Aproximación metodológica a la digitalización de documentos textuales históricos y su aplicación al diseño de un sistema de información hipermedial sobre el teatro español de los Siglos de Oro

Jesús Robledano Arillo
Profesor Ayudante Doctor del Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid.
Departamento de Biblioteconomía y Documentación
C/ Madrid, 126-128. 28903 Getafe (Madrid)
jroble@bib.uc3m.es
Teléfono: 91 6249249
Fax: 91 6249212

Arturo Martín Vega
Profesor Titular del Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid.
Departamento de Biblioteconomía y Documentación
C/ Madrid, 126-128. 28903 Getafe (Madrid)
arturom@bib.uc3m.es
Teléfono: 91 6249256
Fax: 91 6249212

José Antonio Moreiro González
Catedrático del Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid.
Departamento de Biblioteconomía y Documentación
C/ Madrid, 126-128. 28903 Getafe (Madrid)
jamore@bib.uc3m.es
Teléfono: 91 6249238
Fax: 91 6249212

Resumen

Se presenta una aproximación teórica y metodológica a la digitalización de documentos textuales históricos. Se ilustra su aplicación a través de la presentación de un estudio de caso: la fase de selección y digitalización de fuentes del Proyecto *TESORO*. La metodología se fundamenta en una serie de concepciones sobre la naturaleza material del documento con valor histórico que se articulan sobre la idea de la valoración de los aspectos físicos y formales del documento manuscrito e impreso como conformadores de lo que denominamos información intrínseca del

documento. Este tipo de información es esencial para la interpretación y estudio de los documentos, por lo que su captura digital se considera tan relevante como la captura del propio contenido intelectual vehiculado en aquellos.

Palabras clave

Fondos digitales, digitalización, documentos textuales, metodología.

Introducción

La finalidad del presente trabajo es dar a conocer la aproximación teórica y metodológica empleada para el desarrollo de una de las fases de un proyecto de investigación realizado por un grupo de profesores del Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid, entre los que se encuentran los comunicantes, titulado *Edición electrónica del teatro español del Siglo de Oro para la difusión del español y la formación a distancia (TESORO)*. Este proyecto ha sido financiado por el Ministerio de Ciencia y Tecnología a través del *Programa de Fomento de la Investigación Técnica* (Programa PROFIT 2000-2003, Acción Genérica en Sociedad de la Información, Ref.: FIT-150200-2001-75). Su objetivo fundamental fue contribuir a la creación de contenidos en lengua española en Internet utilizando la literatura de los Siglos de Oro españoles, y en concreto las obras de teatro, por su riqueza con relación a las posibilidades hipermedia que permiten. Para el cumplimiento de dicho objetivo se marcaron las siguientes tareas y metas:

1. La digitalización, el almacenamiento, procesado y difusión del patrimonio histórico-artístico del español, mediante la digitalización de textos de obras de teatro de los siglos de oro españoles y de la información relacionada con dichos textos, tales como biografías de los autores, documentación sobre la historia de las representaciones, los personajes y los lugares citados en las obras o la bibliografía producida a lo largo de la historia sobre los autores.
2. El procesamiento técnico de los documentos digitalizados, mediante marcado XML y hojas de estilo asociadas.
3. La preparación de un curso a distancia sobre el teatro español del siglo de oro.

En esta comunicación, nos detenemos en los presupuestos teóricos y en la metodología empleados para abordar la fase de digitalización de los textos teatrales que fueron incluidos en el proyecto. Presuponemos el conocimiento por parte de los lectores de una serie de conceptos relativos a la tecnología de la imagen digital aplicada a documentos, por lo que no nos extendemos en la explicación de los términos de dicha tecnología que vamos a emplear a lo largo de este texto.

