Técnicas de producción masiva de contenidos para bibliotecas digitales.

Francesca Marí-Domènec y Alejandro Bia-Platas

Biblioteca Virtual Miguel de Cervantes,
Universidad de Alicante,
Apartado de correos 99, E-03080, Alicante, España
{francesca.mari, alex.bia}@cervantesvirtual.com
http://cervantesvirtual.com/

Resumen. La Biblioteca Virtual Miguel de Cervantes[1] es probablemente el proyecto de biblioteca digital de lenguas hispanas con el mayor número de libros en línea en la actualidad. Casi tres años después de su puesta en funcionamiento, la Biblioteca Virtual Miguel de Cervantes se ha convertido en la página web de consulta obligada para los amantes de la literatura, la historia y la investigación humanística en el mundo hispano. Otra característica de la Biblioteca Virtual Miguel de Cervantes es el variedad de medios digitales que ofrece, que incluye vídeo, audio, imágenes gráficas y libros tanto en versión facsímil como en texto digital. Estos últimos destacan por la calidad de su formato de presentación como por su funcionalidad.
La finalidad de este artículo es la de dar a conocer los procesos que integran el sistema de producción de estos libros digitales. Trataremos temas como la selección de textos y el sistema de catalogación creado a medida para nuestra biblioteca. Analizaremos el proceso de digitalización: criterios, herramientas informáticas utilizadas, problemas del tratamiento tanto de los formatos textuales como digitales y métodos de control de calidad. Haremos un breve repaso de los criterios de corrección y edición de los libros así como de la creación y adaptación de los programas informáticos para el procesamiento de los mismos.

1 Modelos de producción

La Biblioteca Virtual Miguel de Cervantes ha elaborado diferentes procedimientos de producción, uno para cada tipo de recurso multimedia (texto, facsímiles digitales, etc.) [1]. En este artículo[2] estudiaremos el flujo de trabajo para la producción de libros en formato texto [2][3][4] (ver figura 2) y de libros en versión facsímil digital [5] (ver figura 3). Con respecto al software que hemos utilizado para poner en práctica estos modelos de producción, se trata de una integración exitosa de software comercial con programas de desarrollo propio de la biblioteca

[1] http://cervantesvirtual.com/
[2] Versión completa de este artículo con diagramas y figuras en http://cervantesvirtual.com/research/articles/index.shtml

virtual. Como filosofía general hemos intentado no reinventar la rueda, es decir, no invertir esfuerzos en desarrollar software cuando existe una alternativa comercial que cumple los requisitos. Sin embargo hemos tenido que diseñar software propio para ciertas tareas críticas.

2 Recursos humanos

Nuestra biblioteca está formada por un equipo interdisciplinario de técnicos cuyo número oscila en torno a un centenar de personas (ver la distribución porcentual por áreas en la figura 1).

Un 6% de nuestro personal son bibliotecarios, un 15% se dedica a la digitalización de textos e imágenes, un 60% a corrección y marcado de textos, un 9% a investigación, desarrollo y mantenimiento informático y un 5% a diseño gráfico. El 3% restante corresponde a la administración del proyecto.

Con respecto a la formación de los correctores que forman el grupo especializado más grande dentro de la biblioteca, un 70% de ellos son filólogos, el 52% de filológica hispánica, un 15% poseen formación en historia y geografía y el 15% restante proviene de diversas disciplinas humanísticas como la educación, filosofía, psicología, sociología y traducción.

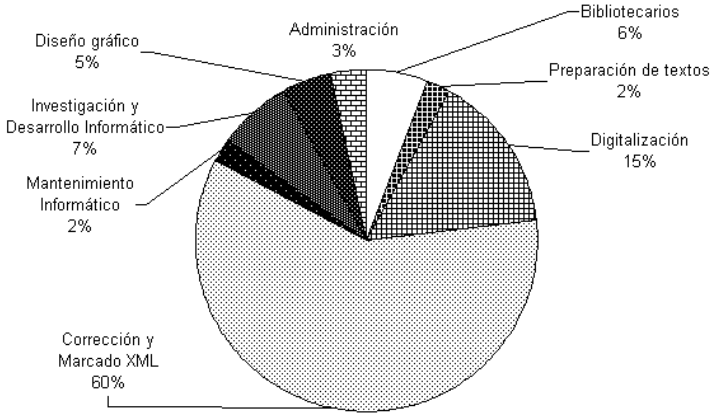

Fig. 1. Recursos humanos en la Biblioteca Virtual Miguel de Cervantes.

