

PROCESOS DE DIXITALIZACIÓN
E RECUPERACIÓN DE MATERIAL EN GALICIA

Daniel Buján Núñez

Centro Superior Bibliográfico de Galicia (CSBG)

A dixitalización do patrimonio bibliográfico preséntase como un proceso asociado á recuperación, conservación e difusión do patrimonio bibliográfico (e documental), neste caso que nos ocupa, no ámbito galego, que consiste, actualmente, na reprodución dixital, almacenamento e difusión, esencialmente por internet, do patrimonio manuscrito ou impreso conservado nos diversos centros galegos con coleccións de especial interese para Galicia ou en centros exteriores que conteñan fondos que poidan ser considerados de interese para a nosa cultura (a produción ligada aos centros galegos en América é o mellor exemplo).

Na presente comunicación preténdese expoñer, dunha forma breve, os inicios, os obxectivos e circunstancias, así como a actualidade e a actual planificación de obxectivos cara á dixitalización e difusión do patrimonio bibliográfico galego, centrándonos de forma especial na política levada a cabo polo Centro Superior Bibliográfico de Galicia (CSBG), só ou en colaboración con outras entidades como o Consello da Cultura Galega, a Fundación Cidade da Cultura ou a Biblioteca da Universidade de Santiago de Compostela, entre outras.

Non é necesario dicir que todo este proceso está intimamente ligado a un desenvolvemento tecnolóxico, no que as posibilidades deste foron abrindo ventás moitas veces inesperadas, e que por esa razón ten influído de forma fundamental na fixación de políticas e obxectivos para este longo labor de recuperación. Paréceme importante incidir neste aspecto, en canto elementos

que non deixan de ser tecnolóxicos, como o da distribución das imaxes por liñas de alta densidade. Nestes momentos está cambiando de forma drástica o mundo do acceso e consulta aos textos do patrimonio impreso, aquí e en todo o mundo.

Os puntos, polo tanto, nos que quero estruturar esta participación son, en avance, os seguintes:

Na primeira parte da exposición hai que sinalar as bases funcionais, de política patrimonial, tecnolóxicas e normativas do proceso:

1. En primeiro lugar, as razóns, vantaxes e posibilidades deste novo medio de conservación e transmisión de patrimonio.

2. En segundo lugar, as bases tecnolóxicas nas que se apoia o proceso, tanto no referente á propia dixitalización como á súa difusión por redes como internet.

3. En terceiro, lugar os aspectos referentes á recuperación da información e á normalización no acceso ás imaxes.

Nestes dous últimos puntos só intentarei sinalar aqueles elementos que teñen maior relación co proceso en Galicia, dado que esta participación se complementa coa do meu compañeiro Xavier Agenjo, un dos mellores expertos nos aspectos de normativa de descrición e acceso ás imaxes e no mundo relativamente novo dos metadatos, que vai estenderse neses temas concretos na súa participación.

Na segunda parte centraríame de forma máis concreta nos procesos de dixitalización do patrimonio bibliográfico levados a cabo en Galicia.

AS RAZÓNS DA DIXITALIZACIÓN

A dixitalización dos fondos do patrimonio bibliográfico e documental preséntase cun múltiple obxectivo:

1. A conservación do patrimonio

É indubidable que o carácter perecedoiro de moito do material sobre o que se asenta a cultura escrita e impresa resulta un importante hándicap cara á súa conservación. Este perigo afecta a algúns materiais de forma especial, como a prensa, o que fai que

cada vez sexa máis difícil atopar coleccións en bo estado e, ao mesmo tempo, que sexa importante realizar o antes posible unha copia de seguridade do material do que teñamos constancia da súa existencia, de forma que, en caso de perda do orixinal, contemos sempre, polo menos, cunha copia da publicación.

Esta política de conservación complétase, ademais, coa posibilidade que a copia dixital ofrece de ser consultada en lugar do orixinal (polo xeral, nunha busca de información, a consulta do orixinal non é precisa), o que supón a posibilidade de restrinxir o uso deste e de, polo tanto, contribuír á súa preservación.