Objetivos de la digitalización del documento textual

Entendemos que la finalidad de la fase de digitalización en el tipo de proyecto que estamos describiendo es conseguir la captura digital de los documentos que van a formar parte del sistema de información con un nivel de corrección suficiente como para que las imágenes o textos digitales obtenidos cumplan unos requerimientos de calidad determinados. Esos requerimientos deben estar suficientemente claros al inicio del proceso, pues serán utilizados para la toma de decisiones sobre la tecnología y los parámetros de captura a emplear en la obtención de las versiones digitales. Partimos de la base de que no sólo el conocimiento de la tecnología a aplicar para la obtención de versiones digitales de documentos en su origen no digital garantiza una captura correcta, sino que también es imprescindible considerar, comprender y precisar con exactitud las características de esos documentos originales que deberán estar reproducidas con suficiente nivel de corrección en sus versiones digitales.

A través de la implantación de un sistema de información que permita el acceso a imágenes digitales se persigue evitar en la mayor medida posible el acceso al objeto original, que es sustituido por la versión digital que el usuario obtiene en pantalla o por impresora. Cuando se trata de difundir documentos con valor histórico a través de versiones digitales, pensamos que éstas deben aportar información de esos documentos con unos niveles de calidad –o fidelidad original copia digital– que garanticen su utilidad para aquellas personas que pueden hacer uso del sistema de información en el que se integran. El usuario debe poder encontrar en esa versión aquella información del original que necesita obtener. Esto es, al comienzo del proyecto de digitalización hay que tener suficientemente claro qué es lo que se quiere difundir, o lo que es lo mismo, qué información de estos objetos va a ser posible ofrecer a través de sus versiones digitales, y el nivel de corrección de la información que sobre éstos se ofrecerá; y si esa información y calidad satisface a las personas que harán uso de esas versiones digitales.

Nos encontramos, en consecuencia, con una serie de variables que deben ser consideradas en dicha fase inicial; y que se articulan en torno a dos ejes: por una parte, la información del documento que es posible reproducir a través de un medio digital; y por otra parte, la información del documento original que los usuarios necesitan para el uso que van a dar a esos documentos, y que debería estar representada en las versiones digitales que se le ofrecen en el medio digital puesto a su alcance. Estas variables son muy dependientes de la concepción que los responsables del proceso de digitalización tienen de la naturaleza de los propios documentos que van a ser difundidos y de las necesidades de los usuarios con respecto a éstos.

Bajo estas premisas, entendemos que uno de los principios en los que debe ampararse un proceso de digitalización es el conocimiento de la naturaleza del tipo de documentos que va a ser capturado, y una vez entendida ésta, definir clara y objetivamente el tipo de información de éstos que requiere ser capturada, así como el nivel de fidelidad que se va a ofrecer en su reproducción.

Naturaleza del documento textual con valor histórico y el alcance de su digitalización

Otro presupuesto teórico fundamental que contemplamos es la idea de que los documentos textuales con valor histórico no deben ser entendidos exclusivamente como contenido intelectual, sino también como artefactos físicos conformados por una materialidad que aporta igualmente información útil al investigador.

El contenido intelectual es conocimiento (ideas o datos) expresado de forma intencionada a través de un sistema de representación simbólico o icónico concreto, esto es, de un lenguaje, que se manifestará físicamente en un texto de carácter lingüístico o gráfico representado sobre un soporte (papel, pergamino, papiro...) El documento, como objeto material, presenta unas características físicas propias de la tecnología y materiales de escritura empleados en su realización, de la forma física particular de representación del texto lingüístico o gráfico que vehicula, y de procesos de alteración física causados por su manipulación posterior o por las condiciones en las que ha sido almacenado o procesado a lo largo del tiempo. El conjunto de estas características es lo que denominamos información intrínseca del documento, que se manifiesta por la forma física que presentan los elementos materiales que conforman el documento en un momento dado. Dichos elementos sufren procesos de transformación a lo largo del tiempo, cuya manifestación física puede aportar información relevante sobre la tecnología empleada para la elaboración del documento o las circunstancias materiales a las que ha sido sometido éste durante su período de existencia. Las transformaciones pueden ser intencionadas o debidas a procesos de deterioro propios del envejecimiento o la agresión ambiental sufridos por el documento.

