de información, incluyendo en ella la búsqueda a texto completo realizada sobre el contenido obtenido de los procesos de OCR. De hecho, hasta la versión 6.0 del ILS DIGIBIB, la solución adoptada era la de catalogar cada una de las cabeceras de prensa con un registro bibliográfico en MARC21, mientras que la gestión de los números se realizaba mediante un modelo propio, almacenado en una base de datos relacional.

Como puede verse a partir del análisis de los atributos de las entidades WEMI tal y como se van consignando en el proceso de control de calidad, los registros MARC 21 distribuyen esa combinación de atributos, que pueden pertenecer a las entidades Obra-Expresión-Manifestación-Ítem, entre registros de autoridad de título uniforme, registros bibliográficos y registros de fondos y localizaciones:

- Patrón de publicación (Expresión)
- Frecuencia (Expresión)
- Numeración y cronología (Manifestación)
- Cambios de título (Obra - Expresión)
- Cambios de subtítulo (Manifestación)
- Fusiones (Obra – Expresión)
- Separaciones (Obra-Expresión)
- Continuaciones (Obra-Expresión)
- Presencia de índices y suplementos (Manifestación)
- Cambios de editores (Manifestación)
- Cambios de impresores (Manifestación)
- Cambios de tamaño y número de páginas (Manifestación)
- Estado de conservación (Ítem)
- Signatura (Ítem)

Era posible, por supuesto crear para cada número de una publicación periódica un registro de ejemplar para cada una de las unidades físicas. En un registro de este tipo hubiera sido posible mantener el nombre de la unidad bibliográfica en el campo 844, pero no se hubieran podido mantener otros datos de la descripción bibliográfica absolutamente necesarios para las funcionalidades de búsqueda y recuperación.

Por tanto, los números de prensa, a través de esos procesos complementarios o mediante la gestión realizada en DIGIBIB, pueden tener su propia información descriptiva, como su fecha de publicación, notas en las que se describe su estado de conservación y si se identifica lo que

falta, e incluso se les asocia tanto las imágenes digitalizadas, como los ficheros resultantes de los procesos de OCR, se obtienen descripciones de que combinan atributos propios de la manifestación con atributos propios del ítem, todo ello codificado en un registro bibliográfico en formato MARC 21. Teniendo en cuenta todo ello, es relativamente sencillo generar de forma automática un registro bibliográfico "virtual" para cada número de publicación, haciendo un "mapeo" de toda esa información, a los campos correspondientes del formato MARC21, y así se procedió a llevar a cabo en DIGIBIB desde su versión 6.0.

La consideración de cada número como un registro bibliográfico, con su propia entidad, y su descripción catalográfica en el formato MARC21, tal y como se hace en la actualidad para los proyectos de digitalización de prensa que realiza DIGIBÍS, permite la indexación de los metadatos propios de los números, e inclusive de los textos resultado de los procesos de OCR, de tal forma que se puede buscar y mostrar la información de un número concreto y de sus objetos digitales, tanto delimitando por valores en sus metadatos, por ejemplo, la fecha de publicación, como delimitando las búsquedas a texto completo, resultando estas funcionalidades especialmente útiles, en aquellos casos en que el recurso continuo sea muy extenso, como ocurre con frecuencia en la prensa.

Información heredada de la publicación

Parte de los atributos o campos de los registros de números se heredan de la descripción bibliográfica de la publicación en su totalidad. Un ejemplo de campo "heredado" es el campo 752, en el que se indica la cobertura geográfica de la publicación expresada jerárquicamente y que de forma lógica es compartida por todos sus números. Este hecho ofrece la posibilidad de realizar búsquedas de números incluyendo como filtro una localidad, lo cual es fundamental en prensa. Del mismo modo, se han tratado los campos 6XX. Es cierto que al realizar la "copia" de todas y cada una de las materias relacionadas con una publicación a todos sus números se crearán relaciones que en realidad serían inexistentes para un número en concreto, pero a pesar de ello, cabe pensar que es mayor la ventaja de incluir dichas relaciones entre materias y números, máxime cuando las entradas de materia que se utilizan habitualmente en prensa no pueden ser muy específicas, lo cual reduce el margen de imprecisión, pero permite realizar búsquedas de números delimitando por una o varias materias. También se hereda de la publicación, mediante copia, los campos y posiciones (p.e., 041) que describen el o los idiomas del contenido a través de los cuales se pueden filtrar resultados de búsqueda.

