LEXIS. Vol. XVII. Nº 1. 1993

LA BIBLIOTECA ELECTRONICA EN EL ARCHIVO DIGITAL DE MANUSCRITOS Y TEXTOS ESPAÑOLES

Francisco Marcos Marín
Universidad Autónoma de Madrid

PRESENTACION

La conmemoración del Quinto Centenario del Descubrimiento y Colonización de América por España ha sido planteada con una perspectiva de futuro, que, en el plano lingüístico, requiere una acción conjunta sobre nuestra lengua común, tanto en su realidad presente como en su historia, al mismo tiempo que un tratamiento adecuado a la realidad tecnológica del inminente siglo XXI. La solución pasa por la creación de unos instrumentos de base que son los archivos digitales, colecciones de textos completos en formato electrónico, para poder recuperar también electrónicamente la información que contienen.

TRES TIPOS BASICOS DE ARCHIVOS DIGITALES

En nuestra exposición vamos a introducir una terminología también convencional. En ella, limitaremos el término general de *archivos digitales* a los que se constituyen con la intención de almacenar un corpus de *textos e imagen*, mientras que utilizaremos la palabra *corpus* para los archivos de texto sólo, sin imagen. Dentro de este segundo grupo tendremos que distinguir dos grandes grupos, el *corpus oral* y el *corpus escrito*.

Dentro de las actividades que ha llevado a cabo el área de Industrias de la Lengua de la Sociedad Estatal del Quinto Centenario, estas dos líneas tienen nombre propio: *ADMYTE*, el *Archivo Digital de Manuscritos y Textos*

Españoles, es un archivo digital en el sentido restringido que acabamos de proponer, mientras que el *Corpus de Referencia del Español Contemporáneo,* que incluye un corpus oral y un corpus escrito, es un ejemplo del segundo grupo de archivos, en sentido amplio.

DESCRIPCION GENERAL DE ADMYTE

La preservación del patrimonio cultural es una de las principales preocupaciones de los pueblos modernos. Los libros, especialmente los más antiguos, manuscritos e incunables, por su rareza y por las vicisitudes que han sufrido en su existencia, están expuestos a graves peligros. España, a lo largo de su historia, ha perdido incluso bibliotecas magníficas, como la de Hernando Colón, que hoy sería sin duda la mejor biblioteca románica medieval del mundo y que se fue destruyendo simplemente por incuria e ignorancia.

Pero no basta con conservar, también es necesario que esas obras cumplan su función al servicio de los lectores, del público culto interesado, en general. Para ello sería preciso ponerlas a disposición de éste, lo que inevitablemente acarrearía su deterioro y hasta su destrucción, lentamente.

Por fortuna, la tecnología actual nos permite cumplir el objetivo de poner a disposición de los estudiosos de distintos campos toda la literatura castellana medieval, entendiendo de modo amplio el término "literatura", como colección de los textos escritos, e incluso añadir a ello unos instrumentos de trabajo, que ya existen o están en muy avanzado desarrollo, con los cuales los investigadores podrán realizar su labor en condiciones de total seguridad para las obras que estudian y de máximo rigor científico, junto con una comodidad de que han carecido hasta ahora.

Para ello, la Sociedad Estatal para la Ejecución de Programas del Quinto Centenario, en el Area de Industrias de la Lengua, y la empresa MICRONET, S.A., de amplia y reconocida experiencia en el campo del almacenamiento y recuperación de información en CD-ROM o disco láser (término que iremos introduciendo a partir de este momento), con la colaboración de las universidades Autónoma y Complutense de Madrid (España), de California en Berkeley (EE.UU.), de Madison (Wisconsin, EE.UU.) y de Toronto (Canadá), y de distintas bibliotecas y entidades públicas y privadas, iniciaron conjuntamente el proyecto ADMYTE, Archivo Digital de Manuscritos y Textos Españoles.

El disco CD-ROM, siglas que corresponden a "Compact Disk, Read Only Memory", es decir, disco compacto de sólo lectura, es exactamente igual que los discos de música, sólo que en vez de contener sonido contiene textos e imágenes. La máquina lectora, el lector de CD-ROM, se acopla a un ordenador y puede usarse también, con altavoces, para reproducir discos compactos de música. Para hacernos una idea, en un disco que pesa trece gramos y tiene una duración en principio indefinida caben unas nueve mil páginas de texto e imagen.

ADMYTE, que se presenta como una colección de discos láser de tipo CD-ROM, en la versión más moderna que la técnica puede ofrecer en este tiempo, y que puede ser utilizado por los investigadores que dispongan de un sencillo ordenador personal, tipo MS-DOS, un monitor VGA y una lectora de discos láser, se divide en dos series de desigual extensión: la primera está constituida por un solo disco, el disco 0, o disco instrumental, destinado a los investigadores con requisitos más complejos, mientras que la segunda, de los discos 1...n, está destinada al público más amplio, que incluye, naturalmente, a los propios investigadores. He aquí la estructura general de la colección:

Volumen 0:

BETA/BOOST. Bibliografía Española de Textos Antiguos. Base bibliográfica con más de cuatrocientos cincuenta campos y once tablas, interactiva. (Colaboración de la Universidad de California en Berkeley y la Universidad Complutense de Madrid. Este proyecto cuenta con el apoyo del National Endowment for the Humanities y, parcialmente, de IBM USA y de IBM España).

TACT. Programa de recuperación de información textual con sistema de creación de la propia base de datos textual. (Colaboración con la Universidad de Toronto de Canadá. Este proyecto cuenta con el apoyo del Centre for Computing in the Humanities, CCH, de la Universidad de Toronto, y, parcialmente, de IBM Canada Ltd.)

TEXTOS-MAD. Colección de Textos Medievales, cedidos por el Hispanic Seminary of Medieval Studies. (Colaboración con la Universidad de Wisconsin en Madison, proyecto que se realiza con la ayuda del National Endowment for the Humanities).

