DIGITALIZACIÓN DE DOCUMENTOS :

Una opción de preservación

Ángel Borràs

Trancripción de la cinta grabada durante la conferencia

Hola buenos días, durante esta intervención vamos a intentar aproximarnos a lo que es un proyecto de digitalización de documentos a efectos de conservar el contenido. Muchas veces se piensa que esto es un problema informático, y lo es hasta cierto punto. La base de la problemática que presenta la digitalización de documentos está en toda una serie de aspectos que son más documentales, de gestión. de previsión, que no son tan informáticos, es decir, tenemos que prever que vamos a digitalizar, pero para qué vamos a digitalizar.

Evidentemente, antes de iniciar un proyecto de digitalización, tenemos que saber qué vamos a obtener, quién va a utilizarlo, para qué va a utilizarlo y en qué condiciones se va a encontrar en el momento en que utiliza este sistema; por ejemplo, no es lo mismo digitalizar una página de una revista para volver a publicar impresa esa misma página de revista, no es lo mismo digitalizarla para que alguien tenga que leerla en pantalla, no es lo mismo digitalizarla a efectos de que podamos obtener una reproducción de calidad de fotocopia, y no es lo mismo digitalizarla si lo que queremos hacer es obtener un registro de base de datos.

Entonces, antes de empezar, tenemos que saber para qué vamos a digitalizar, quién va a utilizar este sistema que va a contener información digitalizada, y en el momento en que esta persona, este usuario, esté utilizándolo, en qué situación se va a encontrar, va a estar en la calle?, va a estar sentado en una biblioteca?, va a estar en su despacho?, va a estar en su casa conectado en línea? ..., y esto tenemos que tenerlo claro porque va a determinar cómo vamos a tener que proceder.

Por tanto, podríamos hablar de una fase preliminar que es establecer unos objetivos del sistema, y entre estos objetivos del sistema tenemos que analizar cuáles van a ser los usuarios potenciales de este sistema, pero no un análisis superficial, es decir "ah no, estos van a ser el público de mi biblioteca", qué público? van a ser profesores que van a

estar utilizando un sistema en su despacho, porqué un profesor querría utilizar mi sistema digitalizado y no la revista original, o el libro original, o el plano original, o el esquema, o el dibujo, o la litografía. Entonces, tenemos que tenerlo claro. A menudo se llega a la conclusión de que se necesita digitalizar por un problema del día a día, tal vez estamos en una biblioteca, tenemos una colección de revistas que es valiosa y tenemos decenas de alumnos que cada día llegan y nos la piden para fotocopiar una fotografía para incluirla en un trabajo y observamos que con esa manipulación constante se va degradando la revista y acaba por estropearse. Entonces pensamos "ah, ostras¡ podría digitalizar esta colección de revistas, así tengo un sistema informático y cuando los alumnos me piden fotocopias pues que lo saquen impreso desde el ordenador" Es una aproximación correcta, lo que pasa es que desde un punto de vista informático esto va a ser muy costoso. Compensa hacer esta digitalización solo para que determinados alumnos saquen fotocopias? Es un problema que nos tenemos que plantear. Quién más utiliza esas revistas? Investigadores, gente que es ajena a nuestro entorno, es decir, tenemos que prever si nos vamos a limitar a una funcionalidad específica o si vamos a poder dar una serie de funcionalidades tanto propias como para otra gente. Esto es quizá la parte más difícil, no nos engañemos, si nosotros nos limitamos o no hemos estudiado suficiente a nuestros usuarios o no hemos estudiado suficiente a la colección y nos gastamos o hacemos una inversión de recursos en obtener una digitalización que después no nos va a servir, vamos a tener un problema. Y al decir que no nos va a servir significa que los usuarios no lo utilicen para aquello por lo que lo hemos programado o por lo que lo hemos decidido digitalizar, si una vez digitalizada esa colección de revistas los alumnos continúan prefiriendo acceder a la revista original significa que esa digitalización que hemos hecho no es correcta. Bien porque no tiene la calidad suficiente, bien porque el interfaz de consulta no es el adecuado, bien porque el tiempo de respuesta no es el correcto.