La utilidad de la información intrínseca nos parece esencial para la interpretación del documento textual manuscrito o impreso, pues permite obtener datos sobre: el acto de creación del documento; aspectos o circunstancias de tipo técnico, institucional, social y personal que envuelven dicho acto; intencionalidad estética en la apariencia visual; y el devenir de ese documento a lo largo del tiempo. La conjunción de estas características aporta la forma visual del documento, esto es, su apariencia, y, consecuentemente también, su valor estético.

Si tenemos en cuenta el valor de la información intrínseca para el investigador, hemos de considerar que el alcance de la digitalización es no sólo captar el contenido intelectual sino esa materialidad física sobre la que se vehicula aquél. Entonces, una reproducción fiel del documento implica que su reproducción digital represente con fidelidad los atributos materiales de éste que son accesibles a través del sentido de la vista y que han sido estimados como relevantes para el uso de ese documento como fuente para la investigación. Los elementos físicos del documento desde el punto de vista material a considerar son: el soporte, el medio de escritura o representación gráfica empleado, los trazos que conforman el texto escrito o elementos gráficos, la información textual o gráfica incrustada en soporte (marcas de agua), y cualquier otro elemento añadido al soporte (sellos, recortes...)

El conocimiento de la materialidad del documento es importante también para soslayar el problema que representan los procesos de deterioro que ha podido sufrir el docu-

mento a la hora de la captura digital. Es relativamente frecuente encontrar en documentos con cierta antigüedad la presencia de procesos de deterioro, o incluso de transformación intencionada, que afectan al soporte en papel y a las tintas, y que dificultan la legibilidad de los textos sobre el propio original al impedir o dificultar el acceso y la identificación de los signos empleados. En ocasiones el documento sólo conserva restos de la representación material original del mensaje, de nula o escasa visibilidad a simple vista. Estos problemas de legibilidad se verán incrementados en las versiones digitales si no se aplican los parámetros de captura digital adecuados. Para ello es importante conocer el tipo de deterioro y el grado en que afecta éste al texto. La representación digital debe posibilitar la legibilidad de los textos materializados en el documento cuando ésta es precaria en el original. Así, la consecución de una copia digital legible ayudará a la descodificación correcta del mensaje original, puesto que el lector podrá identificar los signos lingüísticos y gráficos que fueron empleados.

Conseguir un documento digital legible, en muchos casos, implica la consideración de estos aspectos físicos, que pueden llegar a impedir o dificultar por un lado, la obtención de un texto suficientemente claro en pantalla o papel impreso, y por otro, la consecución automática de texto en formato digital, mediante un sistema OCR (Reconocimiento Óptico de Caracteres). La legibilidad de un documento puede ser mejorada a través de la aplicación de funciones de proceso de imagen digital. En estos casos, el proceso de captura digital puede ayudar a aumentar la legibilidad del original sin requerir actuación física sobre el objeto, mediante el empleo de procesos de *restauración óptica o digital*. Con la aplicación de este tipo de técnicas se puede llegar a hacer visible una forma aproximada de la representación del mensaje original, permitiendo la descodificación de los signos y el acceso al mensaje[1]. El proceso de captura digital, en principio, no supone la disminución de la legibilidad del documento original, y sí así lo hiciera es necesario aportar al usuario en el sistema de visualización de las imágenes herramientas que le permitan llegar, al menos, al estado de legibilidad del original en el momento de la digitalización. La restauración óptica no tiene que implicar necesariamente pérdida de información, pues la realización de versiones digitales con legibilidad mejorada no es un proceso excluyente, sino complementario a la realización de versiones que representan fielmente los atributos físicos del original, si se permite que el usuario acceda a ambas imágenes.