2. As posibilidades de recuperación do patrimonio

O feito de non ter a necesidade de contar co orixinal facilita a obtención de copias e o feito de poder contar con exemplares dixitalizados, que sería, doutra forma, imposible contar con eles se fose preciso adquirir ou contar co orixinal para a súa consulta.

Tamén facilita a integración de coleccións, especialmente naqueles materiais (volvemos ao caso da prensa) que, polas súas características, non sempre están completos e que, ás veces, se poden completar integrando diversas coleccións, mantendo ao mesmo tempo as coleccións na súa situación física.

Non é necesario engadir que procesos deste tipo inciden tamén, de forma especial, na colaboración entre organismos e no desenvolvemento de normas e proxectos comúns de salvagarda e difusión do patrimonio.

3. A difusión do patrimonio

As copias, tanto que sexa en microfilme como en soporte dixital, permiten, coa súa fácil reprodución, multiplicar os puntos de consulta dun impreso sen a necesidade de contar co orixinal. Coa dixitalización, que permite a súa transmisión por canles de comunicación coma a de internet, estas posibilidades vense multiplicadas ata o infinito.

O SOPORTE TECNOLÓXICO

Este aspecto da dixitalización é, posiblemente, o que nos últimos anos sufriu unha maior evolución.

Por unha parte, no propio proceso de dixitalización, que conta actualmente con equipamento especialmente adecuado para a dixitalización de impresos –anteriormente adoitábase utilizar o mesmo microfilme como etapa intermedia cara á dixitalización. Por outra parte, o desenvolvemento de programas e normas para a compresión de imaxes, que vai ter unha importancia vital cara á súa transmisión por redes como internet.

Nestes momentos, as características da dixitalización que se está a producir no CSBG, que se seguen como norma xeral para os impresos (algúns materiais coma a cartografía precisan resolucións máis altas), son:

Características da dixitalización de textos en b/n: valores mínimos:

Resolución: 300 ppp

Profundidade de bits: 8 bits (256 niveis de grises)

Características da dixitalización de textos en cor: valores mínimos:

Resolución: 300/600 ppp

Profundidade de bits: 24 bits (16,7 millóns de tons de cor)

O formato de saída será TIFF sen compresión

Para o acceso á imaxe a través de internet realízase outra copia, neste caso comprimida, de forma que pese o menos posible (entre 50 e 200 k), pero que sexa perfectamente lexible.

As características normalmente utilizadas das imaxes para visualizar en internet son as seguintes:

Resolución: 200 dpi (b/n)

Profundidade de bits: b/n

Compresión: TIFF G 4

A transmisión por redes de comunicación é, sen dúbida, o elemento que vai ter maior transcendencia na actualidade. Nas primeiras fases do proceso de dixitalización de patrimonio prevíase que as copias se consultasen directamente no ordenador persoal ou, no mellor dos casos, a través de intranets no propio centro. Non se trataba, en calquera caso, dunha función demasiado diferente da que cumpría no seu momento a microfilmación: copia de seguridade e posibilidade de consulta en lugar do orixi-

nal. Incluso nos primeiros momentos de internet esta función de difusión a través da rede non estaba plenamente operativa, en función, sobre todo, do enorme peso, mesmo comprimidas, das imaxes, para a velocidade que a rede primitiva tiña para a transmisión de información.

Actualmente, non obstante, as características técnicas das redes de transmisión de datos permiten acceder a través de internet a imaxes dun peso medio (e mesmo alto, aínda que máis lento) a velocidades xa perfectamente operativas. Con isto cambia radicalmente o panorama da consulta. Por sinalar só algunhas das moitas vantaxes:

- posibilidade de consultar o patrimonio dende o terminal persoal, evitando os desprazamentos. Dende o propio domicilio ou lugar de traballo, en calquera lugar (conectado) do mundo, pódese acceder a todas as coleccións accesibles por internet (cada vez máis, en progresión xeométrica);

- rapidez das consultas e do acceso aos documentos;

- posibilidade de consulta múltiple: moitos usuarios ao mesmo tempo;

- horarios e tempos ilimitados. Pódese acceder en calquera momento de calquera día;

- posibilidade de obter copias ao momento, tanto en formato dixital como en papel;

- posibilidades acrecentadas de busca e localización de imaxes, ao poder aplicar metadatos e programas de análises e busca de información.