Vamos a tener un problema, primero porque la empresa nos dirá "oye, es que me habías dicho trescientas" y tú "ah bueno es que no se qué", y al final vas a tener una colección incompleta, un proceso quizá no del todo correcto, etc. por tanto vamos a delimitarlo muy bien, también para poder calcular los costes económicos. Muy bien, ya sabemos qué partes de la colección exactamente son las que queremos digitalizar, después tenemos que saber quién las tiene, es decir, quién posee la custodia de estas colecciones o parte de colecciones que vamos a tratar, es decir, tal vez aquellos documentos tan

valiosos que queremos escanear, digitalizar, están en la sala noble del rectorado de no sé qué y la gestión de esta colección administrativamente no es de la biblioteca sino que es de no se quién, bueno, este tipo de cosas tenemos que saberlas inmediatamente después de saber si un documento vamos a digitalizarlo o no. Quién es el propietario de la gestión de ese documento? Cuando sepamos quién es el propietario, a lo mejor es nuestra biblioteca y no tenemos que hacer ningún trámite, pero es importante porque, por ejemplo, se han dado casos en los que se pretende digitalizar un incunable, el incunable digamos está en la biblioteca, quien va a hacer el proceso de digitalización coge el incunable y quiere sacarlo de la biblioteca, y el responsable político de la entidad le dice "no, este incunable no sale de aquí" o "si sale tiene que hacerse un seguro", o "si sale tiene que hacerse en determinadas condiciones de transporte"; por tanto, todo esto tenemos que preverlo antes.

Podemos encontrarnos con planos enormes que si no se manipulan de una determinada manera se pueden estropear, manipularlos cuesta dinero, porque quizás tengan que embalarse de una manera especial para poder sacarlo de la biblioteca, y no puede ser que nos encontremos con este problema en el momento en que vamos a sacarlo y que alguien nos diga "no, esto no se puede hacer así" porque entonces esto tendrá implicados unos costes que no habremos previsto.

Muy bien, supongamos que ya tenemos nuestro mapa de la colección, sabemos a quién pertenece cada uno de los componentes de este mapa, quién es el encargado de la gestión fisica de cada una de estas partes de la colección y si la manipulación de estas colecciones fisicas tiene o requiere algún tratamiento especial.

Una vez ya hemos determinado exactamente qué vamos a hacer, tenemos que plantearnos, irnos por un momento al otro extremo, y decir, bueno, cuando lo tenga digitalizado de hecho voy a tener un montón de archivos, cómo voy a hacer que el usuario lo consulte? El producto de esta digitalización lo voy a poner en una base de datos? Lo voy a poner en Internet? Voy a ponerlo en un disco duro y que la gente me pida el documento y yo ya me espabilaré para encontrarlo? Claro, esta pregunta como antes hemos hecho un análisis de usuarios, sabemos cómo el usuario quiere explotarla, por tanto tendremos que determinar el modo de consulta, es decir, si esto va tener que ir a una base de datos tendremos que buscar una base de datos que pueda soportarlo o

alguien que la fabrique, o un producto comercial que sea capaz de gestionarlo. No nos podemos encontrar al final que tenemos un montón de documentos digitalizados y que no sabemos qué hacer con ellos. Entonces no podemos improvisar, en el momento en que tengamos todos los ficheros informáticos tenemos que tener una aplicación para poder gestionarlos.

Muy bien pues, ya sabemos algo más. Una vez ya sabemos cómo vamos a gestionarlo, lo que tenemos que hacer es prever las manipulaciones que van a requerir cada uno de los documentos para poder ser insertados en el sistema informático, me explico, dentro de esta colección que hemos determinado como objeto de nuestro sistema, seguro que no todos los documentos son uniformes, me explico, podemos tener diferentes formatos, podemos tener diferentes texturas, diferentes materiales, diferentes grados de conservación, evidentemente los requisitos para poder escanear un mapa mural no son los mismos que los requisitos para escanear un sello de correos, son documentos los dos con un elevado contenido gráfico, los dos tal vez con el mismo número de colores, a lo mejor los dos están fabricados con el mismo material, pero si queremos escanear el sello de correos nos encontraremos con unos problemas diferentes que si queremos escanear el mapa mural. Por tanto tenemos que determinar todas las posibles variantes de formatos que nos vamos a encontrar, de materiales, de grados de conservación, más que nada para que podamos después agruparlos, es decir clasificarlos para poder manipularlos en bloque, todos los documentos que son Din A-3 los ponemos juntos porque entonces probablemente podremos escanearlos en bloque, todos los documentos que son Din A-0 en otro, además porque a lo mejor tiene que hacerlo personas diferentes, para escanear un Din A-0 necesitas un escáner enorme y para el Din A-3 tienes tú un escáner en tu biblioteca.