3 Catalogación

El nacimiento de la biblioteca digital no significa, de ninguna manera, el rechazo de las técnicas utilizadas para la creación y organización de una biblioteca tradicional. Por el contrario, creemos que una biblioteca digital debe cumplir

las mismas funciones que una biblioteca tradicional y, además, agregar otras nuevas que sólo son posibles mediante el uso de las nuevas tecnologías. Así es que nuestra biblioteca tiene también un proceso de registro y catalogación. La llegada de los libros físicos a nuestras dependencias sigue diversos cauces. Por una parte, cada uno de los responsables de las diferentes áreas hace una selección de los libros y materiales que tiene pensado incluir a la hora de crear o ampliar un portal o sección. Cuando trabajamos en colaboración con otras instituciones, por ejemplo, la Biblioteca de Cataluña o El Colegio de México, son éstas las que generalmente marcan cuáles son sus objetivos respecto a contenidos, y por lo tanto cuáles son los materiales a catalogar. Al mismo tiempo, nuestro programa de digitalización de obras de autores clásicos hace que el Área de Catalogación siga también una rutina de búsqueda de obras esenciales, obras clásicas que deberían estar en cualquier biblioteca. El Área de Catalogación está formada por seis especialistas en documentación y biblioteconomía que seleccionan las mejores ediciones y verifican que estén libres de derechos de autor. También en el Área de Catalogación se gestiona la cesión de derechos de autor y los permisos para publicar en Internet todo tipo de materiales: fotografías, grabaciones de audio, folletos, guiones cinematográficos, entrevistas, etc.

¿Cómo se registran y catalogan los libros y demás recursos? [6] Ingresando en un programa informático datos tales como autor, fecha nacimiento y muerte, edición original, mención de publicación, procedencia del original, información sobre la cesión de derechos, el tipo de obra, el tipo de material, el soporte, la materia, etc., en suma, todos aquellos datos que se usan en una biblioteca tradicional. Existe un libro de registro electrónico y una base de datos de catalogación que se comparan automáticamente para detectar inconsistencias. En el área de catalogación empieza el proceso y también acaba: todos los materiales son registrados y se hace una ficha de catalogación que los acompaña por su recorrido a través de las diferentes áreas de producción de la biblioteca y cuando el trabajo de digitalización está terminado, antes de almacenar o devolver el material a las instituciones o bibliotecas de origen, se hace una última revisión cotejando el material original con el digitalizado y se le da el visto bueno.

Cabe destacar la enorme importancia de esta parte del proceso de producción. El registro es esencial por dos motivos, primeramente porque clasificar el material como texto, imágenes, facsímil, vídeo, audio, etc., determinará el método de producción a seguir [2][7][8], y el hecho de que se lo registre nos servirá para identificarlo después y conocer en qué parte del proceso se encuentra. El buen flujo del trabajo dependerá de esta identificación y del posterior seguimiento.

La arquitectura de datos de nuestro sistema de catalogación se basa en el formato MARC [9][10], aunque para algunos proyectos especiales como el portal que hemos diseñado para la Biblioteca de Palacio Real (Patrimonio Nacional) hemos usado formatos derivados del esquema de marcado de textos TEI [11]. Nos referimos al formato del TEI-header (cabezal con metadatos de los documentos TEI), aunque para manuscritos hemos usado una ampliación de éste llamada

MASTER[3] [12]. La tendencia es abandonar el formato MARC en favor de otros formatos más modernos y adecuados para obras exclusivamente digitales, como es el caso de los formatos basados en la norma TEI o el Dublin Core [13].

Desde el punto de vista informático, los datos se almacenan y gestionan mediante un sistema de aplicación de bases de datos de construcción propia que usa una base de datos relacional. Este sistema aporta una interfaz que permite las operaciones tradicionales de alta, baja y modificación de registros entre otras. Este sistema se usa para la gestión interna de los datos de catalogación. Para búsquedas y consultas externas a través de Internet, los datos de esta base se exportan a otra base de datos orientada a objetos (basada en el producto ObjectStore), y las búsquedas se realizan mediante otro programa de diseño propio que construye una estructura TRIE en memoria del servidor que es muy rápida y eficiente en las búsquedas [14]. Este programa esta realizado en Java, al igual que la mayoría de nuestra programación de servidor[4].

Este sistema permite hallar un título aunque se busque incompleto: p.ej. acepta "Quijote", "Don Quijote", "Ingenioso Hidalgo", es decir, subconjuntos del nombre completo. Ignora palabras poco significativas como artículos y preposiciones. Permite búsquedas por título, autor, materia y época.

A este buscador para información de catálogo se suman otros como un buscador por número de referencia o localizador de la obra, un buscador sobre contenidos (busca palabras en el propio texto de las obras) y un buscador sobre páginas de navegación de la biblioteca (busca en páginas web de portales y secciones de nuestro sitio web). A estos ha de sumarse próximamente un nuevo buscador de contenidos que además de efectuar búsquedas sobre el texto de las obras permite buscar utilizando condiciones basadas en el marcado estructural XML-TEI [15]. Este buscador permitirá realizar búsquedas de texto muy potentes, indicando dentro de qué partes estructurales debe encontrarse la cadena buscada. Podremos buscar libros donde "Galdós" aparezca como autor, o dentro del título, o como parte del texto de la obra, siendo estas búsquedas muy diferentes. Por ejemplo, podremos buscar obras donde la palabra "Paloma" aparezca exclusivamente como personaje de una obra de teatro, y no con otros usos. Estas condiciones de búsqueda basadas en el marcado estructural permiten hacer búsquedas muy precisas y sofisticadas.

