PRIMERA PARTE

DOCUMENTOS DIGITALIZADOS Y SUS VENTAJAS. UN ESQUEMA

1. Introducción

En este apartado se describirán las ventajas que ofrecen los textos digitalizados, con el propósito de resaltar su validez y efectividad, a la hora de realizar una investigación en particular. Tomando en cuenta que la esta tesis forma parte de un proyecto de digitalización e investigación más amplio, este capítulo se podría considerar como una justificación más detallada del trabajo. También se abordarán algunas consideraciones técnicas sobre los alcances didácticos de los textos digitalizados, así como sobre cuestiones relacionadas con los derechos y deberes de la digitalización, de acuerdo a una serie de normas internacionales estandarizadas. Por último, se indicará con un ejemplo particular, cuál fue el procedimiento técnico de digitalización de los documentos contemplados para la tesis.

2. ¿Qué es un documento digitalizado?

Un documento digitalizado es aquel que ocupa un espacio virtual y cuya consulta se realiza por medio de la tecnología del momento; en este tipo de documento se facilita el acceso, la generación y diseminación de la información. Un documento virtual puede ofrecer servicios y opciones de búsqueda de información específica de forma rápida y segura, en contraste, el documento físico ocupa un espacio real y no ofrece ninguna

opción de búsqueda de tal efectividad como la digital. Una vez que se tiene la tecnología adecuada, el acceso fácil permite disponer con mayor libertad la consulta y uso de la información a las necesidades del usuario, por ejemplo, un libro no puede tener varios lectores al mismo tiempo, mientras que un documento electrónico tiene dicha capacidad. Además quienes lo consultan pueden navegar, al mismo tiempo, y gozar de los servicios de Internet o buscar vía web cierta información dentro del mismo texto. De esta manera la búsqueda se adapta a los requerimientos de quienes lo consultan. La riqueza del conocimiento será directamente proporcional a la accesabilidad del documento de cualquier tipo, hasta los de difícil adquisición e incluso textos únicos se pueden hacer accesibles mediante este proceso.

3. Documentos digitalizados y ventajas para la investigación

La digitalización de los textos apoya la investigación de muchas maneras, entre ellas, el texto digitalizado es más fácil de almacenar, transportar y maniobrar. El costo por mantener gran cantidad de libros dentro de un espacio físico puede ser mayor que guardarlos de manera electrónica: hay reducción de espacio, no hay uso de hojas, de copias, etc. Por supuesto, siempre hay que tener cuidado de mantener los programas y archivos al día puesto que la tecnología cambia con mucha rapidez. Esto da origen a una forma de conocimiento y catalogación muy distinta. Incluso las técnicas de recuperación, clasificación y catalogación tradicionales de una biblioteca no se utilizan en este contexto, existe una diferencia tal, que la biblioteca digital es considerada como otra disciplina de investigación multidisciplinaria porque en su desarrollo existe la participación de distintos investigadores de computación, ciencias de la información y expertos en cada área del conocimiento. Por esta razón, los servicios digitales aumentan

con referencia virtual –apoyo de expertos bibliotecarios a usuarios vía Internet–, recomendación de libros –que puede ser automática y personalizada–, libertad en anotaciones –posibilidad de hacerlos resaltar, pueden ser términos o fragmentos de textos de cualquier tamaño, ya sea en revistas o en libros digitales– recursos tecnológico de visualización –posibilidad de ver en dos y tres dimensiones–, prototipos de investigación en la búsqueda de videos –ya sea en la búsqueda de una escena en particular o frase de en audio–, acceso a datos geográficos, entre otras.[1]

Aparte de las ventajas mencionadas, el texto digital ofrece una mayor versatilidad ya que no solamente facilita la consulta de diversas clases de material de apoyo que en pocas ocasiones se muestran en las publicaciones de los libros. Este es el caso de un material musical sonoro y el de video. Dentro de las publicaciones es común que introduzcan ciertos anexos como lo serían las fotos, grabados, partituras, gráficas, fórmulas, etc., pero el texto digitalizado nos puede mostrar ejemplos de movimiento y sonido que no son propios de un libro, aunque ahora, gran cantidad de libros contiene un DVD o un CD con ejemplos visuales o auditivos o de ambos. Otra ventaja sería que mediante la impresión, el material digital se puede convertir en diversos tipos de apoyos pedagógicos como son los acetatos, o se puede abreviar con fines también pedagógicos.