Después, evidentemente, a la que queremos separar por formatos nos podemos encontrar con problemas de encuadernación, es decir, documentos heterogéneos que están compuestos por diferentes formatos, puede ser que un documento sea un grupo de hojas grapadas, más un plano, más un esquema, más diecisiete fotografías, claro, si vamos a clasificar por formatos probablemente tendremos que romper esta unidad documental y coger fotografías que son parte de un documento y ponerlas con fotografías que son documentos per se, el plano que forma parte de este documento lo ponemos con el resto de planos..., evidentemente tenemos que establecer los parámetros

para poder reconstruir de nuevo estos documentos una vez ha finalizado el proceso de digitalización. Por tanto este análisis detallado de los documentos físicos que tenemos es muy importante.

Otro criterio a tener en cuenta además del formato físico es el grado de legibilidad de estos documentos, si nosotros vamos a obtener una imagen digital de un documento que no se puede leer ya el original, pues mejor no la obtenemos porque no vamos a poder hacer nada con ello. O tal vez si tenemos documentos que tienen un fondo azul y son letras blancas sobre fondo azul pues seguro que requieren un tratamiento especial y tenemos que separarlos. Por tanto, estas agrupaciones lo que tienen que buscar es la homogeneidad en todos los sentidos, cuanto más parecidos sean los documentos dentro de una agrupación más fácil será procesarlos en bloque, por tanto más fácil o más barato será obtener la digitalización.

Bien, después evidentemente un aspecto que tenemos que considerar también en esta primera fase es el tema del Copyright. Vamos a poder digitalizar para aquel objetivo que nos habíamos propuesto sin problemas? es decir, probablemente si lo que queremos es escanear o digitalizar una revista para hacer un producto en línea a través de Internet, la editorial probablemente nos lo va a prohibir, si queremos digitalizar para dar un servicio de copias a nuestros usuarios a lo mejor la editorial nos da permiso, pero estos permisos tenemos que tenerlos claros antes de empezar el proyecto, para todos y cada uno de los tipos documentales que vayamos a tratar y para todos y cada uno de los formatos que vayamos a tratar. Más que nada porque sino, una vez hecho todo el proceso nos pueden prohibir utilizar el sistema y por tanto habremos tirado todo el dinero.

Bien, fijaos que hasta ahora de informática no hemos tocado nada, es un tema de planificación y previsión y es un tema bibliotecario porque nadie más que una persona que está al cargo de un centro de documentación puede establecer el valor que tienen estos documentos.

Muy bien, una vez tenemos clara esta primera parte, y esto significa hacerlo todo por escrito, no vale decir bueno esto ya lo miraré más tarde, hay que tener las cosas claras antes de empezar. Muy bien, ya que sabemos qué vamos a hacer y sobre qué lo vamos a

hacer, vamos a hacer un plan de aproximación a esta digitalización. Primero, tenemos que hacer un plan de conservación. Si nosotros queremos digitalizar para evitar que se deterioren unos documentos no sirve decir yo digitalizo y después los documentos que se estropeen, tenemos que establecer estos parámetros de conservación física de los documentos que vamos a digitalizar, más que nada porque a lo mejor durante el proceso de digitalización también se deterioran y es posible que un documento muy valioso y muy delicado después del proceso de digitalización tenga que restaurarse total o parcialmente. Quizás un grupo de revistas para poder ser escaneadas con calidad tienen que desencuadernarse y después hay que volver a encuadernarlas. Es decir, si tenéis revistas en volúmenes encuadernados y al hacer una fotocopia el libro no se puede abrir bien y sale mal la fotocopia, pues si se hace en un escáner pasa exactamente lo mismo, por tanto si queremos que tenga la calidad suficiente puede ser que se tenga que desencuadernar la revista pero a lo mejor no se puede desencuadernar por un tema de política de conservación de la biblioteca, entonces todo este plan de preservación física hay que establecerlo antes de empezar.

-....?

-mira si, explico, para escanear diferentes tipos de materiales, por ejemplo las transparencias, digamos que podríamos establecer dos tipos de materiales: unos que son reflectivos, es decir cualquier material que refleje la luz que para leerlo utilizamos la luz reflejada, y los transmisivos, que son las transparencias, aquellos que para poder leerlos tenemos que utilizar la transmisión de la luz. Los tipos de escáner para diferentes tipos de documento tienen que prever el comportamiento de este documento respecto a la luz. Para las transparencias son escáneres especiales, el problema que tienen los escáneres de transparencias es que no pueden automatizar, el proceso de carga no puede ser muy rápido porque el calor del escáner puede estropear el material. Entonces habitualmente son escáneres fríos, digamos que tienen una protección o lo que sea para evitar que el plástico se estropee. El tema este de los tipos de escáner que vamos a utilizar en función del documento lo consideraremos un poco más adelante, aun nos falta saber muchas cosas respecto a nuestra colección antes de poder decidir con qué escáner vamos a hacerlo.