4 Digitalización

El siguiente paso es la digitalización del material (ver figura 2). El objetivo de esta tarea es obtener, a partir de los textos en papel, ficheros de texto digital con el menor número posible de errores. Esta etapa se divide en dos procesos consecutivos: escaneo y reconocimiento de caracteres ópticos. Para ambos utilizamos el programa OmniPage versión 11 que logra un reconocimiento casi perfecto en

[3] El proyecto MASTER trata de la gestión de metadatos para manuscritos antiguos y fue llevado a cabo por un grupo de universidades europeas, Oxford entre ellas.

[4] Utilizamos Java, Java Servlets para servicios interactivos cliente-servidor, y Java-Web-Server como software de servidor web.

textos de impresión moderna, donde el papel no tiene manchas y las letras son uniformes y continuas. Como este programa se sirve de un diccionario de castellano (o del idioma de elección) para reducir la cantidad de errores, no funciona igual de bien con textos de castellano antiguo que no siguen las normas gramaticales modernas. Si el propio libro, no sólo el texto, es antiguo, tampoco se obtienen buenos resultados. Esto sucede cuando el papel es amarillento (a veces con manchas), y los tipos de letra son irregulares, con discontinuidades en la impresión, y donde a veces aparecen símbolos antiguos para representar letras que ya no se usan como es el caso de la doble s, la s larga y otras grafías antiguas. En todos los casos, haya muchos o pocos errores, será necesaria una minuciosa corrección del texto contrastando con el original, pero queda claro que habrá más errores cuanto más antiguo sea el libro y el vocabulario de la obra.

Teniendo en cuenta los datos de catalogación y el estado en que se encuentra el material, se decide qué herramientas van a ser necesarias para digitalizarlo: escáner plano con o sin alimentador automático de hojas sueltas, escáner cenital, máquina fotográfica digital o escáner para microfilmes. Una vez escogido el tipo de escáner o cámara digital se ajustarán los parámetros de procesamiento de OmniPage con el objetivo de mejorar los resultados del reconocimiento de caracteres ópticos. Este es un procedimiento crítico que no se debe descuidar, ya que permitirá que los textos lleguen a los correctores con el menor número de errores de OCR posibles. Comprobar y mejorar los resultados del OCR es una actividad rentable, pues de ello dependerá en parte la rapidez posterior en la edición de los textos. Para ello, se creará un archivo de capacitación que sirve para mejorar el programa de OCR que lo incluirá en su diccionario para reconocer los nuevos términos en posteriores ocasiones. Una vez finalizado el proceso de digitalización se han de comprimir todos los ficheros creados del escaneado y OCR (imágenes y textos), con el fin de archivarlos en CD-ROM para su preservación y reutilización futura. De las imágenes se guardará tanto la copia de mejor calidad (TIFF), como la imagen recortada y adaptada para publicación web de menor resolución (JPEG comprimido con pérdida de calidad). Este archivo en CD-ROM forma parte de nuestro plan de preservación digital.

Los digitalizadores, al igual que se hizo en el Área de Catalogación, deberán rellenar una ficha con una serie de datos como, por ejemplo, el tratamiento del color del texto, el brillo, el tamaño, la orientación de los cuadernos, la calidad del original, la calidad del OCR, número de imágenes digitalizadas, retoques que se han realizado (márgenes externos, enfoque, tonos, eliminación de manchas, sellos, reconstrucciones), si existen folios en blanco, repetidos o sin numeración, etc. Los correctores podrán cotejar tanto el informe que se ha hecho en el Área de Catalogación como en la de digitalización para saber qué tipo de corrección se ha de realizar. A los correctores únicamente les llegará el texto impreso (original o en fotocopias) y el fichero de texto digitalizado. El Área de Digitalización cuenta ahora con un equipo de 13 personas, algunas especializadas en la manipulación de ciertos tipos de escáner: cenital, de microfilmes y microfichas, de negativos o cámaras digitales.

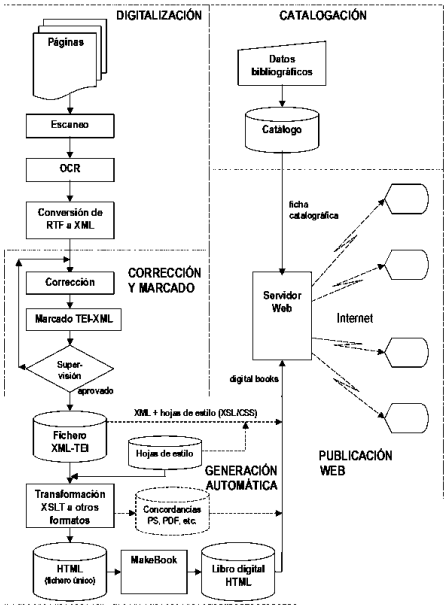

Fig. 2. Diagrama de producción de textos en XML.