Por otro lado, usualmente el contacto entre un investigador y un documento implica que uno de los dos tenía que transportarse de manera física. En el caso de textos valiosos el investigador viajaba hasta la biblioteca o archivo donde se encontraba la información que necesitaba. En el caso de libros o documentos de los que se tienen copias múltiples, siempre es posible un préstamo interbibliotecario, siempre y cuando el usuario mantenga la relación necesaria con una institución que tenga este tipo de

1 Disponible: http://biblio.udlap.mx/

convenios. Tener acceso a una biblioteca conlleva tiempo, dinero y relaciones ya que no todos los libros están a la disposición de cualquier usuario. Es necesario ser miembro de una comunidad específica, la universitaria, por ejemplo, para tener derecho a este tipo de servicios los cuales también tienen un alto costo para la institución que posee el libro o documento, o para el individuo que necesita el documento para sustentar o proseguir con su investigación. Alguna de estas restricciones existen en la distribución de ciertos documentos digitales cuyo acceso es restringido por derechos de autor o por derechos del que consulta el documento. Una muestra sería el New Grove Dictionary of Music que existe en versión digital, pero sólo puede ser consultarlo en las instituciones que pagan los derechos de acceso.

Actualmente es muy claro que gracias a la red electrónica el texto digital se mueve a mayor velocidad y menor precio, suprimiendo la salida física del individuo o del documento y provocando a su vez el manejo más democrático de la información. En general, gracias a su mayor disponibilidad, el documento digital puede estar al alcance de grandes comunidades de usuarios. Hay otras ventajas en las que el texto digital apoya a la investigación ya que, como se ha mencionado arriba, facilita ciertos tipos de búsqueda que antes eran difíciles o incluso imposibles. Los productos de investigación son el resultado de procesos cognoscitivos complejos en los que interviene de una manera preponderante la memoria del investigador.

El discurso final del investigador presenta los datos como éste los ha sintetizado y ordenado y en todo momento las conclusiones dependen de la impresión que el estudioso tiene sobre la materia que se discute. La persona que encuentra un documento digitalizado tiene mayores y mejores posibilidades para consultarlo y manejarlo, al contrario de los textos impresos. Una de éstas posibilidades es la búsqueda de nombres,

vocablos específicos, temas, obras, etc. Ésto, a su vez, permite una rápida revisión sobre la manera como un tema ha sido tratado en el texto o libro que se está citando. En el caso de libros impresos, las búsquedas relacionadas con nombres y vocablos suelen incluirse en el índice de nombres, materias, títulos de obras que aparecen al final de varios libros. Se trata de una práctica reciente y no una regla. Además, no todos los libros tienen estos índices, mismos que incluyen sólo la selección del autor y del editor, pero éstos no pueden prever todos los temas ni todos los vocablos que pueden interesar a un lector. Por ejemplo, el libro de *European Sacred Music* de John Rutter (Oxford University Press, 1996) sólo tiene índice de orquestaciones, mientras que el libro *Black Music in América* de JoAnn Skowronski (The Scarecrow Press, 1981) sólo tiene índice de autores; en cambio, el libro *Music Education* de Ernest E, Harris (Gale Research Company) tiene de manera ordenada y separada el índice de autor, el índice de título y el índice de temas. En el caso del libro de John Rutter, si un lector necesita saber si el libro contiene cierto tema, tendrá que revisarlo completamente para poder darse cuenta, de igual manera con el libro de JoAnn Skowronski: si el investigador quiere saber si habla o tiene cita de una obra en especial, tendrá que hacer una revisión más profunda que sólo la de buscar al autor de dicha obra. Incluso en el caso de Ernest E. Harris, que tiene índices más completos, puede llegar un investigador a buscar su lista larga de vocablos que no se incluyeron en el índice de temas, pero eso no quiere decir que en libro no se hable de éstos términos o que el autor no haya abordado los temas. También se puede asumir que el índice contiene todos los nombres de los autores que aparecen, pero ¿se puede estar completamente seguro de ello? Además de que las listas resultan incompletas no aparecen en todos los libros, de hecho en los libros publicados en español rara vez aparecen.

Digitalizando éstos documentos, existe la posibilidad real de saber si abordaron a un autor en especial y cuántas veces citaron una obra. Asimismo se pueden consultar, de manera fácil y rápida, vocablos no incluidos en los índices. La optimización del tiempo es considerable y la calidad de la búsqueda de palabras se efectúa con mayor facilidad, sin el riesgo que tiene el lector en pasar de largo el término tan solicitado dentro de una lectura con papel.