La metodología aplicada a las fases de selección y digitalización de fuentes del proyecto TESORO

Bajo los presupuestos teóricos descritos en los epígrafes anteriores hemos desarrollado una metodología en la que diferenciamos las siguientes fases: selección de fuentes, caracterización física del grupo de documentos a digitalizar, definición de las técnicas y parámetros más adecuados para la captura digital y el control de calidad de los productos digitales, definición de procesos de restauración digital a realizar en las versiones de legibilidad mejorada de los documentos, pruebas de rendimiento, definición del protocolo de digitalización, proceso de digitalización, obtención de las diferentes versiones digitales del documento, control de calidad y corrección de resultados o repetición de procesos.

Durante la fase de selección de fuentes, se decidió el carácter y la acotación temporal de las obras que serían objeto de captura digital. Fue necesario concretar las fechas marco entre las cuales se iba a producir la elección de los autores dramáticos cuyas obras iban a ser digitalizadas. El primer problema que se suscitó fue la acotación temporal respecto al teatro español del Siglo de Oro, denominación bajo la que subyace cierta ambigüedad cronológica. Como resultado de las discusiones se delimitaron las fechas comprendidas entre 1500 y 1700 por dos razones. En primer lugar, la crítica moderna acepta que la Edad Dorada de la literatura española, o Siglo de Oro español, abarca tanto el Siglo XVI como la mayor parte del Siglo XVII (al menos hasta 1681). En segundo lugar, dados los objetivos del proyecto, era interesante considerar, como posible análisis, las producciones dramáticas menos conocidas y estudiadas dentro del límite temporal de los Siglos de Oro; esto es: las obras teatrales correspondientes también al Siglo XVI.

Una vez definido el período de producción de las obras que habría que digitalizar, el equipo de recopilación se encargó de realizar una nueva comprobación de las obras ya digitalizadas sobre el teatro español de los Siglos de Oro con el fin de evitar el solapamiento de actividades y recursos ya existentes desarrollados a título personal o colectivo. Durante este proceso se constató un tratamiento desigual en los campos de investigación. Por ello se seleccionaron como posibles unidades de análisis del proyecto, aquellos textos de autores teatrales poco conocidos y estudiados fuera del ámbito profesional, y escasamente representados en los recursos de acceso electrónico. Ello dio lugar, en principio, a centrar la selección de textos en torno a tres bloques: autos sacramentales, obras de teatro escritas por mujeres y tragedias del Siglo XVI.

Para cada obra se localizaron al menos dos versiones: la edición crítica y la edición *princeps*. Dentro del marco fijado para la selección de los primeros textos objeto de tratamiento, y considerando los tres bloques temáticos antes mencionados, se tuvieron en cuenta las siguientes variables: la facilidad de acceso a las obras que presentaran una cobertura representativa de los Siglos de Oro, la significación teatral de las obras y la calidad de conservación de la obra para un eficiente tratamiento informático. El resultado provisional ha sido, como experiencia piloto, la digitalización y la marcación en XML de un grupo de cinco obras: tres autos de Gil Vicente: *Auto de la Visitación o Monólogo del Vaquero, Auto de los Reyes Magos* y *Auto Pastoril Castellano;* la *Tragedia de la Destrucción de Constantinopla* de Gabriel Lobo Lasso de la Vega; y *Las Muñecas de Marcela* de Álvaro Cubillo de Aragón.

Durante las dos siguientes fases -caracterización física del grupo de documentos y definición de las técnicas y parámetros de captura digital- se empleó una metodología de *benchmarking* aplicada a la digitalización de documentos que ha sido desarrollada y difundida por investigadores del *Cornell University Department of Preservation and* Conservation[2]. La aplicación de esta metodología hizo posible obtener los parámetros de captura digital adecuados a los objetivos marcados, en cuanto a resolución espacial y profundidad de bit. En este proceso trabajamos con una muestra de documentos que mostraban unas características formales representativas del núcleo de obras a digitalizar.