Como ya sabemos toda una serie de datos y vamos a enfrentarnos a una manipulación física de cada uno de esos tipos documentales tenemos que establecer unos parámetros de manipulación. Quién va a coger y manipular un determinado documento y qué

normas tiene que seguir para poder manipularlo. Tenemos que establecer una serie de recomendaciones que indiquen cuál es la deteriorabilidad de ese documento, qué factores de manipulación pueden alterarlo. Por ejemplo una cosa que ya ha salido es, si yo tengo un documento plástico, como recomendaciones puedo establecer "no someter a temperaturas elevadas", puedo establecer no doblar el documento, manipular siempre sobre soporte rígido. Y este tipo de recomendaciones tengo que hacerlo para todos los tipos documentales. Por ejemplo si son negativos, no tocarlos con los dedos, no ensuciarlos, no desencuadernarlos. Si son mapas a lo mejor no plegarlos; es decir, todas las condiciones de manipulabilidad de los documentos tienen que estar escritas en esto que será el manual de actuación.

Ya que sabemos cómo vamos a manipular estos documentos, habrán emergido una serie de restricciones a la hora de digitalizarlos, es decir, tendremos que buscar el dispositivo adecuado para digitalizar estos documentos. Evidentemente hay muchos tipos de escáneres, por ejemplo hay escáneres para diapositivas y negativos, por tanto todas las diapositivas y negativos tendrán que procesarse por este tipo de escáner, tendremos que buscar por tanto alguien que pueda hacer todo este proceso. Si tenemos medios transparentes tendremos que buscar un escáner de transparencias, si tenemos formatos muy grandes tendremos que buscar alguien que pueda disponer de un escáner de ese formato. Antes, por tanto, de empezar a hacerlo físicamente, para cada uno de los formatos tenemos que haber seleccionado quién, cómo y cuándo lo va a hacer, eso implicará a menudo pedir ya un presupuesto.

Como ya sabemos quién lo va a hacer, cómo lo va a hacer, con qué restricciones lo va a hacer, sobre qué documentos lo va a hacer, vamos a empezar a determinar una serie de cosas como es que vamos a dimensionar el dispositivo informático para que pueda sostener esta colección digitalizada. Para poder dimensionar este dispositivo informático tenemos que tener en cuenta también varios factores. Uno evidentemente es el número de usuarios que van a estar utilizando el sistema. El segundo es qué espacio de almacenamiento van a ocupar las imágenes que vamos a obtener.

- ...?

-es capacidad de disco del ordenador para poder soportar la colección digitalizada. Para el cálculo de capacidades hay varios factores que afectan al tamaño físico que va a ocupar un archivo de imagen. Entre todos los aspectos a considerar para el cálculo de

espacio tendremos que ver evidentemente número y tamaño de documentos. Evidentemente la imagen de un documento de un metro cuadrado va a ocupar más espacio que la imagen de un documento de un palmo cuadrado. Otro aspecto que hay que tener en cuenta es la resolución de almacenamiento, esta resolución de almacenamiento la vamos va a determinar nosotros a priori según la función que asignemos a este documento digital. Una fotografía escaseada para ser publicada en una revista necesita una resolución de unos tres mil puntos por pulgada, si esta imagen es una diagnosis médica a lo mejor necesitaremos una resolución de unos quince mil puntos por pulgada, si esta imagen es de calidad de fotocopia a lo mejor es 150 puntos por pulgada. Las diferencias son brutales, por tanto esto lo determinaremos a partir de aquel esquema preliminar donde dictábamos las funcionalidades del sistema. Lo que está claro es que si nosotros escandamos para publicación vamos a consumir muchos más recursos que si escandamos solo a efectos de lectura en pantalla o fotocopia, pero del orden de treinta veces más, por eso era importante antes tenerlo muy claro. Digamos que el número de puntos por pulgada lo que viene a decir es dividir el documento en una matriz de puntos, por ejemplo imaginamos que tengo una letra "o" y que esto es una pulgada, si yo escojo dos puntos por pulgada dividiría esto en una matriz de dos puntos y entonces el almacén digital de esta "o" sería: primer punto, hay imagen? si. Segundo punto, hay imagen? si. Tercer punto, hay imagen? si. Cuarto punto, hay imagen? si. Es decir, tendría un cuadrado negro, evidentemente esta resolución es insuficiente.