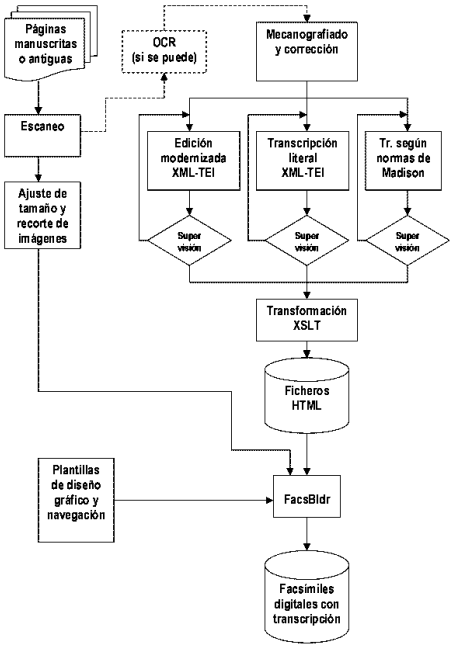

Fig. 3. Diagrama de producción de facsímiles digitales.

4.1 Conversión a XML

OmniPage no contempla al XML como un posible formato de salida, sin embargo nuestro proceso de corrección está pensado para ser hecho en XML. Llegado este punto nos planteamos una pregunta: ¿Cómo comenzar un texto XML de la manera más fácil posible? Hemos construido un programa que convierte automáticamente la salida en RTF de OmniPage a XML-TEI con un marcado básico. Obviamente, en un fichero RTF no existe la información suficiente como para transformarlo en un fichero XML con un marcado sofisticado, pero sin embargo, se puede obtener un fichero XML de partida con un marcado simple. El programa RTF2XML es un conversor programado en lenguaje C++ con Lex que fue diseñado siguiendo las especificaciones de RTF versión 1.5. Detecta párrafos, letras en negrita e cursivas así como saltos de página. A todos estos elementos se les agrega las correspondientes marcas XML: `<p></p>`, `<hi type="bold"></hi>`, `<hi type="italic"></hi>` y `<pb/>`. Además se crea un cabezal TEI con sus marcas parcialmente completadas[5]. Por último se convierten todas las letras con acento, las eñes y otros caracteres especiales, en entidades ISO 8859 (p.ej.: *á* en `&aacute;`). El documento XML obtenido de esta manera es un documento TEI válido, es decir, que cumple las reglas de la DTD que utilizamos y es un buen punto de partida para las etapas que siguen ahorrándonos el trabajo de importar un texto plano al editor de XML y tener que marcarlo desde cero. Si se trata de una obra en prosa donde la mayor parte de los elementos estructurales son párrafos, tendremos una buena parte del marcado resuelto. Por el contrario, si se trata de una obra de teatro, o aún peor, de teatro en verso, cuya estructura es de las más complicadas que podemos encontrar, seguramente tendremos que cambiar casi todas las marcas ya que en lugar de párrafos `<p>` tendremos diálogos `<speach>`, personajes `<speaker>`, y líneas de verso `<l>`.

No todos los textos a procesar siguen estos pasos. Algunos provienen de editores de textos como MS-Word o Wordperfect, ya sea porque se mecanografiaron directamente o porque tras la digitalización, se importaron en formato RTF para editarlos o corregirlos. Al pasar por estos editores seguramente se le dio formato a los títulos, a las nota de pie de página y a otros elementos estructurales del texto. Muchas veces hacemos esto sin siquiera ser conscientes de que estamos marcando elementos estructurales. Lo hacemos cada vez que asignamos a un título un estilo de encabezado y lo asociamos con un nivel jerárquico. Esta información de marcado se puede aprovechar. Para ello, hemos construido el programa HTML2XML que a partir del código HTML generado por estos editores produce un documento XML donde no sólo se reconocen los párrafos y unos pocos elementos más, sino que también se reconocen los títulos y su nivel jerárquico, pudiéndose, a través de ellos establecer divisiones estructurales en el texto las que corresponden a capítulos, subcapítulos, etc. Al igual que en el caso de RTF2XML, se reconocen cursivas y negritas, pero además también subíndices, superíndices y centrados.

[5] Recordemos que la norma TEI establece que un documento TEI consta de dos partes: un cabezal TEI o "teiHeader" con los metadatos (información sobre el documento), seguido del cuerpo de la obra entre marcas "text" [3].

En los párrafos se aprovechan indicaciones de alineación izquierda, derecha y centrada. Basándose en marcas de formato especiales puede incluso detectar indicadores de número de página. Se convierten también a XML los hiperenlaces y la inserción de imágenes gráficas. Este programa puede incluso distinguir entre simples párrafos y líneas de verso en base a ciertas características del formato, mediante técnicas similares a las utilizadas en otro trabajo sobre la aplicación de técnicas de extracción de información a las bibliotecas digitales [16]. Debido a que en estos ficheros HTML hay más información estructural, el programa de conversión puede generar XML con un marcado más rico que el que se puede obtener partiendo de RTF.

De todos modos en ninguno de los dos casos citados se obtiene un marcado XML completo. Sólo se obtiene un marcado parcial pero que ahorra mucho tiempo y esfuerzo si lo comparamos con empezar a trabajar a partir de un texto plano. Estos dos programas de conversión no son de propósito general sino que están diseñados para generar documentos XML-TEI específicos para la DTD que utilizamos en la biblioteca virtual, que es un subconjunto de la norma TEI adaptado a nuestras propias necesidades de marcado [17]. Con ellos tenemos dos modos diferentes de dar origen a un libro en XML de forma rápida y eficiente.