UNITE. Conjunto de programas para la construcción de ediciones críticas automatizadas. (Colaboración con la Universidad Autónoma de Madrid.

Este proyecto ha contado con el apoyo de IBM España, IBM Deutschland, EUROTRA-España y, especialmente, la Alexander von Humboldt Stiftung).

Volúmenes 1 ...n:

Textos transcritos con código ASCII, con marcas o membretes estandarizados.

Imágenes (facsímiles de textos en blanco y negro y reproducción en color de miniaturas).

Desarrollo de una interfaz específica para el medio Windows (3.0 y posterior), por MICRONET, S.A. para la recuperación de textos e imágenes de alta resolución, desde CD-ROM.

La propuesta de ADMYTE parte de esa necesidad primera de conservación del patrimonio escrito y, además, de la conciencia de que se ha producido un cambio en los métodos de trabajo de los investigadores. El ordenador es ya un instrumento generalizado entre los humanistas e insustituible para algunas funciones, como el manejo y selección de la inabarcable bibliografía o de la mayor riqueza de datos disponibles. Al aumento del caudal informativo corresponde ineludiblemente el desarrollo de nuevas técnicas para abarcarlo y asimilarlo.

Por esta razón, los investigadores que realizan ADMYTE son conscientes de que es el momento de desarrollar sus investigaciones complementarias, hasta ahora parciales, en busca de una síntesis que reúna experiencia y conocimientos para constituir un nuevo tipo de *vademecum* de este momento tecnológico, un vademécum informatizado para el investigador de la Edad Media y el Humanismo.

Los procedimientos de trabajo perfeccionados y reunidos no sólo sirven para ser aplicados a la lengua o la literatura del antiguo reino de Castilla y León, y su expansión peninsular, sino para las distintas lenguas, no sólo hispánicas o románicas, sino de cualquier lugar del mundo. Lo mismo debe decirse de los procedimientos técnicos desarrollados para la digitalización de manuscritos o la transcripción automática de incunables: están a la disposición de investigadores de cualquier lengua y cualquier país o etapa histórica.

Todo ello nos lleva a un proyecto basado en la lengua española, pero realizado por un equipo internacional e interdisciplinar, en la medida de lo

necesario. Un proyecto que, justo es resaltarlo, se ha llevado a cabo por primera vez en la historia. Los investigadores participantes tienen entre sí, a veces, diferencias significativas; pero en lo que afecta a esta investigación, todos convergen en un mismo punto o, mejor, en tres puntos que resumen sendas series de resultados:

1. Completar y desarrollar instrumentos concretos de investigación que constituyen un sistema experto de tratamiento textual y recuperación de información contenida en todo tipo de textos.

2. Presentación de un modelo aplicable, en principio, al español medieval; pero expandible a otras épocas y otras lenguas con suma facilidad.

3. Recuperación del patrimonio cultural español a través de la localización, catalogación, preservación y estudio de una parcela fundamental del pasado histórico como son los textos medievales y sus soportes, manuscritos e incunables.

Es preciso señalar las consecuencias de un proyecto de tal envergadura. Para empezar, podemos llevar el estudio de la cultura española a cualquier lugar del mundo, lejos de las grandes bibliotecas, por el módico precio de un disco láser. En segundo lugar, cabe recordar que la inclusión de facsímiles limitará sobremanera el uso de los originales, con el consiguiente beneficio que esto supondrá para una mejor conservación de los mismos. En tercer lugar, debe notarse que España será precursora en el uso de una tecnología que no tardarán en aplicar los estudiosos de otras lenguas y períodos históricos. Finalmente, las técnicas de tratamiento de imágenes, desarrolladas originariamente para mejorar las fotografías obtenidas desde los satélites espaciales, permitirán que el estudioso "restaure" los manuscritos de modo electrónico y que, así, podamos recuperar un enorme conjunto de textos que hoy consideramos ilegibles.

DIGITALIZACION Y TRATAMIENTO DE IMAGENES: MANUSCRITOS E INCUNABLES

Las técnicas que empezaron a ser desarrolladas por los técnicos de la Agencia Espacial norteamericana, la NASA, en los años setenta, comercializadas poco después, y que permiten la conversión de una imagen en un patrón digital, se aplican ya desde hace tiempo a los estudios de ciencias naturales y biológicas y han pasado a los humanísticos.

Con el crecimiento de los últimos años en las capacidades de memoria de los ordenadores y el aumento de las posibilidades de almacenar estos grandes ficheros que resultan de la digitalización, esta técnica se ha ido acercando al usuario, hasta que el disco óptico ha permitido un abaratamiento definitivo de los costos y la posibilidad de que los individuos se beneficien de ello en la investigación personal.

La digitalización de los manuscritos e incunables castellanos reúne dos ventajas esenciales: preserva el patrimonio bibliográfico español y pone a disposición de los estudiosos reproducciones que, debido a las posibilidades de la electrónica, ofrecen imágenes más nítidas y facilitan por ello el trabajo de investigador.

La colaboración de la Biblioteca Nacional en este proceso resulta fundamental, porque garantiza que se realiza con la máxima seguridad y sin daño alguno para los ejemplares digitalizados.

ADMYTE incorpora los textos digitalizados, que el estudioso puede reproducir cómodamente mediante una simple impresora láser estándar, con lo que se convierte en el primer elenco completo de una época para una lengua moderna.

Es conveniente señalar que la digitalización puede realizarse a través de imágenes fotográficas, lo que además permite mejorar las condiciones de lectura de los manuscritos, mediante el uso de los auxiliares más oportunos (luz ultravioleta e infrarroja, por ejemplo). La primera parte de este servicio no es desconocida de la Biblioteca Nacional y tanto la Bibliothèque Nationale de París como la British Library poseen experiencia más que suficiente sobre estos procedimientos. La novedad es la posibilidad de ofrecer una reproducción digitalizada, además de las formas habituales de fotocopia y microfilme, la primera de las cuales es notoriamente dañina para los libros o documentos cuando se aplica directamente y será sustituida por la reproducción láser, que asegura una mayor calidad.