Una limitación importante, y que tuvo una fuerte incidencia en el desarrollo del proyecto, fue la imposibilidad de poder manejar los originales de los documentos seleccionados. En su lugar, la institución depositaria nos hizo llegar copias en papel de los docu-

mentos obtenidas a partir de microfilm. Las copias pasaron a convertirse en el objeto de la digitalización para la obtención de los facsímiles digitales. Las pruebas se hicieron sobre dos categorías de documentos relevantes para el proyecto. Por una parte, se emplearon las copias facsímiles en papel de originales impresos en el siglo XVII que nos fueron facilitadas. Se trataba de estudiar la posibilidad de obtener facsímiles digitales que pudieran ser presentados al usuario en pantalla, con suficiente legibilidad y corrección en la representación de las características materiales relevantes para el investigador. También debíamos comprobar si era viable obtener directamente texto digital aplicando un software de OCR a estos facsímiles. Por otra parte, se utilizaron textos obtenidos de ediciones críticas recientes. En este segundo caso, se trataba de comprobar el rendimiento del software de OCR (reconocimiento óptico de caracteres) con tipos de fuentes actuales.

Con respecto a la primera categoría de documentos, como las copias en papel fueron obtenidas a partir de microfilm, se plantearon serios problemas de rendimiento, tanto para la obtención de facsímiles digitales legibles y fieles al original, como para la obtención de texto digital producto de un proceso automático a través de OCR. Los problemas han derivado de los siguientes factores:

- Los documentos originales que fueron microfilmados manifestaban procesos de deterioro, consistentes fundamentalmente en trasposición de tintas y en manchas. La transposición de tintas es debida a un proceso de degradación conocido como "corrosión de tintas". Este proceso obedece a la acidificación y corrosión que provocan las tintas ferro-gálicas, empleadas en la impresión, sobre el soporte de papel. Las trasposiciones de tintas dan lugar a la visibilidad de las letras de la cara opuesta de la hoja. Al microfilmar y fotocopiar posteriormente del microfilm, se produce un fuerte aumento del contraste del documento, que acarrea una pérdida considerable de margen de tonos. La pérdida tonal deriva en que los trazos de las letras llegan a alcanzar valores tonales muy parecidos a los de las manchas y trasposiciones. Este fenómeno dificulta tanto la legibilidad del texto como el reconocimiento automático realizado por el software de reconocimiento óptico de caracteres (OCR). Las pruebas de rendimiento dieron como resultado porcentajes entorno al 50% de texto reconocido con corrección, pero considerando como unidad la letra, no la palabra. Si consideramos como unidad la palabra, el porcentaje baja considerablemente[3].
- El problema de la reducción y alteración del margen de tonos de los originales, resultado de tener que trabajar necesariamente con copias de "tercera generación", acarrea la deformación de los tipos de letra, con lo que la imagen digital resultante no permite ofrecer con fiabilidad la apariencia de las fuentes empleadas en las ediciones de época.

Como consecuencia, el nivel de calidad de los facsímiles digitales no se consideró suficiente como para justificar su integración en el sistema de recuperación en esta primera fase del proyecto. No obstante, se ha contemplado para una segunda fase la construcción de filtros de proceso automático de las imágenes digitales, que permitan mejorar los problemas tonales y corregir las trasposiciones de tintas y manchas que dificultan la legibilidad y el rendimiento del OCR.

El segundo tipo de documento son versiones críticas de las obras teatrales. Estamos ante documentos impresos modernos, sin problemas de preservación y con tipos de letra que reconoce el producto de OCR con un nivel de corrección bastante alto. Los textos digitalizados han sido a partir de estas ediciones modernas, eliminando todo el texto correspondiente al aparato crítico de la edición durante la fase de OCR.