5 Corrección

¿Cómo llega el material a los correctores? Los archivos en texto llegan a la encargada del control de producción y son asignados automáticamente, después de hacer una valoración de la dificultad de lectura y señalar qué criterios de edición han de ser utilizados en la corrección, todo ello a través de un sistema de gestión de ficheros (workflow). Los libros físicos son repartidos por los coordinadores de las diferentes salas de corrección.

Los correctores disponen de un sistema de manuales electrónicos de orientación y ayuda en donde se explican los diferentes criterios de edición y de marcado XML-TEI accesibles por intranet. Existen normas de marcado XML-TEI para revistas, para tesis doctorales, para libros convencionales, así como para ediciones multimedia. Hemos diseñado también manuales con criterios de foliación para las ediciones facsimilares, y con criterios de corrección para los siglos XVI y XVII (basados en [18]), siglo XVIII, y siglos XIX y XX (basados en [19]). Se han creado también criterios de corrección para literaturas específicas tales como la literatura gauchesca o la literatura paraguaya. Los correctores disponen también en línea por intranet de material de consulta como son diccionarios de español, inglés, francés, un libro de estilo, un Tesoro como el Covarrubias [20], etc. Todos los criterios de edición han sido revisados por especialistas en cada uno de los siglos señalados. Contamos con la colaboración de Evangelina Rodríguez para los siglos XVI y XVII y de Enrique Rubio Cremades y Juan Antonio Ríos para los textos de los siglos XVIII, XIX y XX.

Se cuenta con un total de 44 correctores, cuatro filólogos que se dedican a la supervisión de todos los textos y a la revisión del marcado XML-TEI, dos a la fijación y actualización de criterios y tres al montaje de aquellos libros que vienen

únicamente en formato imagen. Algunos correctores están especializados en la corrección o el marcado de obras de un tipo específico como por ejemplo: literatura hispanoamericana, de historia, en lengua catalana, revistas y periódicos, teatro, trabajos de investigación como tesis y homenajes a hispanistas, etc. Los correctores anotan, durante la corrección de una obra, toda una serie de observaciones, problemas y dudas que serán después analizadas y resueltas por el Área de Supervisión. Con este personal producimos, como término medio, unas 175 novedades al mes.

Además, poseemos una unidad audiovisual que produce vídeos digitales de entrevistas a autores y obras en audio digital. Estos siguen otros métodos de producción que no trataremos en este artículo.

El trabajo de corrección, edición y supervisión de textos es el proceso de edición más costoso que se realiza en nuestra biblioteca debido a su meticulosidad, donde se intenta además, obtener ediciones propias mejoradas con respecto a las que se encuentran habitualmente en el mercado.

5.1 Edición de texto XML

Después de probar varios editores de XML, hemos elegido XMetaL principalmente por las ventajas que ofrece su interfaz de usuario. XML-Spy es también un excelente editor que no sólo soporta DTDs sino también varios tipos de "Schemas"[6], pero su interfaz es más adecuada para aplicaciones de procesamiento de datos que para la edición de textos literarios. Wordperfect 9 por su parte es un muy buen editor de textos para uso personal o de oficina pero su modo de trabajo en XML deja mucho que desear. Emacs es otro editor gratuito de libre distribución con muy altas prestaciones. Posee un modo para XML y poderosas herramientas de apoyo a la edición. Su único inconveniente es la interfaz que es poco amigable y bastante menos intuitiva que la de otros editores.

5.2 Problemas encontrados

La principal dificultad no fue técnica sino humana: cambiar la mentalidad de los correctores acostumbrados a los editores de texto comerciales y ayudarles a adaptarse a las nuevas herramientas requeridas para el marcado en XML. Sin embargo, el tiempo de aprendizaje fue menor de lo esperado. En esto, la formación filológica de la mayoría de nuestros correctores ha jugado un importante papel, hallándolos predispuestos a comprender el nuevo enfoque estructural.

En otros aspectos las herramientas de marcado XML actuales todavía presentan algunas desventajas para el idioma español, las cuales seguramente se solucionarán a corto plazo con el advenimiento de nuevas versiones. Una de ellas es la falta de diccionarios para la corrección ortográfica y ayudas gramaticales

[6] Los Schemas son una nueva forma de definir la estructura de un tipo de documento, como alternativa a las tradicionales DTDs, y con algunas ventajas respecto de éstas como el poder usar tipos de datos.

para el castellano, las cuales son fáciles de encontrar y aún más fáciles de usar en los editores de texto más tradicionales.

Una solución temporal a este problema fue utilizar herramientas paralelas para la corrección ortográfica, particularmente aquéllas que aceptan diccionarios de gran tamaño provistos por el usuario (por ejemplo ISPELL[7]) [21]. Esto nos permitió construir diccionarios especializados para el castellano antiguo de diferentes épocas (siglo de oro, siglos XVIII, XIX, y XX) que mejoraron la corrección de textos antiguos [22][23].