Obtención de positivos fotográficos

La naturaleza y estado actual de los textos originales que es necesario digitalizar ha obligado a trabajar con una técnica mixta, uniendo procesos fotográficos y electrónicos. Dada la extrema delicadeza de algunos originales, ha sido necesario preparar un sistema para poder obtener la máxima informa-

ción posible de cada página de cada documento, con el manipulado más corto posible, sin contacto físico y sin exposición a temperaturas elevadas o radiaciones perjudiciales.

Después de muchas pruebas y ensayos con distintos materiales y distintos sistemas de digitalización, se llegó a la conclusión de que era necesario realizar un paso previo a la digitalización de las páginas: la obtención de originales fotográficos. Sólo cuando las características de los textos que se iban a digitalizar, por su tamaño, generalmente, impedía hacerlo desde diapositiva, se ha procedido al escaneado directo, a través de un escáner adaptado a las características del libro antiguo, para no dañarlo.

El paso a imágenes digitalizadas a partir de soportes en celuloide no es novedad: se aplica ya con éxito en historia del arte y en documentación en general. El desarrollo de la técnica para esta aplicación concreta de biblioteconomía supone un avance notable en el servicio de bibliotecas, al conseguirse, entre otras, las siguientes ventajas:

1. No existe contacto físico que pueda dañar el documento.

2. El documento está fuera de su ubicación habitual el tiempo más corto posible, facilitándose así la custodia adecuada de todos los originales.

3. Los documentos se exponen a la luz durante un período de tiempo extremadamente corto (aproximadamente 15 segundos por página).

4. No hay cambios bruscos en la temperatura o humedad de conservación.

5. No es necesario forzar la encuadernación, ya que no es preciso abrir totalmente los volúmenes.

6. Los procesos de digitalización se realizan sobre imágenes fotográficas, lo que nos permite la obtención de las imágenes electrónicas digitales posteriormente y sin presencia de los originales.

7. Es posible realizar exploraciones múltiples digitales y diversos procesos en páginas muy degradadas sin volver a procesar las páginas originales.

8. Se obtiene un respaldo complementario de los documentos, ya que proporciona un archivo fotográfico en color, de gran calidad, que puede ser utilizado por la Biblioteca Nacional para otros trabajos posteriores.

9. Es posible aumentar la legibilidad de algunos documentos muy deteriorados, ya que mediante combinaciones de luz, filtros y emulsiones sensibilizadas de forma especial, se pueden obtener resultados mejores que con otras técnicas.

Después de decidir la técnica que se iba a utilizar, hemos llevado a cabo un proceso de selección del material fotográfico, que nos ha hecho utilizar una película reversible de color en formato 24 x 36 mm., de grano ultrafino, y una resolución y rendimiento a los colores muy superiores a los requeridos en los procesos posteriores de digitalización. Las cámaras utilizadas son capaces de garantizar la exposición correcta de cada página de forma automática, mediante un sistema de medición de luz a través del objetivo capaz de evaluar diferentes zonas del encuadre seleccionado. El arrastre de la película es automático y se puede realizar un enfoque automático de cada página antes de cada toma.

También ha sido necesario preparar útiles especiales para sostener todo el material necesario mientras trabaja simultáneamente, con los elementos de iluminación necesarios y los soportes de los originales. Para ello se construyó un atril especial capaz de soportar las dos cámaras y todos los elementos básicos para asegurar la realización de las diapositivas en condiciones óptimas. De esta forma se pueden fotografiar al mismo tiempo las dos páginas de un libro abierto, en el mínimo tiempo posible y sin forzar la encuadernación del mismo.

Se ha diseñado un proceso de clasificación de originales (una vez reveladas las diapositivas y comprobada su calidad) para poder realizar seguidamente los procesos de digitalización.

Digitalización de imágenes

La digitalización de las imágenes fotográficas de las páginas de los libros constituye el puente entre los procesos fotográficos e informáticos utilizados en el proyecto ADMYTE.

Para realizar correctamente el paso de digitalización se ha recurrido a digitalizadores o "scanners" de diapositivas de alta calidad, de muy elevada resolución (4.096 puntos por pulgada, 3.850 x 5.800 puntos en cada diapositiva) y capaces de diferenciar 16.777.216 colores (24 bit por punto, 8 por cada color RGB). Suponiendo que la imagen que se va a digitalizar ocupara

toda la superficie de la diapositiva y trabajando a la mitad de la resolución máxima del digitalizador, obtendríamos una imagen de 1.925 x 2.900 puntos, con 3 octetos o bytes (24 bit) por punto, lo que supone ocupar 16'75 Mb. de información por cada página. Aunque los requerimientos finales del proyecto no incluían imágenes de tan elevada resolución, ha parecido conveniente almacenar las imágenes originales con esta calidad por varias razones:

* Permite realizar procesos de corrección de color y reducción de la imagen sin pérdidas apreciables de calidad.

* Proporciona un almacenamiento inalterable en disco óptico que permite la realización posterior de otros proyectos y trabajos de investigación.

* Hace posible crear una base de datos con imágenes de alta definición en la Biblioteca Nacional sin volver a "tocar" los textos originales.

Para realizar el almacenamiento de estas imágenes ha sido necesario desarrollar nuevos algoritmos de compresión de imágenes en color, basados en procesos matemáticos que utilizan la transformada rápida de Fourier. Estos sistemas especiales de compresión, sin pérdida de calidad, han permitido reducir la ocupación de cada imagen en color a sólo 800.000 octetos. Los procesos de compresión son imprescindibles para construir un archivo manejable en disco óptico WORM, ya que de esta forma podemos almacenar 1.100 imágenes en cada disco de trabajo de 940 Mb., junto con la información necesaria para su localización y datos relativos a la diapositiva original.