Otro ejemplo de campo "heredado" de la información de la publicación, es el campo 773 de un número, que relaciona a éste con el bibliográfico de la publicación a la que pertenece. En este caso, el campo 773 del número no se "copia" a partir de los campos 773 que declara la publicación, sino que se cumplimenta con información declarada en otros campos de la misma: el subcampo $d del campo 773 se forma por la concatenación de los subcampos del 260 de la publicación, y los subcampos $t, $w, $x y $z a partir de los valores del recurso continuo guardados en sus campos 245, 001 y 022 $a, respectivamente.

Información proveniente del propio número

El campo 008 del registro de un número se genera, lógicamente, a partir de la información disponible de ese número, rellenándose las posiciones 7-14 con la fecha exacta del número, indicando en la posición 6 del campo el código 'e', que expresa que se trata de una fecha detallada. De este modo, en la búsqueda y recuperación de información DIGIBIB utiliza esta información para realizar búsquedas por una determinada fecha de publicación o por un rango de ellas, obteniendo resultados altamente precisos. El resto de información del 008 de

un número es heredada de la información de la publicación, como es el caso de la lengua, que es la misma para todos los números, salvo que en el proceso de validación se especifique otra circunstancia.

Otro campo digno de mención es la generación del título del número, el campo 245, cuyo subcampo $a se compone de la concatenación del título de la publicación, es decir su campo 245 $a, y de la cronología y enumeración del número. De esta manera, todos los registros de números de una misma publicación comparten el título, como es lógico, pero permiten generar listados de números por su cronología lo que ofrece grandes ventajas en la presentación y ordenación de los resultados de una búsqueda.

La información descriptiva del estado de conservación de un número, como manchas, arrugas, partes desprendidas, papel quebradizo, etc., junto con errores de numeración, paginación, etc., introducidos durante los procesos de digitalización y validación, puede llegar a ser información realmente útil para determinar el estado en el que se encuentra la prensa, motivo por el cual se han incluido estos detalles en los campos 5XX del registro de número. De igual modo se puede proceder con los cambios de editor, impresor, los cambios de tamaño, número de páginas, etc. Una vez incluidos estos cambios en los registros de número podrá saberse no sólo que a partir de un determinado número se produce un cambio sino a qué números concretos afecta este cambio. Obviamente, esto es de gran importancia para la búsqueda de información y para la reconstrucción exacta de la historia bibliográfica de un determinado título.

Junto con la generación automática de un registro bibliográfico para un número, se incluye la generación de un registro de ejemplar con información como el de la Biblioteca que conserva el ejemplar (campo 852) “copiada” del registro de ejemplar que tenga asociado el registro bibliográfico de la publicación; la información del patrón de periodicidad, declarada en los campos 853/855 y los enlaces a las imágenes digitalizadas asociados al ejemplar en forma de campos 856.

Repositorio OAI

Las descripciones de números así obtenidas se pueden exportar a otros esquemas de metadatos normalizados como *Dublin Core* o *Europeana Semantic Elements,* lo que amplía enormemente la difusión de una biblioteca virtual o digital, permitiendo su recolección por parte de recolectores OAI como Hispana, Europeana u OAIster.