6 Producción de facsímiles digitales

Hasta este punto hemos tratado el proceso que sigue un libro en formato texto, sin embargo, la presentación de las obras facsimilares sigue un camino muy diferente. Cuando se trata de manuscritos o de impresos muy antiguos, a los que por un lado no se les puede aplicar el OCR y por otro, resulta interesante verlos en su aspecto original, optamos generalmente por la producción de un facsímil digital, a veces acompañado de una o más transcripciones en formato texto. Esto sigue otro modelo de producción diferente al de los textos digitales. En este caso, las transcripciones hay que mecanografiarlas ante la imposibilidad de efectuar el reconocimiento automático de caracteres. La imposibilidad de aplicar el OCR con éxito se debe a varias causas: las características físicas del papel (generalmente envejecido y manchado), por tratarse de un manuscrito o de un impreso de tipografía antigua e irregular y por el hecho de tratarse de léxico antiguo (castellano antiguo, catalán antiguo, latín, etc.) lo que impide el uso de técnicas automáticas de corrección de errores, como las que se aplican para mejorar la salida del proceso de OCR, que se basan en diccionarios modernos. El problema de la corrección de textos en castellano antiguo lo hemos resuelto mediante la construcción de diccionarios de época y el uso de programas de corrección ortográfica abiertos a la incorporación de diccionarios del usuario [24].

Después de la digitalización de las imágenes, éstas se pasan al Área de Edición Digital donde se procede al montaje de las mismas, diseño gráfico de marcos, índices y botones de navegación, y en algunos casos se aplican procesos de foliación automática. Para ello se usan programas comerciales como Dreamweaver 4 y programas de producción propia como FacsBuilder.

7 Supervisión

La calidad de nuestros textos es la mejor hoja de presentación de nuestra biblioteca, por ello, el Área de Supervisión y Control de Calidad sigue una metodología

[7] Ispell en un programa para corrección ortográfica con varios años desde su creación. El original fue estricto en lenguaje ensamblador del ordenador PDP-10 en 1971, por R. E. Gorin. La versión C fue escrita por Pace Willisson del MIT y Walt Buehring de Texas Instruments agregó la interfaz para Emacs y lo colocó en la red. red. Hay versiones de este programa de corrección ortográfica para la mayoría de los sistemas operativos actuales.

basada en la revisión de los puntos críticos en cualquier edición y las calas continuas en los textos. Las obras llegan a esta sección después de que el corrector haya acabado de corregir y marcar en XML la obra y se haya hecho una copia de seguridad. Al finalizar su trabajo, el Área de Supervisión proporciona al Área de Corrección una serie de listados que sirven de retroalimentación de los errores.

8 Generación automática de diferentes formatos

Una vez finalizada la supervisión, obtenemos un fichero XML-TEI único por cada obra. Entramos entonces en la fase de generación automática de los diferentes formatos y servicios que finalmente publicaremos en la web: HTML, PDF y Concordancias TACT. Arms describe este tipo de proceso de conversión de forma simple en el capítulo 9, pg. 163-166 de su libro [25]. Nuestra versión del proceso de transformación se muestra en la parte de abajo de la figura 2.

8.1 Hojas de estilo XSL y MakeBook

Los libros digitales en formato HTML se obtienen a partir de una doble transformación de los ficheros XML-TEI. Primero una transformación XSL genera un fichero HTML único (con marcas de formato especiales embebidas) por cada fichero XML-TEI. Luego un parser llamado MakeBook convierte ese fichero en un libro digital, generando un fichero para la tabla de contenidos y el índice alfabético de primeras líneas de verso (éste último sólo en el caso de libros de poesía), y un fichero por cada parte (capítulos, artículos, etc.) de la obra. También se extraen las notas a pie de página y se las coloca en pequeños ficheros externos, dejando en su lugar hiperenlaces a las mismas. MakeBook fue pensado también para agregar las cabeceras y pies de página de cada parte del libro, con los botones e hiperenlaces de navegación que permiten recorrer las partes, volver al índice, y subir y bajar de título en título con un click de ratón. El resultado es un libro electrónico, que no es otra cosa que un conjunto de páginas web enlazadas con una topología anillo-estrella: un anillo bidireccional de enlaces que recorren los capítulos, más una página índice central con enlaces bidireccionales a cada capítulo (ver figura 4).

Tanto las hojas de estilo XSLT como MakeBook fueron desarrollados en la Biblioteca Virtual. MakeBook toma como entrada el fichero HTML con marcas de formato que genera la transformación XSLT. Además, se usa un juego de plantillas para dar un formato y funcionalidad específicos a los archivos HTML generados. El propósito de usar plantillas es dar una apariencia uniforme a todos los libros de la biblioteca, así como facilitar el mantenimiento, al permitir que los cambios de formato se apliquen fácil y uniformemente a toda la colección de libros (por ejemplo, títulos de la página HTML, colores de fondo y botones de navegación pueden cambiarse desde las plantillas).