De la misma forma como el intercambio de libros sucede, primordialmente, en el seno de las universidades y centros de estudio, son estas instituciones las que han desarrollado los sistemas, criterios y políticas que se han desarrollado siguiendo la evolución tecnológica, la facilidad para almacenar y consultar datos, y el crecimiento y popularización de la red electrónica.[2] Esta red permite que los usuarios, sin importar su localización, tengan acceso a utilizar directamente una amplia gama de materiales convertidos en material digital que promueva la comprensión y el conocimiento, aumentando el acceso y ayudando a preservar colecciones raras o frágiles proporcionando sustitutos digitales, aumentando la eficacia de los procedimientos de la biblioteca y construyendo la colaboración con otras bibliotecas a fin de lograr poseer una masa de gran volumen de artículos digitales para formar un componente significativo de información digital.

Junto con la expansión de los procesos aquí descritos también se han desarrollado nuevos sistemas de clasificación. Dentro de este tipo de numeración, un primer planteamiento sería realzar el acceso a las colecciones de la biblioteca como: colecciones de audio, copiado digital, etc. Cada vez que un material es convertido en digital, es porque existe una selección de los materiales como parte integral de un

2 Disponible: http://www.bn.org.pl/dig-bib_eng.htm (Adquirido: 06 / 04 / 04)

proyecto especial. Dicha numeración, será a su vez, parte integral de las actividades de una biblioteca y será realizada de acuerdo con el acto de copyright.

Posterior al acceso de la información, el uso y la reproducción de documentos y artículos de colecciones convertidas en material digital estarán bajo las políticas de biblioteca en el acceso y la carga del usuario. De esta manera, implicará la atribución del trabajo al autor y el reconocimiento a la biblioteca como fuente. Otro punto a destacar es que las versiones finales digitales siempre deberán representar los materiales originales del modo más fidedigno posible, a tal punto, que las imágenes, texto, etc., no sean manipuladas, excepto para compensar defectos en el componente. [3]

Organizaciones importantes ofrecen sus modelos de digitalización y procesos. La planeación es el factor de inicio para la digitalización y en él se delimita y se selecciona el material que se ofrecerá a los usuarios: libros, documentos frágiles o únicos.[4] La numeración es organizar y realizar el acceso que se deberá crear para la colección. La organización incluye el establecimiento de las prioridades del orden y la numeración estará de acuerdo con el acto de derechos de autor –copyright– y las restricciones que se marquen dentro de ellos. En países como Australia, el formato y el programa a utilizar se especifica que deberá ser compatible con Personal Computer (PC).[5]

4. Condiciones legales para la digitalización de documentos y proyectos académicos

3 Una política digital de la preservación de la biblioteca nacional de Australia. Disponible: http://www.nla.gov.au/policy/digpres,html (Adquirido: 06 / 04 / 04)
4 Ibid.
5 Ibid.

Cabe destacar que los fines que tiene este proyecto de tesis de licenciatura es para la investigación, por lo tanto, no tiene fin lucrativo. Debido a estos fines de investigación, la primera fase del proyecto no se realizará con la universidad, sino que será un proyecto particular. De este modo, no hay una obligación directa para resolver los derechos de autor, puesto que no estará al alcance de cualquier usuario, sino que únicamente el acceso lo tendrá la dueña del proyecto y quien le ayuda. Si la autora de este trabajo decide hacerlo público, habrá una nueva etapa donde se revisarán los permisos correspondientes a los derechos de autor, así como también los respectivos formatos y normas que establece esta institución educativa. Esta actividad queda comprendida dentro de la numeración de la biblioteca:

> Cuando es posible, el permiso de convertir el material a digital será negociado cuando se adquiere el material. La biblioteca apuntará convertir los materiales para los cuales las restricciones del copyright han expirado; y para aquellos en que las restricciones del copyright todavía se aplica, donde a digital estos materiales del copyright constituyen una parte esencial de la colección. Dado el coste de negociar el permiso del copyright después de la adquisición, las actividades de la numeración de la biblioteca se centrarán inicialmente en el material libremente de las restricciones del copyright.[6]

Para los materiales que son patrimonio cultural se utiliza una guía para la buena práctica en la representación digital. Esta guía está enfocada en las artes y humanidades: guía NINCH (National Initiative for a Networked Cultural Heritage). Esta guía fue diseñada para todos los sectores que tienen a su cargo recursos culturales que serán convertidos al formato digital para establecer una red de la comunidad. La guía contiene la información de las decisiones que se harán a lo largo del ciclo de proyectos digitales: del planeamiento, de la selección, y del copyright del proyecto, con la numeración en todos los formatos, la "gerencia de activo Digital" y la preservación. Además contiene