Todas las bibliotecas virtuales o digitales de prensa histórica que se mencionan en esta comunicación participan en Europeana, a través de Hispana. Hispana es un recolector de metadatos, pero también un repositorio OAI[43]; con este doble mecanismo recolecta metadatos de repositorios OAI españoles y es, a su vez, recolectada por Europeana, formando parte de este entramado como agregador nacional español. En tanto que repositorio, Hispana ofrece no sólo los registros bibliográficos de las publicaciones como un todo, sino también los registros de números, siendo éstos últimos los que tienen asociadas las imágenes digitalizadas. Es indudable que la generación de estos registros de número ha facilitado en gran manera la participación de bibliotecas digitales de prensa en proyectos nacionales e internacionales garantizando al mismo tiempo las funcionalidades típicas de búsqueda en la prensa histórica. Es decir, una vez recolectados los registros de número se siguen posibilitando las búsquedas

[43] http://goo.gl/ACjpR

por fechas o rangos de fechas, materias, ámbito geográfico, etc., en los sistemas de información que han recolectado esta información. Dicho de otro modo, los registros de ejemplar se han generado conforme a unos requisitos funcionales que pueden perfectamente mantenerse una vez recolectados.

Europeana Data Model y Linked Open Data

Durante la redacción de esta comunicación se ha procedido a instalar la versión 7.0 de DIGIBIB, tanto en la Biblioteca Virtual de Polígrafos[44], como en la Biblioteca Virtual de Prensa Histórica. Pronto el resto de las bibliotecas que hemos mencionado anteriormente pasarán a esta nueva versión. La versión 7.0 de DIGIBIB es ya –casi- una biblioteca Linked Open Data, puesto que es capaz de generar registros conforme al Europeana Data Model, versión 5.2.1[45], que no deja de ser un sabor de Linked Open Data (aparte de uno de los pilares de la fase Danubio de Europeana). Europeana Data Model reutiliza cuatro *namespaces*, RDF, OAI-ORE, SKOS y DC Terms y, desde luego, está previsto enriquecer, agregar[46], por utilizar la terminología de Europeana, los correspondientes registros con vocabularios tales como GeoLinkedData[47], *dataset* de nombres y lugares geográficos elaborado por el Instituto Geográfico Nacional, así como la vinculación con la Lista de Encabezamientos de Materia para Bibliotecas Públicas, que transformadas en SKOS, ha implementado la Subdirección General de Coordinación Bibliotecaria en un registro CKAN justamente para este fin. Están previstos también enriquecimientos mediante el uso de VIAF[48]. Se dan así los primeros pasos para que esos 7 millones de páginas a las que hemos hecho referencia en esta comunicación, no sólo estén en Hispana o en Europeana, sino que además aparezcan en la nube de Linked Open Data.

Xavier Agenjo, Bibliotecario. Director de Proyectos de la Fundación Ignacio Larramendi y de la Polymath Virtual Library [xavier.agenjo@larramendi.es]

Jesús Domínguez. Director del Departamento de Informática. DIGIBÍS, Producciones Digitales [jesus.dominguez@digibis.com]

Francisca Hernández. Bibliotecaria. Consultora de DIGIBÍS, Producciones Digitales [francisca.hernandez@digibis.com]

Fernando Román. Analista Programador. DIGIBÍS, Producciones Digitales [fernando.roman@digibis.com]

[44] Agenjo Bullón, Xavier; Hernández Carrascal, Francisca. *La Biblioteca Virtual Ignacio Larramendi desde la perspectiva LOD y EDM*. En: *I Seminario Internacional de la Biblioteca de Galicia : Santiago de Compostela, 7-9 de abril de 2011*. [http://goo.gl/uFrXx]

[45] http://goo.gl/ojlL

[46] Agenjo Bullón, Xavier; Hernández Carrascal, Francisca. *Perspectivas europeas en el desarrollo funcional de los sistemas de información: la agregación de datos del Europeana Data Model*. En: *FESABID'11 : XII Jornadas Españolas de Documentación : Málaga, 25-27 de mayo*. [http://goo.gl/Ig4BV]

[47] http://goo.gl/S65B9

[48] http://viaf.org/