O ACCESO Á INFORMACIÓN

Para poder consultar estas imaxes é preciso a creación de elementos de análise e busca que permitan a súa recuperación por parte dos investigadores. Estes elementos de descrición teñen diversas formas e responden a diferentes estruturas de busca. Sobre eles inciden, ademais, as novas formas de normalización de recu-

peración, os metadatos, estruturas de descrición da información contida nas coleccións dixitais co fin de permitir o aproveitamento e recuperación dos datos contidos nelas.

Estas estruturas de información anexas ás imaxes, baseadas nunha descrición en formato recuperable a través de buscas, polo xeral, aínda que non unicamente, en bases de datos, son indispensables para a recuperación da información contida ou implícita na documentación, que, como imaxe que é (estamos a falar sempre de copias dixitais dos orixinais), carece dunha estruturación que permita a recuperación normalizada dos seus contidos.

Na colección dixital do CSBG utilízanse varias formas de información adxuntas ás imaxes:

- Unha forma que podemos considerar «automática», que consiste en aplicar un programa de OCR (recoñecemento de caracteres) sobre o texto e posteriormente un programa de busca en texto libre aplicado sobre esta información xa en formato texto. É o sistema que se está a aplicar para a prensa, fundamentalmente pola dificultade que representa unha indización ou catalogación dos contidos, ás veces mínimos, do inxente número de páxinas existentes de prensa.

- Unha forma que podemos considerar «documental», que consiste na elaboración de campos ou formatos que permitan a busca e posterior acceso ás imaxes.

Esta estruturación realízase de dúas formas:

- A primeira, semiautomática, a través dos propios nomes dos ficheiros.

A estrutura dos nomes dos ficheiros, que se está a aplicar para as imaxes dixitalizadas do CSBG, é como segue:

No caso da prensa:

tít_anomesdía_n.º da publicación.

xxxx_xxxxxxxxxx_xxxxx

(nome da carpeta en que se han de gardar as imaxes das páx. dun n.º dunha publicación periódica).

tít_anomesdía_n.º da publicación_n.º de páx.

xxxx_xxxxxxxxxx_xxxxx_xxxxx

(nome da imaxe dunha páx. dunha publicación periódica).

As imaxes de cada páxina dun número dunha publicación periódica van gardadas dentro de cadansúa carpeta. Todos os arquivos dunha mesma obra almacénanse nun mesmo directorio. Inclúese un arquivo a maiores, chamado «documento.txt», no que se realiza a descrición da obra dixitalizada (polo menos título, autor, ano de publicación e descrición física).

As imaxes procedentes de monografías estrutúranse do seguinte xeito:

- Asignación dun código único da forma CODG. Este utilízase como clave única na base de datos específica correspondente e tamén como nome do directorio físico, que almacena todas as imaxes das páxinas do volume.
- As imaxes correspondentes a un volume van nomeadas en orde alfabética na forma CODG_NNNN_AAAAA, na que CODG é o código do volume, NNNN é o número de páxina físico (non a numeración interna) e AAAAA é a información relevante á páxina (por exemplo, se é cuberta –cub01...–, o número da paxinación interna –p0001...–, anuncios –anu01...–, etc.).

- A segunda forma de acceso ás imaxes é a que se realiza a través das bases de datos bibliográficas a partir dun campo de enlace entre o formato MARC correspondente e o arquivo coas imaxes do texto

Á parte das informacións e estruturas de recuperación anexas ás imaxes existe unha serie de páxinas de consulta que intentan facilitar a localización da información buscada. Segundo o tipo de documento elabórase un tipo diferente de sistema de busca:

- para o caso de prensa periódica elabórase unha listaxe de cabeceiras que permite a busca por título, lugar de edición e por ano. A páxina atópase en www.csbg.org/prensagalega;

- para o caso de monografías elaborouse un lugar web específico de «autores galegos», no que se pretende reunir para cada autor, e nunha páxina dedicada a este, a información existente na rede e impresa e o acceso á súa obra. A páxina atópase en www.autoresgalegos.org;

- para a cartografía estableceuse un acceso con busca xeográfica (www.csbg.org/cartografia).