Cualquier cambio en las plantillas hace que sea necesario reprocesar todos los ficheros HTML para que los cambios se reflejen en todos los libros de la biblioteca. Si bien este proceso es automático llegó un momento en que debido a

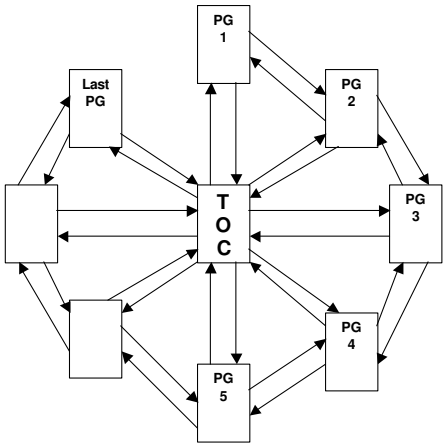

Fig. 4. Topología anillo-estrella de los enlaces de un libro digital.

los miles de libros que posee la biblioteca se hizo bastante lento, por lo que hoy en día se ha transferido parte de las funciones de las plantillas a programas del servidor (Java Servlets) que agregan las cabeceras y pies de archivo al momento de servir los ficheros HTML. De este modo, un cambio en estos elementos no implica reprocesar todo el lote sino que afecta instantáneamente a los próximos libros a servir.

Como ejemplo de cómo pueden crearse formatos diferentes a partir de libros XML similares pero de colecciones distintas vea los libros de la Biblioteca Virtual Miguel de Cervantes (http://cervantesvirtual.com/) y compárelos con los de la Joan Lluís Vives (http://lluisvives.com/). Ambos son portales de la misma biblioteca (mismo servidor y mismo proceso de producción), el primero en español, y el segundo en catalán y con un formato de presentación diferente.

Por otro lado también usamos transformaciones XSL para generar libros en formato PDF. Para procesar las transformaciones XSL usamos el parser Saxon que ofrece extensiones propias a la norma XSLT 1.0 de gran utilidad práctica, aunque existen muchos otros buenos parsers XSLT (como el XT de James Clark).

8.2 Conversión a TACT

Por último, hemos construido un parser que permite generar concordancias basadas en TACT [26] y TACTweb [27]. Este es un servicio interactivo que posee la biblioteca pensado para ser usado por estudiantes de literatura e investigadores. TACTweb ofrece este servicio interactivo cliente-servidor por Internet, por medio de formularios HTML del lado del cliente y un programa CGI (TACTweb) que procesa las consultas del lado del servidor. TACTweb trabaja con concordancias generadas por TACT, siendo éste un programa MS-DOS muy usado por filólogos y lingüistas para el análisis de textos y concordancias. Ambos

programas fueron desarrollaron en la Universidad de Toronto, Canadá. Las etiquetas usadas por TACT son similares a las usadas en el marcado estilo COCOA. Nosotros hemos desarrollado un parser (XML2TACT) que automáticamente convierte nuestros textos XML-TEI al formato COCOA que usa el programa MakeBase de TACT como entrada para generar una concordancia. De esta manera, la generación de concordancias TACT a partir de nuestros textos XML se realiza de forma totalmente automática.

También hemos desarrollado nuestro propio programa de concordancias para generar concordancias en CD-ROM [28]. En este caso el modelo cliente-servidor no era posible ya que necesitábamos un servicio de concordancias que estuviera basado sólo en archivos HTML estáticos grabados en CD-ROM. CD-Concord ofrece una interfaz HTML basada en frames dónde el usuario puede seleccionar palabras aisladas y ver sus usos en el texto seleccionado. Comparado con las concordancias TACT, este servicio carece de búsquedas por expresiones regulares, pero proporciona un servicio muy útil e independiente de la web para estudiantes e investigadores.

References

1. Bia, A.: The use of multimediality to enhance the accessibility to digital library resources: The multicultural-scope of the services offered by the Miguel de Cervantes digital library project. In: Digital Resources for the Humanities 2001 (DRH2001), University of London, London, England (2001) 9–14
2. Bia, A.: Automating the Workflow of the Miguel de Cervantes Digital Library. In: ACM 2000 Digital Libraries conference (Fifth ACM Conference on Digital Libraries), Menger Hotel, San Antonio, Texas, USA (2000) (presented as poster).
3. Bia, A., Sánchez-Quero, M.: Diseño de un procedimiento de marcado para la automatización del procesamiento de textos digitales usando XML y TEI. In De-la-Fuente, P., Pérez, A., eds.: JBIDI 2001, II Jornadas Bibliotecas Digitales, Almagro (Ciudad Real), Spain (2001) 153–165
4. Bia, A., Sánchez-Quero, M.: Marcado de textos literarios en XML-TEI en la Biblioteca Virtual Miguel de Cervantes. Revista Interamericana de Nuevas Tecnologías de la Información **6** (2001)
5. Bia, A.: A Versatile Facsimile and Transcription Service for Manuscripts and Rare Old Books at the Miguel de Cervantes Digital Library. In: JCDL'01: Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, USA (2001) 477
6. Societat Catalana de Documentació i Informació: Formatos bibliográficos, su compatibilidad y conversión: casos de usuarios de Sistemas Automatizados de Bibliotecas. Socadi, Barcelona (1992)
7. Bia, A.: Technical Aspects of the Production Process of Digital Books Using XML-TEI at the Miguel de Cervantes Digital Library. In: ACH/ALLC 2001. The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, The 2001 Joint International Conference, New York University, New York City (2001) 142–143
8. Bia, A.: Automating the Production of Facsimiles and Transcriptions for Manuscripts and Rare Old Books at the Miguel de Cervantes Digital Library.