Procesos de tratamiento de imágenes

La mayor parte de las imágenes que se incluirán en los discos compactos (CD-ROM) serán en blanco y negro, por lo que podemos realizar procesos especiales de tratamiento encaminados a aumentar su legibilidad. La digitalización en color nos permite modificar determinados tonos, realizar procesos digitales de filtrado, etc., previos a la reducción de las imágenes a blanco y negro.

Los sistemas de proceso de imágenes empleados permiten hacer más legibles los documentos eliminando o reduciendo las manchas de humedad, el tono amarillento, las agresiones naturales, etc.

El estado de los distintos documentos hace que aproximadamente el 95 por ciento de los mismos se pueda tratar completamente con procesos globa-

les, que afectan a todo el documento, mediante programas preparados especialmente para este fin. El 5 por ciento restante debe tratarse con procesos electrónicos manuales y, a veces seleccionando únicamente la zona afectada del documento, con un tiempo de ocupación de personal muy calificado elevadísimo.

Los procesos de tratamiento de imágenes incluyen sistemas estadísticos de reducción de colores a su valor medio, de control zonal de tonos, de sustitución de colores y puntos, de realce de contornos y contrastes, etc., con los que se obtienen resultados espectaculares al aclarar los fondos y oscurecer las tintas.

Reducción y cambio de colores a blanco y negro

Los procesos de las imágenes concluyen con su paso a blanco y negro y su reducción al equivalente a 150 puntos por pulgada (lo que permite obtener una copia impresa de calidad). Durante este proceso es necesario realizar una conversión de los colores según unas ciertas normas, imprescindibles para mantener la legibilidad en los textos no escritos con tintas negras (rojo, azul, etc.) y en las ilustraciones.

Durante las comprobaciones que se llevan a cabo con los ficheros resultantes se realizan copias en papel mediante impresora de tecnología láser. Una vez corroborada su calidad, se procede a su compresión según normas CCITT Grupo IV y su almacenado, con una ocupación media de 39 Kb.

Tratamiento de las iluminaciones e ilustraciones

En algunas páginas existen ilustraciones en color o iluminaciones con distinto grado de detalle. La reproducción adecuada de estas páginas requiere tratamientos especiales, puesto que se han de incluir en los discos CD-ROM conservando el color original. En la mayor parte de los casos únicamente se trabaja con la zona que contiene la ilustración (de forma ampliada), con lo que se aumenta la posibilidad de apreciar fácilmente cada detalle. Se han creado para el tratamiento de estas imágenes en color programas especiales que permiten realizar los procesos de corrección de forma interactiva, es decir, viendo directamente en la pantalla los resultados obtenidos con las modificaciones.

Transcripción de los textos

La transcripción de los textos se realiza solamente en caso de imposibilidad de utilizar una transcripción existente, ya sea por no corresponder a la misma edición de la obra o bien por no poder obtener los permisos necesarios para su uso.

Realiza la transcripción un equipo de especialistas que utiliza herramientas informáticas adecuadas, las cuales permiten generar las grafías especiales necesarias para poder realizar una transcripción paleográfica correcta.

En esta fase del proyecto se utiliza la impresora láser para obtener copias de los documentos que se han procesado siguiendo los pasos anteriores y que sirven además para realizar en el texto las indicaciones que se utilizarán más adelante, en el producto resultante del proyecto, para marcar el comienzo y fin de cada página.

La transcripción de los textos se beneficia, además de la cooperación de todos los investigadores que deseen ceder sus transcripciones en forma electrónica, de la posibilidad de leer ópticamente textos editados y textos mecanografiados, con ello reducimos notablemente la necesidad de teclear los textos para la introducción de los datos en forma electrónica y mejora sensiblemente la calidad de los textos ASCII incluidos en los discos, por la corrección de las ediciones o transcripciones mecánicas leídas electrónicamente. Es innegable que la calidad del texto transcrito es un requisito esencial en ADMYTE.

ANTECEDENTES

Este apartado está dedicado a exponer con un mínimo detalle de investigaciones previas sobre los distintos aspectos lingüísticos, filológicos y documentales que abarca el Archivo Digital, especialmente las que han sido realizadas por los participantes. Este material de trabajo para distintos especialistas (bibliotecarios, filólogos, editores, documentalistas, lingüísticas) se presenta en el disco 0, aunque puede aplicarse a los contenidos de cualquier disco. Se advertirá que, en la mayor parte de los casos, los proyectos de mayor envergadura realizados hasta la fecha corresponden precisamente a los investigadores reunidos en el cuerpo redactor de ADMYTE.

BETA: Bibliografía de Textos Españoles

Una de las labores más urgentes que debe acometer un país con una rica tradición cultural es la de catalogar y preservar los testimonios que reflejan su pasado; ésta es también una de las tareas principales para los especialistas en la Edad Media española. Nuestra experiencia en el campo de los manuscritos medievales y de los libros incunables nos ha llevado a recordar una y otra vez la necesidad de esta labor: desconocemos la existencia de numerosos libros españoles medievales custodiados en bibliotecas públicas y privadas de todo el mundo (hemos de recordar que la mayor parte de las bibliotecas españolas, y entre ellas las más importantes, como la Nacional, Palacio, Real Academia Española, Colombina de Sevilla o Universitaria de Salamanca, no disponen aún de catálogos completos); muchos de nuestros manuscritos sufren un fuerte deterioro y están amenazados por diversos peligros (es especialmente grave el problema de las tintas corrosivas, aunque no deben olvidarse otros elementos: el propio paso de tiempo, incendios, inundaciones, mutilaciones, aplicación de reactivos, polillas u hongos); finalmente, es preciso señalar que, a pesar de la legislación en defensa del patrimonio nacional, son muchos los códices e impresos que escapan al control del Estado por desconocer su existencia (disponemos de abundantes y escandalosos ejemplos dentro y fuera de España). BETA, o BETA/BOOST, siglas inglesas originales de la *Bibliography of Old Spanish Texts* pretende solucionar gran parte de estos problemas.