6 Ibid.

bibliografía, informes de la entrevista y el instrumento de la entrevista que están disponibles. El sitio de red donde se encuentra esta guía está en la Universidad de Nueva York (NYU). Lorna Hughes es director auxiliar para las humanidades que se encarga de digitalizar información de NYU.[7]

La Federación de la Biblioteca Digital (DLF)[8] esta formada por un consorcio de bibliotecas y agencias que están relacionadas y comienzan el manejo de las tecnologías de electrónica e información para ampliar sus colecciones y servicios. El DLF provee la dirección para las bibliotecas por medio de sus miembros identificando estándares y las mejores prácticas para realizar colecciones digitales y para el uso y acceso de red, promueve la investigación, el desarrollo que coordina el uso de las bibliotecas de las tecnologías de la electrónica y la información y siembra proyectos y servicios que ayuden al comienzo de estas bibliotecas.

Los archivos de las bibliotecas, especialmente de las académicas, están invirtiendo gran cantidad de recursos económicos para la adquisición y realización de la información electrónica, pues estas colecciones electrónicas crecen cada día más y se vuelven más difíciles de manejar. Pocos sistemas de gerencia existentes entre las bibliotecas proporcionan herramientas adecuadas para el desarrollo local de las bases de datos. De ahí la necesidad de comenzar a desarrollar un sistema estándar de definiciones de los datos y un esquema común (de nombres, de definiciones, de relaciones semánticas para los elementos relacionados con la identificación, el acceso y otorgar la autorización de estos recursos).[9]

---

7 http://www.ninch.org/guide.html .
8 Por sus siglas en inglés, DLF: Digital Library Federation.
9 Un taller inicial fue sostenido de mayo el 10, 2002 en Chicago, publicado en septiembre el 30 de 2002 según los datos de: www.cordis.lu/digicult

En 2001, el Instituto de los Servicios del Museo y de la Biblioteca (IMLS) convocó a un foro[10] para considerar y promover estándares de "buenas prácticas"para que pudiera ser acogido por las comunidades de las bibliotecas para dirigir el desarrollo de las colecciones digitales. Se proporcionaron estatutos para la realización digital de materiales impresos e imágenes, indicando nivel deseable de calidad –sobretodo para las imágenes.[11]

Al año siguiente, IMLS convocó a un foro[12] para discutir acerca de las ediciones para bibliotecas digitales. Miembros del foro 2002 desarrollaron un plan para delimitar la calidad de las colecciones, la evaluación de los mecanismos, los patrones y las mejores prácticas que para su promoción. IMLS proveen un sistema de principios de alto nivel por los cuales los estándares específicos y las buenas prácticas pueden ser calificados.[13]

Se entiende por "buenas prácticas" a las colecciones digitales de calidad, evaluando diferentes aspectos: de buenos objetos –artículos que abarcan colecciones–, de metadata –información referente a los artículos y colecciones–, y del buen manejo del proyecto. Dos cualidades lo determinan: la interoperabilidad y la persistencia.[14]

Cabe hacer la diferencia entre ambas instituciones: DLF se encarga de delimitar los principios de guía y la estructura de la organización, el IMLS rige los estándares

10 2001 Tela-Sabio - El Digital Se divide: Una conferencia sobre bibliotecas y museos en el mundo Digital. Febrero de 2001
Disponible: www.imls.gov/pubs/forumframework.htm
11 Disponible: www.cordis.lu/digicult
12 2002 Tela-Sabio - Comunidades Constructivas Digital. Marcha de 2002
Disponible: www.imls.gov/pubs/forumframework.htm
13 Ibid.
14 Disponible: http://www.diglib.org/standards/imlsframe.htm

específicos y las buenas prácticas. "El DLF finalmente anima al IMLS que considere una enmienda modesta al marco. El marco acentúa colecciones digitales" [15]

En el informe de la reunión del DLF en la preservación que cambia prácticas de formato recomienda una "prueba patrón mínima" para los textos impresos convertidos formato digital. Se recomienda por su importancia, el análisis razonado y las implicaciones que se precisan para el material siguiente junto la indicación respectiva para su consideración.