La siguiente etapa consistió en la realización de protocolos de captura para OCR, una vez descartada la inclusión de facsímiles digitales. Estos protocolos agilizaron el proceso de digitalización, permitiendo que diferentes operadores trabajaran con una secuencia de operaciones ya demostrada como eficaz y eficiente. El protocolo de digitalización utilizado ha resultado en un tiempo de proceso automático con un promedio de 30 segundos por página. El proceso que ha requerido más tiempo es el de la revisión del texto producido por el OCR, pues necesariamente tiene que ser realizada por un operador humano, se debe hacer de manera muy minuciosa, cotejando, cada uno de los caracteres del texto obtenido con los caracteres del original. El tiempo medio de la revisión humana se ha situado en torno a los tres minutos por página. Para acelerar el proceso de marcado XML se acordó un formato uniforme de separadores de elementos estructurales del documento en los correspondientes ficheros que contenían el texto digital de las obras.

Conclusiones

En esta comunicación nos hemos centrado en mayor medida en un intento de reflexión teórica sobre la digitalización de documentos con valor histórico, destacando una serie de argumentos a favor de la fidelidad en la representación digital tanto del contenido intelectual como en el contenido intrínseco del documento, sin menoscabo de la legibilidad de los textos. Hemos querido ejemplificar su aplicación práctica con el caso de la digitalización del proyecto TESORO, destacando de manera especial la incidencia de esos principios en las decisiones tomadas en el transcurso del proyecto. Pensamos que la aplicación de las tecnologías de la imagen digital a la difusión del patrimonio documental requiere una reflexión y un trabajo teórico en el que deben confluir conocimientos de diferentes campos, tales como la Diplomática, Paleografía, Historia, Tecnología de la Imagen Digital, Archivística, Documentación, Tipografía, Conservación y Restauración.

Bibliografía

AYRIS, P. (1999). "Guidance for Selecting Materials for Digitisation". En: Joint RLG and NPO Preservation Conference. Guidelines for Digital Imaging. http://www.rlg.org/preserv/joint/ayris.html. [Consulta: 25/09/2002].

DEPARTAMENTO de Preservación y Conservación (Cornell University Library). (2002). Llevando la teoría a la práctica. Tutorial de Digitalización de Imágenes. http://www.library.cornell.edu/preservation/tutorial-spanish/contents.html. [Consulta: 25/09/2002].

FREY, F. (2002). File Formats for Digital Masters, Guide 5 to Quality in Visual Resource Imaging. http://www.rlg.org/visguides/visguide5.html. [Consulta: 25/09/2002].

HAZEN, D.; Horrell, J. y Merrill-Oldham, J. (1998). Selecting Research Collections for Digitization. http://www.clir.org/pubs/reports/hazen/pub74.html. [Consulta: 25/09/2002].

KENNEY, A. R. (2002). "Digital Benchmarking for Conversion and Access". En: Kenney, A. R. y Rieger Y. O. Moving Theory into Practice: Digital Imaging for Libraries and Archives. Mountain View, CA: Mountain View: Research Libraries Group, 2000. p. 24-60.

KENNEY, A. R. y Chapman, S. (1996). Digital Imaging for Libraries and Archives. Ithaca: Cornell University Library.

KENNEY, A. R. y Rieger Y. O. (2000). Moving Theory into Practice: Digital Imaging for Libraries and Archives. Mountain View: Research Libraries Group.

MENNE-HARITZ, A. y Brubach, N. (2001). The Intrinsic Value of Archive and Library Material. List of Criteria for Imaging and Textual Conversion for Preservation. http://www.uni-marburg.de/archivschule/intrinsengl.html. [Consulta: 25/09/2002].

PRESCOTT, Andrew. The Electronic Beowulf and Digital Restoration. URL: http://www.uky.edu/~kiernan/eBeowulf/ajp-llc.htm [Consulta: 25/09/2002].

Sellink, M. Iron gall ink corrosion - the role of collection keepers. http://www.knaw.nl/ecpa/ink/html/colman.html. [Consulta: 25/09/2002].