BETA/BOOST es un catálogo general de fuentes primarias —impresas y manuscritas— de textos españoles medievales escritos en castellano o en cualquiera de sus dialectos. Por ahora, nuestro campo de acción se ha limitado fundamentalmente a textos literarios, sin dejar de lado las obras de carácter histórico, legal, científico o religioso, ya que el hombre de la Edad Media las consideraba también literatura (además, se trata de un enorme corpus de gran importancia para filólogos e historiadores). Por el momento, sólo hemos prescindido del documento puro (de carácter notarial, generalmente), cuya inclusión habría requerido un esfuerzo adicional que superaría con mucho nuestras fuerzas (un catálogo general de documentos medievales resulta imposible de hacer en este momento, pues constaría de muchos cientos de miles de entradas recogidas en innumerables centros).

DOSL: DICTIONARY OF THE OLD SPANISH LANGUAGE

Entre las numerosas universidades norteamericanas con Departamento de Lengua Española, la Universidad de Wisconsin, en Madison, ha destacado

por la presencia de algunos de los más célebres especialistas en literatura medieval peninsular; entre ellos, cabe señalar a Antonio Solalinde o Américo Castro. En Madison, en 1931, Solalinde fundó el Seminary of Medieval Spanish Studies, institución que sigue atrayendo a hispanistas de todo el mundo, que tienen en Madison una cita obligada. El motivo que atrae al estudioso es el enorme volumen de información de que dispone el Seminario por los grandes proyectos de que se ocupa: entre éstos, el más ambicioso es, sin duda, el diccionario de español medieval en el que se trabaja desde hace cincuenta años, el *Dictionary of the Old Spanish Language,* también denominado *DOSL/ DEA.*

Una de las tareas básicas fue la de localizar las fuentes primarias de las que se iba a partir. Esta labor resultaba especialmente difícil por carecer de un catálogo general que brindase datos exactos sobre los códices e impresos conservados, su localización y su contenido. La *Bibliography of Old Spanish Texts* nació en 1975 con el propósito de colmar esta laguna, aunque su gran utilidad para el conocimiento de los textos castellanos medievales la convirtió muy pronto en algo más que una herramienta del DOSL: En la actualidad, BETA/BOOST es el primer libro al que acude cualquier estudioso que precisa un conocimiento directo de la Edad Media a través de sus fuentes primarias.

El Seminario ha tenido también otra influencia notable, la de creación de un estándar para la transcripción: los textos se transcriben según el libro preparado por David Mackenzie, *A Manual of Manuscript Transcription of the Dictionary of the Old Spanish Language* [1986: 4ª ed.], y las distintas entradas se crean de acuerdo con el patrón marcado en Victoria Burrus, *A Procedural Manual for Entry Establishment in the Dictionary of the Old Spanish Language* [1986: 3ª ed.].

UNITE

UNITE es un conjunto de programas que comparan diferentes versiones de un mismo texto, con el objetivo de obtener una versión unificada a partir de las comparadas. UNITE no está planteado como la solución definitiva a los problemas de la crítica textual (extensión de los textos, diversidad de grafías, existencia de diferentes versiones, etc.), sino como una herramienta que libere al editor humano de labores rutinarias y centre su trabajo en aquellas fases del proceso en las que sea indispensable la actuación del experto. Las versiones previas de UNITE, en lenguaje PASCAL, para ordenadores IBM

con sistema operativo V/CMS o para sistema UNIX, trataban un máximo de seis versiones de textos en verso. La versión para ADMYTE, escrita en C, para el sistema operativo MS-DOS, compara textos en verso de cualquier tipo y formato. Está en realización la versión MS-DOS para comparar textos en prosa.

La primera de las características de UNITE es que los textos no llevan ningún tipo de marca, etiqueta o membrete incorporados por el editor al texto que copia: no se requiere ninguna labor de pre-edición.

La segunda es que presenta una amplio abanico de posibilidades de automatización, con lo que el usuario puede delimitar perfectamente el campo de actuación del programa. Además, el sistema personaliza cada intervención, pidiendo un nombre al usuario y organizando todos los ficheros de variables según las instrucciones de ese usuario, lo que permite el trabajo de varios investigadores, cada uno con sus preferencias y, de acuerdo con ellas, sus ficheros de parámetros y de resultados claramente diferenciados e identificados.

El paquete estándar está diseñado para comparar diez versiones, aunque se incluyen también distintos programas ejecutables si se van a comparar más, sin más molestias para el usuario que modificar su fichero de parámetros. La unidad de comparación es la estrofa, entendida simplemente como un conjunto de versos separados por blancos. Este tipo de unidad permite detectar y solucionar el problema de los versos descolocados dentro de ella. No es necesario que las versiones tengan el mismo número de estrofas ni que éstas estén ordenadas por su numeración, ya que se incluyen utilidades para conformar y ordenar los textos. Tampoco es necesario que las estrofas tengan el mismo número de versos ni, a diferencia de versiones anteriores, que este número tenga más limitaciones que las de la capacidad del disco. El conjunto de programas está desarrollado en lenguaje C, y se presenta al usuario en un entorno integrado, cómodo y atractivo. El usuario puede personalizar todo el proceso desde la instalación, incluyendo la selección de su editor preferido. Todos los ficheros de personalización de UNITE pueden ser modificados a conveniencia del usuario siempre que se respeten las indicaciones que aparecen el principio de cada uno.

Salida de resultados

El proceso de unificación automática genera varios ficheros de resultados. El primero de ellos almacena la versión unificada, mientras que el se-

gundo registra una serie de datos referentes a los tres procesos vistos anteriormente y que resumen su ejecución.

El *fichero que almacena la versión unificada*, al igual que los textos comparados, está dividido en estrofas separadas cada una de ellas por una línea en blanco. Además de las palabras que fueron seleccionadas para la versión unificada, figuran también las variantes que no fueron seleccionadas, con un número asignado para identificar la versión original en la que aparecían. Opcionalmente también pueden aparecer las estrofas originarias de cada versión delante de la estrofa unificada.