In: Digital Resources for the Humanities 2001 (DRH2001), University of London, London, England (2001) 155–156 (presented as poster/demo).
9. Estévez-Ballester, A.: Formato USMARC: versión 1. Servicio Central de Bibliotecas de la Universidad de Cádiz, Cádiz (1999)
10. Biblioteca Nacional de España: Formato IBERMARC para registros bibliográficos. Biblioteca Nacional, Madrid (1996)
11. Sperberg-McQueen, C.M., Burnard, L., eds.: Guidelines for Electronic Text Encoding and Interchange (Text Encoding Initiative P3), Revised Reprint, Oxford, May 1999. TEI P3 Text Encoding Initiative, Chicago - Oxford (1994)
12. Burnard, L., Robinson, P.: Vers un standard européen de description des manuscrits: le project Master. In André, J., Chabin, M.A., eds.: Les documents anciens. Volume 3 of Document numérique. Hermes Science Publications, Paris (1999) 151–169
13. DCMI: Dublin Core Metadata Initiative home page. http://dublincore.org/ (Last visited: September 2002)
14. Bia, A., Nieto, A.: Information Retrieval in Digital Libraries: efficient catalog searches using tries. http://cervantesvirtual.com/research/articles/tries.pdf (2000)
15. DeRose, S.: XML and the TEI. In Mylonas, E., Renear, A., eds.: Text Encoding Initiative: Anniversary conference; 10th — November 1997, Providence, RI. Volume 33(1) of Computers and the Humanities 1999; /2., Norwell, MA, USA, and Dordrecht, The Netherlands, Kluwer Academic Publishers Group (1999) 11–30
16. Bia, A., Muñoz, R.: Aplicación de Técnicas de Extracción de Información a Bibliotecas Digitales (Applying Information Extraction Techniques to DLs). In Ferro, M.V., ed.: Proceedings of the XVI Conference of the SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural). Volume 26., University of Vigo, Spain, SEPLN (2000) 207–214 (published in: Procesamiento del Lenguaje Natural, journal of the SEPLN).
17. Bia, A., Carrasco, R.C.: Automatic DTD simplification by examples. In: ACH/ALLC 2001. The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, The 2001 Joint International Conference, New York University, New York City (2001) 7–9
18. Arellano, I., Cañedo, J., eds.: Crítica textual y anotación filológica en obras del siglo de oro: actas del Seminario Internacional para la edición y anotación de textos del Siglo de Oro, Universidad de Navarra, Pamplona. Castalia, Madrid (1991)
19. Blecua, A.: Manual de crítica textual. Castalia, Madrid (1990)
20. de Covarrubias, S.: Tesoro de la lengua castellana o española. 4a edn. Editorial Alta Fulla, Barcelona (1989) Esta edición reproduce la que Martín de Riquer, de la Real Academia Española, publicó en 1943, según la original de 1611, regularizando la puntuación y la ortografía, y añadiendo unos utilísimos índices.
21. López, M., Rodríguez, S., Carretero, J.: Herramientas ortográficas de libre distribución para la lengua castellana. In: II Congreso Hispalinux. (1999)
22. Bia, A., Sánchez-Quero, M.: Building Spell-Checking Facilities for Ancient Spanish. In: ACH/ALLC 2001. The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, The 2001 Joint International Conference, New York University, New York City (2001) 9–11
23. Bia, A., Sánchez-Quero, M.: Creación de diccionarios de castellano antiguo para la revisión ortográfica de textos y estudio estadístico de la evolución del uso de las palabras. In: XVII Encuentro Internacional de la Asociación de Jóvenes Lingüístas (AJL), Universidad de Alicante, España (2002)

24. Bia, A., Sánchez-Quero, M.: Building ancient Spanish dictionaries for spell-checking of DL texts. In González-Rodríguez, M., Suárez-Araujo, C.P., eds.: LREC 2002, Third International Conference on Language Resources and Evaluation. Volume VI., Las Palmas de Gran Canaria, Spain (2002) 1832–1837
25. Arms, W.: Digital Libraries. MIT Press, Cambridge, Massachusetts (2000)
26. Bradley, J.: TACT Design. In Wooldridge, T.R., ed.: A TACT Exemplar. Volume 1 of CCH Working Papers. Centre for Computing in the Humanities, Toronto (1991) 7–14
27. Bradley, J., Rockwell, G., Stevens, A.: TACTweb, an experimental software to access TACT databases through Internet. http://tactweb.humanities.mcmaster.ca/index.htm (1997)
28. San-Martín, I., Pomares, J.: CD-Concord: a concordance generator based on HTML and JAVA. Technical Report 2001-03, Miguel de Cervantes Digital Library, University of Alicante, Alicante, Spain (2001)