THE Iron Gall Ink Corrosion Web Site. http://www.knaw.nl/ecpa/ink. [Consulta: 25/09/2002].

THE project Digital Image Archive of Medieval Music (DIAMM). http://www.diamm.ac.uk/index1.html. [Consulta: 25/09/2002].

KNOX, K, et. al. (2001) "Multispectral Imaging of the Archimedes Palimpsest". En: Proceedings of the 54th Annual PICS Conference, IS&T, p. 206-210.

Notas

[1] Las técnicas de restauración óptica y digital de documentos cubren un amplio espectro de casos de aplicación y de tecnologías. El objetivo de estas técnicas es posibilitar la captura y representación digital de información que contenía el documento y que ya no es visible para el ser humano, debido a desvanecimiento severo de las tintas empleadas en la impresión o escritura, tachaduras, borrados, sobreimpresiones, manchas o trasposiciones de tintas. El uso de luces infrarrojas o ultravioleta y de dispositivos de captura digital (cámaras fotográficas digitales o escáneres) convenientemente adaptados puede hacer posible la recuperación de esta información perdida. Algunos ejemplos de utilización de estas técnicas en digitalización de documentos históricos son: *The project Digital Image Archive of Medieval Music (DIAMM)*, la digitalización del Beowulf, o el palimpsepto de Arquímedes.

[2] El término *benchmarking* se emplea aplicado al proceso de definición de los requerimientos de captura digital necesarios para la correcta digitalización de un fondo documental determinado. Esta práctica considera factores como los atributos físicos de los documentos, las necesidades de los usuarios, las necesidades de preservación de los

fondos, los recursos económicos necesarios y el tiempo de ejecución. Las tareas fundamentales de un proceso de *benchmarking* son:

- Definición de los requerimientos en cuanto a calidad visual de las reproducciones digitales. Se trata de analizar los atributos físicos de los documentos que son considerados relevantes para su reproducción digital, tales como el color y la necesidad de su reproducción exacta, el detalle visual significativo más pequeño, el tipo de soporte (traslúcido, opaco), el rango de densidades de los originales, la presencia de grano en la imagen fotográfica, etc.
- Definición de las necesidades de los usuarios actuales y futuros.
- Medida. Se trata de medir objetivamente tales atributos y estudiar cuales son los parámetros de captura digital y equipos necesarios para capturar y poder reproducir digitalmente las características de esos atributos. Los atributos se deben correlacionar a través de fórmulas, por ejemplo: "el detalle visual más fino del fondo a digitalizar tiene una anchura de 0,1 mm, y para poder reproducir correctamente ese detalle hacen falta al menos dos pixels, por lo que la resolución espacial de captura más adecuada será de 600 ppp".
- Determinar valores de tolerancia. Se pueden determinar valores de tolerancia respecto a los requerimientos definidos en las fases anteriores, reflejando las consecuencias de esos valores. Por ejemplo se pueden definir menores resoluciones espaciales de captura con respecto al valor ideal, indicando las consecuencias en cuanto a merma de calidad y ahorro de espacio de almacenamiento.
- Confirmación de resultados por pruebas y evaluaciones. Se trata de probar los parámetros de captura decididos como más adecuados, a través de pruebas con un grupo de documentos suficientemente representativo del fondo y de la evaluación por parte de expertos, los profesionales que custodian los fondos y los usuarios.

A partir de la información obtenida en estos estudios se definen los parámetros de calidad requeridos para la conversión a formato digital de los originales del fondo documental, y para la generación de los diferentes derivados de la imagen digital (imágenes a diferentes resoluciones). Para acceder a una descripción detallada de las técnicas de *benchmarking* aplicadas consúltese las obras Anne Kenney que incluimos en la bibliografía.

[3]. Para la captura de OCR se utilizó el programa Text Bridge Pro Milenium.