El *fichero que resume la ejecución de los procesos de unificación* se crea opcionalmente. Esta creación está controlada por un parámetro modificable por el usuario. Su generación es muy conveniente cuando se necesita una aclaración de la ejecución de los citados procesos. Para cada palabra que generan dichos procesos existe una línea del fichero en la que se indican la estrofa y verso en los que figura, el proceso que la generó, la palabra unificada resultado del proceso y las palabras originales (acompañadas del identificativo de versión correspondiente) que generaron dicha palabra unificada; estas últimas aparecen en formato original y no con el que realmente trabajan los procesos de unificación. Es el fichero auxiliar para la elaboración posterior del aparato crítico.

La extracción del *fichero de variantes* permite, además, disponer de todas las variantes que se han producido al realizar el proceso de unificación, con objeto de tener un conocimiento, por separado, de lo que no ha sido unificado y poder disponer de estos datos a la hora de confeccionar el aparato crítico. Téngase en cuenta que, tanto en este caso como en el anterior, se dispone de estos ficheros, opcionalmente, más del fichero resultante de la versión unificada.

La opción de *trabajos con los textos* permite utilizar algunas instrucciones del sistema operativo o de otros programas comerciales para realizar búsquedas de cadenas y obtener información sobre formas léxicas o sintagmáticas que interesen, dentro de su verso. En cada caso se busca dentro de un fichero y se almacenan los resultados en un fichero común al que se van agregando los datos de las búsquedas subsiguientes, si así se desea. Se obtienen concordancias con indicación del texto donde se encuentra la cadena buscada, número de línea y supresión optativa de las diferencias de mayúsculas y minúsculas así como de los espacios en blanco entre palabras que

pueden ir unidas o separadas, caso de los clíticos y de otras formas. El número de opciones y combinaciones es elevado, por lo que vale la pena destacar simplemente la opción entre recoger la concordancia con el texto o sólo la indicación numérica de la línea en la que se encuentra, procedimiento éste más rápido y de interés para recuentos estadísticos. Dentro de ADMYTE, el programa asociado por excelencia es TACT.

TACT

TACT es un programa de recuperación de información textual producido por el Centre for Computing in the Humanities de la Universidad de Toronto (Canadá). El funcionamiento básico de TACT es simple: permite preguntar sobre la ubicación de palabras y sintagmas en un texto. Para ello, aunque trabaje sobre un texto, no lee un simple fichero textual, sino que trabaja con una base de datos textual que fabrica una de sus utilidades, MAKBAS. Para los textos transcritos por el Hispanic Seminary of Medieval Studies y los que, como ADMYTE, siguen estas normas de transcripción, se ha desarrollado una versión especial de MAKBAS, llamada HSMS2TDB.

El usuario interactúa con TACT a través de un entorno integrado en donde puede realizar una serie de operaciones, como ver el texto, extraer la lista de palabras del mismo, seleccionar sobre ella las palabras que quiere destacar para construir concordancias o índices referenciales de menor extensión, estudiar en su distribución por el texto o en su distribución relativa al contexto frente a la totalidad, con fines estadísticos.

Los textos pueden estar escritos en cualquier sistema de caracteres, pues TACT permite la redefinición del juego de caracteres mediante un fichero auxiliar (*XLATTABL.DAT*), o utilizando combinaciones de teclas, lo que permite una gran flexibilidad, especialmente en los sistemas de indexación y de búsqueda.

El sistema funciona, en general, a base de selecciones, con la ventaja de que para ello el usuario puede establecer sus caracteres de control y construir una base de datos para obtener información específica. Puede también personalizar sus preferencias, mediante la creación de reglas personalizadas o mediante la personalización de la base textual, que se puede mantener así en consultas posteriores. Permite amplios márgenes en la definición de contextos o en la selección, incluso desde la aplicación concreta, interactivamente. Podemos interrogar a la base de datos textual para saber no sólo en qué

contexto aparece una palabra, con la posibilidad de elegir ese contexto en líneas o palabras o por el personaje que habla o el capítulo en que aparece, sino hacer consultas más complejas, como el número de palabras del texto que acaban en *—aba*, o los lugares en los que un personaje habla de una palabra concreta o sus variantes formales, mediante el uso de símbolos comodines, o aquellos en donde una serie de palabras ocurren en el mismo párrafo. Mediante reglas podemos agrupar las formas y sus variantes, para construir nuestras propias listas de sinónimos o de alteraciones en el paradigma verbal, por ejemplo, y buscar según ellas. Las reglas pueden conservarse, importarse y exportarse. Las búsquedas permiten condiciones, afirmativas o negativas, combinaciones de requisitos y diversas clases de comodines.

Los resultados obtenidos pueden imprimirse, directamente, o bien guardarse en disco, en formato ASCII, sustituyendo el fichero anterior del mismo nombre, si existía, o añadiéndose a él, en apéndice, según se elija. Otra posibilidad de aprovechar el trabajo de sesiones anteriores es crear un fichero SCRIPT, que contiene un registro de operaciones realizadas que puede repetirse en cualquier sesión futura y que se ha individualizado con un nombre propio.

TACT va acompañado de una serie de programas que refuerzan su utilidad. COLLGEN, por ejemplo, revisa las palabras de un texto para encontrar todos los lugares en los que aparece una combinación de dos o más palabras más de una vez y escribe una lista que muestra qué combinaciones ocurren y con qué frecuencia. Tiene además la ventaja de que puede procesarse por lotes (en modo *batch*).