UN PEQUEÑO RESUMO DO ACTUAL PROCESO DE DIXITALIZACIÓN DO PATRIMONIO

Non convén esquecer, antes de máis nada, que o proceso de dixitalización do patrimonio bibliográfico é un proceso que, en certo modo, continúa un proceso precedente, que é o da microfilmación do patrimonio. De feito, e xa dende os primeiros momentos, vaise proveitar o traballo realizado neste campo (a dixitalización a partir do microfilme é moito máis barata que a partir do orixinal) para a confección da biblioteca dixital. Tal é o caso das coleccións de prensa dixitalizada da Biblioteca Universitaria de Santiago e do Consello da Cultura Galega (prensa galega na Arxentina) ou, máis recentemente, coa colección de «Galicia Nueva» conservada no concello de Vilagarcía.

En calquera caso, o proceso de dixitalización comeza, polo menos no que ao CSBG atinxe, en febreiro de 1997, ano en que se aproba a iniciación dun expediente para a dixitalización, por concurso público, de diversas publicacións conservadas nas bibliotecas públicas nodais de Pontevedra e de Lugo.

As razóns desta elección son varias:

- Non hai que esquecer que, a pesar do seu inhabitual nome, o CSBG é, dende 1979, a biblioteca nacional de Galicia e ten como función, tal e como aparece sinalado na Lei 14/1989, de bibliotecas de Galicia, a recuperación, conservación e difusión do patrimonio bibliográfico;

- O CSBG carece de coleccións propias nese momento. A dixitalización permítelle ao CSBG, por un lado, centralizar nunha colección dixital os fondos dispersos polas diferentes bibliotecas galegas (e de fóra de Galicia con fondos galegos) e, por outro lado, dar o servizo da súa consulta de forma centralizada e conxunta. Existe a posibilidade de contar con copias sen necesidade de contar co orixinal.

- Elíxese nese momento comezar pola dixitalización da prensa polo estado delicado do material e polo perigo de perda das coleccións, o uso das cales precisa cada vez máis restricións, polas dificultades para a súa consulta debido á escaseza e á dispersión dos fondos e pola importancia informativa contida nos propios fondos.

- Elíxese, así mesmo, o século XIX en canto buscasen cabeceiras sobre as que non existan dereitos vixentes de propiedade intelectual. Este coidado manterase durante todo o proxecto: importancia de manter o debido respecto a calquera dereito de propiedade que estea vixente sobre o material dixitalizado e difundido por calquera medio. Por esa razón os fondos actualmente dispoñibles a través das páxinas do CSBG son na meirande parte anteriores a 1900.

A partir desas datas o proceso e dixitalización é continuo, organizándose a colección en diversos grupos.

Cartografía Histórica de Galicia

De momento conta cun total de 371 mapas dixitalizados. Todos eles levan asociado un enlace que nos leva á catalogación bibliográfica do mapa e dende aí pódese acceder ao mapa dixitalizado. Encóntranse todos en formato JPG e pesan unha media de 470 Kb, aínda que hai mapas cun peso dende 39 Kb ata 5364 Kb.

Gravado

O apartado dos gravados na páxina web da Biblioteca Dixital aínda non está activado, pero existe bastante documentación que xa está case lista para a súa introdución e posta en funcionamento. Os gravados pertencen a autores de orixe galega ou de orixe non galega, pero que publicaron gravados seus en publicacións galegas.

Nestes momentos atópanse dixitalizados os gravados que se corresponden coas publicacións periódicas *La Ilustración Gallega y Asturiana* e *Spes*, á parte dalgúns gravados soltos, especialmente do século XVIII.

Para o acceso preténdese, o mesmo que en autores galegos, crear unha páxina para cada autor na que sintetizar a información existente e dar acceso á súa obra dixitalizada.