Definir el "patrón digital para la preservación" –que se refiere al facsímil digital que representa fielmente al texto impreso, incluyendo ilustraciones y textos de impresión extraña–. Los patrones digitales de la preservación deberán tener metadata (los soportes) descriptivo, estructural –se refiere a las estructuras que configuran la información– y administrativo. Estos patrones digitales de la preservación incluyen el texto legible por el proceso:

OCR sin corregir

OCR corregido (exactitud: 99.995%)

Corrección –teclado u OCR– (exactitud: 99.95%)

El objetivo del patrón digital para la preservación es reducir el riesgo de error que implica la producción y mantenimiento de los materiales convertidos a digital, inspirando y animando así la confianza para su utilización. Su conservación será considerada como el objeto digital que anticipó y satisfizo las necesidades del futuro. Pese a todas estas ventajas, no es pensado para promover o definir métodos para crear copias digitales de reemplazo a los documentos de fuente (los originales).

15 Disponible: www.diglib.org/standards/presreformatsum.htm

Se debe tener responsabilidad y eficacia para poder preservar y conservar la herencia impresa. Tampoco es declarada como absoluta mejor práctica que asume que los métodos de numeración no mejorarán. Dentro de un análisis razonado se ponen en consideración las ilustraciones del libro y materiales impresos raros o viejos. Puede ser que se requieran diversas pruebas de patrones, contrario a los textos con edición actual. Dado a la naturaleza propia de los impresos raros se puede fijar una resolución mínima, aunque esto tenga como consecuencia un costo mayor.[16]

5. Nota metodológica sobre el proyecto

Los factores básicos a establecer: "1. grado y naturaleza de ilustraciones en el material de fuente que es convertido a digital. 2. uso previsto del amo digital de la preservación. 3. escala y costo del esfuerzo de numeración."[17]

En el caso de la selección de libros de Revueltas, se empezó con pruebas de escaneo, primero con copias de los libros para mayor facilidad en el manejo de las hojas. Después de buscar escaner adecuado, se encontró que el escáner HP tiene el programa de escaneo OCR exclusivo para textos. Con dicho escáner se compararon versiones primero con las copias de los libros y después con los originales. Después de varias pruebas, se comprobó que tomando los libros originales el resultado del escaneo era mejor que tomando como base las copias, entonces se decidió utilizar los libros originales con cuidado y al término de uso, devolverlos inmediatamente a la persona que posee el documento y que amablemente lo prestó para tal proceso.

16 Dr. Greenstein, 30 / mayo /2001, revisado el 30 / Julio / 2001
Disponible: http://www.diglib.org/standards/imlsframe.htm
17 Ibid.

Primero se guardaron las sesiones en la computadora Personal Computer (PC) marca Dell y finalmente se juntó un archivo que fue enviado como el reporte de un primer libro. De igual modo, los demás libros se escanearon y finalmente se enviaron por correo electrónico, donde se hicieron las correcciones ortográficas de presentación con una computadora Dell con el programa Word XP y nuevamente fueron enviados los archivos por Internet para que fueran grabados en disco compacto para ser revisados.

En los libros donde se escanearon copias por no contar con los originales, al comienzo de las correcciones –múltiples por la calidad de las copias– se rastreó y encontró un original. Entonces se procedió a escanear nuevamente con el libro original, evitando así mayor número de errores y por consiguiente, más tiempo en las correcciones y limpieza del texto.

Durante las correcciones se suprimieron las palabras truncadas por un guión de separación de línea causado por la edición. Esto fue con el fin de optimizar la búsqueda dentro del documento. Si la palabra "Revueltas" en las páginas 5, 46,78, 95 y 109 aparece con un guión por el cambio de renglón, y en diferente sílaba, en el momento de la búsqueda no la rastrearía la computadora a menos que le especifiquemos cada una de las posibilidades que el lugar del guión ocupa: Re-vueltas, Revuel-tas. Esto tendría como consecuencia que desde que se iniciara la búsqueda, no se contarían las cinco palabras truncadas de las páginas propuestas como ejemplo, alterando así el resultado. Con esto no se le quitó fidelidad al texto ni se le alteró el significado, así como tampoco se le aumentaron o quitaron palabras.

Para el cambio de página se hizo una especificación especial por medio de corchetes donde se incluía la página anterior y la siguiente: [6/7]. En el caso de las palabras trucadas con cambio de página, donde se encontraba el mayor número letras

de dicho término se encontraba el lugar de la palabra: así, en lugar de indicar Re-[6/7]vueltas, se señalaba [6/7] Revueltas.