La mayor ventaja de TACT, sin duda, es la posibilidad de crear la propia base textual y definir todos los parámetros que vayan a necesitarse. La base textual contiene, además del texto, índices completos de todas las posiciones de todas las palabras del texto, así como información sobre la estructura formal: dónde empieza o acaba un capítulo en una novela o un libro científico o dónde habla o deja de hablar un personaje en una comedia. De ello se ocupa el programa MAKBAS, que admite todo tipo de informaciones sobre el texto y su etiquetado previo, los caracteres y signos diacríticos usados o aquellas partes que no deben ser tratadas como partes del texto, sino anotaciones, referencias o complementos, todo ello a partir de ficheros en formato ASCII. A veces, podemos intentar crear una base textual demasiado larga para MAKBAS o HSMS2TDB. TACT resuelve esta dificultad mediante MERGEGAS, un programa que permite combinar varias bases textuales en una base muy larga, soslayando así los inconvenientes de un sistema opera-

tivo con tan pocos recursos de memoria como MS-DOS. En su nueva versión, TACT permite hacer uso de la memoria ampliada.

ADMYTE y WINDOWS

ADMYTE ofrece al usuario la posibilidad de combinar un ordenador personal (con una pantalla VGA, como mínimo, por motivos de resolución), un lector de discos láser y una impresora láser para reproducciones inmediatas y seguras. A fin de combinar estos elementos y aprovechar toda la información contenida en ellos se ha desarrollado ADMYTE bajo Windows 3.0, con el fin de aprovechar las ventajas de un entorno estándar. La técnica de ventanas permite ver en columnas paralelas el documento en facsímil y su transcripción. La técnica de menúes desplegables desde una barra permite recuperar la información, bien de un título, de varios agrupados, o de una simple página.

La *búsqueda por palabras* permite seleccionar las páginas de los libros seleccionados en las que se encuentran las palabras que interesan en ese proceso de la investigación, mientras que la *selección por glosario* o la *búsqueda mediante lenguaje de interrogación* son procesos más complejos. Para la selección por glosario debemos imaginar el conjunto como una base de datos, alfabéticos y gráficos, a la que se ha asociado un glosario, construido por el equipo lingüístico e informático de ADMYTE. El *glosario* de la base de datos contiene todas las palabras de búsqueda, ordenadas alfabéticamente. Las referencias se han extraído de los campos índice de los diferentes documentos. La búsqueda se realiza tecleando la referencia en la línea de texto situada justamente encima del Glosario, tras lo cual se pulsa *intro*. Para usar el lenguaje de interrogación es preciso llamar el proceso *Buscar* y teclear la palabra o palabras que se desee hallar, teniendo en cuenta que el lenguaje tiene una rica sintaxis que admite diversas operaciones. Podemos buscar una referencia simple: alfanumérica, numérica, de fechas y genéricas, es decir, con símbolos comodines, que permiten buscar dentro de un patrón más amplio.

También podemos buscar referencias adyacentes, referencias consecutivas separadas por blancos, o bien en una distancia fija, o en un radio más amplio. Estas posibilidades se amplían con los llamados *operadores booleanos*, los de *unión, intersección* y *diferencia: .O, .Y, .NO,* Para alterar la jerarquía de los operadores se pueden utilizar paréntesis, tal como se hace en las operaciones aritméticas y lógicas.

El *Glosario de lemas y formas* es una ayuda útil para las personas que no puedan o no deseen usar las múltiples variantes de las formas antiguas del léxico. Para usarla, dentro del *Glosario de lemas* del menú *Opción*, basta con escribir la forma básica (el *lema*) que se busca y pulsar *intro*. Tras ello aparecerán todas las formas asociadas con ese lema. Si el lema pedido no estuviera en la base de datos, parecería el más próximo alfabéticamente. En el futuro esta opción será sustituida por un diccionario informatizado completo de la lengua medieval y clásica.

UNA BIBLIOTECA DEL DESCUBRIMIENTO EN EL DISCO I

El *disco 1*, realizado entre 1990 y 1992, contiene sesenta y un títulos, la mayoría de ellos de incunables, aunque hay algunos impresos antiguos, todos ellos de la Biblioteca Nacional de España, en Madrid, que corresponden a los apartados: *Enciclopedias, Diccionarios y Gramáticas, Textos Legales, Textos Científicos, Libros de viajes, Crónicas y Biografía, Tratados de Caballería y Nobleza, Textos poéticos*. La intención ha sido la de presentar en este disco inicial lo que podría haber sido una biblioteca de un contemporáneo de los navegantes que partieron de Palos rumbo a las Indias, en el error geográfico sin duda más productivo de la Historia.

Ese mundo del humanismo en el que el latín va dejando su paso al castellano se muestra en las gramáticas de Nebrija, o sus diccionarios, de ambas lenguas, en las *Partidas* alfonsíes, base del ordenamiento jurídico español e hispanoamericano, en las *ordenanzas* promulgadas por los Reyes Católicos, en textos científicos que van desde la medicina a la veterinaria o la cosmografía, desde los autores árabes a los castellanos.

Cuando el viaje constituye la contraseña del fin de siglo, no podían faltar los textos de este tipo, incluido el Marco Polo de Rodrigo de Santaella. Si se trata de hombres, los modelos de la época están en la *Crónica Popular del Cid*, en el *Valerio de las historias eclesiásticas y de España*, de Diego Rodríguez de Almela, o en los *Claros Varones de Castilla*, de Hernando del Pulgar, entre otros. Algunos de estos caballeros leían y aprendían en el *Doctrinal de los caballeros* de Alonso de Cartagena o en el *Nobiliario vero* de Fernando Mejía y todos se solazaban con los textos de Juan del Encima, de Mena o de Iñigo López de Mendoza.

Así se rinde homenaje a los que llevaron la Ley, la imprenta, la universidad y, en suma, la Latinidad, a las tierras por ellos descubiertas, cumpliendo el viejo ideal de Roma, reencarnado en la Hispania Provincia[1].