Manuscritos

O CSBG conta con 73 manuscritos dixitalizados procedentes da Universidade de Santiago, ademais dunha serie de manuscritos dixitalizados procedentes das súas propias coleccións, en concreto:

- Fondos de Manuel Murguía e outros: contén «Apuntes Varios», «Follas Soltas», «Historia de Celanova» e «Historia de Rivadavia». En total 184 imaxes en calidade alta, TIF sen comprimir.
- Outros manuscritos: contén: «A Través del Atlántico», «Curiosidades», «Diario de Pontevedra....». En total, 257 imaxes en calidade alta, TIF sen comprimir.
- Fondos de Wenceslao Fernández Flórez: (todo en calidade alta, TIF sen comprimir):
 - Banda deseñada: *El Malvado carabel*: 87 imaxes.
 - Correspondencia: 297 imaxes.
 - Documentos persoais: 28 imaxes.
 - Manuscritos: 3045 imaxes.

Monografías

Actualmente cóntase con 3184 monografías procedentes na case totalidade da Biblioteca da Universidade de Santiago.

As imaxes dixitalizáronse tanto en formato de alta calidade como en TIF-G4 para a súa consulta na web. Créase tamén un rexistro no Catálogo Colectivo, no que se inclúen os campos 035, onde aparece o Código co cal se identifica a monografía, e o campo 856, no que aparece un enlace que leva directamente á imaxe dixitalizada.

Prensa Antiga

As primeiras publicacións que chegaron ao CSBG procedían da biblioteca nodal de Pontevedra e da de Lugo, así como imaxes dixitalizadas a partir de microfilmes da Biblioteca Universitaria de Santiago e do Consello da Cultura Galega. O estado actual pode resumirse nunha oferta de máis de 150 publicacións, xa dispoñibles dende a web (non todas se atopan completas, boa parte delas recollen o que se conserva nas coleccións dixitalizadas), que suman aproximadamente un total de 750 000 imaxes procedentes de diferentes e moi diversas institucións.

O PROCESO METODOLÓXICO

En canto á metodoloxía de traballo, a organización do proceso distribúese da seguinte maneira:

Supervisións dos procesos de dixitalización: selección do material que se vai dixitalizar, establecemento das pautas para proceder ao escaneado e a forma de entrega do traballo (CD, nalgúns casos en DVD).

Revisión e tratamento de imaxes

Recepción de CD de prensa dixitalizada previamente seleccionada.

Pasar aos PC e visionado, revisión e valoración positiva ou negativa da calidade das imaxes

Tratamento das imaxes: estas deben estar no mesmo formato. Se non están hai que comprimilas para a súa inserción na web (TIF-G4).

Adxudicación dun nome de identificación para cada imaxe e mesmo para cada exemplar, que consiste nun código alfanumérico.

Organización

Realización de copias de seguridade no PC dos datos xa tratados.

Traslado e introdución no servidor dos directorios xa creados para o seu posterior tratamento a través do programa de recoñecemento de caracteres.

Comprobación entre os fondos recibidos e xa almacenados para evitar posibles duplicados.

Informe dos novos datos introducidos.

Actualización da base de datos

Existe unha base de datos creada en Access na que se recollen os datos específicos de toda a prensa que aparecen pola súa vez na web, organizados en 17 campos.

Introdución das novas publicacións a medida que se van recibindo.

Actualización de información.

Investigación en diferentes fontes documentais para introducir unha breve información sobre a historia da publicación, que se vai inserir nesta base de datos.

Catalogación

Nos casos que proceda, creación de rexistros bibliográficos dentro do Catálogo Colectivo de Publicacións Periódicas e introdución dun campo específico cun link que leva directamente á visión da imaxe.

Copias de seguridade: ao final do proceso requírese facer unha copia en DVD por cada publicación.

Todo este proceso realízase en colaboración co Centro Multimedia de Galicia e coa Fundación Cidade da Cultura de Galicia.

RESUMO

O número de imaxes que actualmente forman parte da colección dixital de patrimonio bibliográfico de Galicia:

Prensa: aprox. 750 000 imaxes.

Cartografía: 371 imaxes.

Manuscritos: 21 150 imaxes.

Monografías: 263 602 imaxes (prov. en espera do último envío da Universidade de Santiago).