1. Los datos técnicos sobre los procesos de recuperación y digitalización de imágenes, así como los correspondientes al entorno Windows para ADMYTE, nos han sido facilitados amablemente por los especialistas de MICRONET, S.A. Agradecemos también a Charles B. Faulhaber, Angel Gómez Moreno, Aurora Martín de Santa Olalla, Julián Martín Abad, Manuel Sánchez Mariana y todo el equipo de correctores y revisores su colaboración en ADMYTE. ADMYTE se puede adquirir mediante carta o fax a MICRONET, S.A. Dpto. CD-ROM; c/ María Tubau Nº 7 Edificio Auge III, 6ª planta; 28050 Madrid España. Telf. (91) 3589625; fax: (91) 3589544.

REFERENCIAS

Blecua, Alberto
1983 *Manual de crítica textual* Madrid: Castalia.

Cabaniss, Margaret S.
1970 "Using a Computer for Text Collation", *Computer Studies in the Humanities and Verbal Behaviour,* 3, 1-33.

Cannon, Robert L. Jr.
1976 "COP-COL: An Optimal Text Colation Algorithm", *Computers and the Humanities,* 10, 33-40.

CASE: *Computer Assistance to Scholarly Editing, A User's Guide*
1983 Mississippi State: University.

Dearing, Vinton A.
1984 *Some Microcomputer Programs for Textual Criticism and Editing, Machina Analytica: Occasional Papers on Computer-Assisted Scholarship,* Nº 1, Los Angeles: William Andrews Clark Memorial Library.

Faulhaber, Charles B. et al.
1984 *Bibliography of Old Spanish Texts (Literary Texts, Edition-3)* Madison: Hispanic Seminary of Medieval Studies.

Faulhaber, C.B. y Francisco Marcos Marín
1989-90 "ADMYTE: Archivo digital de Manuscritos y Textos Españoles", *La Corónica,* 18:2, 131-145.

Froger, dom Jacques
1968 *La critique des textes et son automatisation* Paris: Dunod.

Greenia, George D.
1989 "The Libro de Alexandre and the computerized editing of texts", *La Corónica,* 17, 55-67.

Lancashire, Ian & Willard McCarty
1988 *The Humanities Computing Yearbook 1988.* Oxford: Clarendon Press.

Mackenzie, David
1984 *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language (With Spanish translation by José Luis Moure).* 3ª ed. Madison: Hispanic Seminary of Medieval Studies.

Marcos Marín, Francisco
1985 "Computer-Assisted Philology: Towards a Unified Edition of OSp. Libro de Alexandre", *Proceedings of the E[uropean] L[anguage] S[ervices] Conference on Natural-Language Applications, section 16,* Copenhague: IBM Denmark.

Marcos Marín, Francisco
1986a "Metodología Informática para la Edición de Textos", *Incipit,* Buenos Aires, vi, 185-197.

Marcos Marín Francisco
1986b "UNITE: conjunto de programas para el tratamiento filológico de textos en verso", *Procesamiento del Lenguaje Natural, [Sociedad Española para el Procesamiento del Lenguaje Natural]* 4, 43-55.

Marcos Martín, Francisco
1987a *Libro de Alexandre. Estudio y edición.* Madrid: Alianza Universidad.

Marcos Marín, Francisco
1987b "El Libro de Alexandre: Edición unificada por ordenador", *LEA,* IX, 1987, 347-370.

Marcos Marín, Francisco
1988a "Recuperación de información lingüística y tratamiento crítico de textos", *Actas, Simposio Internacional de Educación e Informática,* Madrid, 15 al 18 de junio 1987. Madrid: Instituto de Ciencias de la Educación, Universidad Autónoma de Madrid, 187-196.

Marcos Marín, Francisco
1988b "El Libro de Alexandre: Notas a partir de la primera edición unificada por ordenador", *Actas del I Congreso Internacional*

de Historia de la Lengua Española. Madrid: Arco Libros, 1988, 1025-1064.

Marcos Marín, Francisco
1989 [1991] "UNITE, a Package for Computer Assisted Philological Editing", *Folia Linguistica Historica*, X 1117-143.

Marcos Marín, Francisco
1991b "Computers and Text Editing: A Review of Tools, an Introduction to UNITE and Some Observations Concerning its application to Old Spanish Texts", *Romance Philology*, XLV/1, 1991, 102-122, (Bibliography: 205-237).

FMM y Aurora Martín de Santa Olalla, Charles B. Faulhaber, Angel Gómez Moreno
"ADMYTE: The Digital Archive of Spanish Manuscripts and Texts", *Sesame Bulletin. Language automation worldwide*, 5/2 (summer 1992), 50-61.

FMM y Charles B. Faulhaber
"La conservación y utilización de textos en el futuro inmediato: ADMYTE, el archivo digital de manuscritos y textos españoles", *Hispania*, 75/4, 1010-1023.

FMM y Pilar Salamanca Fernández
1987 "Programas informáticos para la crítica textual", *Telos*, 11, 105-111.

FMM y Jesús Sánchez Lobato
1988 *Lingüística Aplicada.* Madrid: Síntesis.

Meijs, Willem (ed.)
1987 *Corpus Linguistics and beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi.

Oakman, Robert L.
1984 *Computer Methods for Literary Research*, 2nd. ed. Athens, GA: University of Georgia.

Salamanca Fernández, Pilar
1987 "Crítica textual e informática: los programas UNITE", FUNDESCO, *Boletín de la Fundación para el Desarrollo de las Comunicaciones* 73, 8-10.

Shillingsburg, Peter L.
1986 *Scholarly Editing in the Computer Age.* Athens: University of Georgia Press.

Timpanaro, Sebastiano
1981 *La genesi del metodo del Lachmann,* 2 ed. Padova: Liviana Editrice.

Uthemann, Karl-Heinz
1988 "Ordinateur et Stemmatologie. Une constellation contaminée dans une tradition grecque". *Spatial and Temporal Distributions, Manuscript Constellations. Studies in language variation offered to Anthonij Dees on the occasion of his 60th birthday.* Ed. Pieter van Reenen and Karin van Reenen-Stein, 265-277. Amsterdam: Benjamins.