- -...?

-Claro, tienes que definirlo a priori. Más que nada porque determinados escáneres no soportan determinados grados de resolución, por tanto tenemos que saberlo a priori. Es decir, cuanta más sea la resolución más se aproxima la imagen al objeto real. Si esta letra "o" que estaba en un documento de una pulgada cuadrada la dividimos en una matriz de quince mil puntos evidentemente el resultado será absolutamente fiel al original. A medida que disminuyamos la resolución la imagen se irá pareciendo menos a la original, entonces es posible, digamos esta "o" a mucha resolución sería así, a menos resolución veríamos puntitos, a menos resolución veríamos esto y a menos resolución... claro entonces a mayor calidad más resolución, también más nos va a ocupar en el disco, porque cada punto va a ocupar un espacio en disco, por tanto a más puntos más espacio.

Es decir, a una resolución de mil puntos por pulgada esto sería un millón de puntos cuadrados, a una resolución de cien puntos por pulgada esto contendría diez mil puntos, la diferencia es muy considerable. A parte de lo que es la resolución tenemos que tener en cuenta el número de colores en el que vamos a almacenar esta información. Evidentemente si tenemos que almacenar información de uno o dos colores el espacio requerido será menor que si tenemos que almacenar 256 colores o si tenemos que almacenar 16 millones de colores o si tenemos que almacenar 255 millones de colores.

Cuando estamos trabajando en blanco y negro tendremos una ocupación informática de un bit por punto, si estamos en blanco y negro nos va a ocupar ocho veces menos que si está en color, por tanto tendremos no es trivial decidir si un documento lo vamos a escanear en color o si lo vamos a escanear en blanco y negro, más que nada si lo que queremos es obtener fotocopias a lo mejor no tenemos que penalizar el sistema escaneando ese documento en color, si lo que queremos es una lectura artística del gráfico a lo mejor tenemos que forzar el máximo número de colores, por tanto intentar determinar el tipo de tratamiento para cada tipo documental y para cada funcionalidad.

Después del espacio de almacenamiento y del grado de número de colores hay un aspecto puramente informático que es un parámetro de compresión. Este parámetro de compresión nos interesa sólo desde un punto de vista, de los miles de formatos o de mecanismos de compresión hay unos que son estándar y otros que no son estándar, por tanto tendremos desde un punto de vista de gestión documental que ir siempre a un formato de compresión que sea un estándar internacional. Entre estos estándares internacionales nos encontramos con el formato CITT grupo fax-3 y 4 para imágenes de color y después un formato que es JPEG que es del Joint Picture Experts Group que es un estandarizador de facto pero que son las mayores empresas de gestión informática de imágenes que se ha asociado y han establecido este estándar. Cualquier compresión que no siga estos estándares de momento no nos va a poder asegurar el mantenimiento en el futuro, no se si me explico, si nosotros estamos trabajando con estándares seguro que cuando dentro de cinco años el sistema informático se haya quedado obsoleto encontraremos un sistema informático que sea capaz de manipular estas imágenes, si vamos a un formato que no es estándar a lo mejor nos damos cuenta al cabo de cinco años cuando el sistema que tenemos se queda obsoleto que nos tenemos que tragar los ficheros informáticos porque nadie sabe cómo tratarlos, y esto no es anecdótico,

organismos de la administración que habían hecho inversiones de decenas de millones de pesetas en digitalización antes de que salieran estos estándares se han encontrado que han tenido que coger todo el sistema y tirarlo a la basura porque había desaparecido la empresa que había creado el sistema y nadie sabía cómo manipular los documentos, por tanto vamos a restringirnos siempre a formatos y grupos de manipulación informática estándar.

- ...?

-en formato grupo fax 4 un Din A-4 blanco y negro, es decir, dos colores, blanco y negro significa no escala de grises sino blanco y negro. Te puede ocupar entre veinte y treinta Ks, siendo blanco y negro, a la que quieres escala de grises se puede multiplicar por cuatro la ocupación en disco.

Lo más importante de lo que hemos visto hasta ahora es que tenemos las cosas claras, a partir de ahora podemos buscar, ir al proveedor y decirle exactamente qué es lo que queremos, cómo queremos que lo haga y cómo poder controlar la calidad de ese trabajo, más que nada porque sabemos muy bien cómo debe hacerse. A menudo se trivializa el tema de la digitalización y se dice, pues mira sabes que, nos vamos a comprar un escáner entre dos o tres, vamos a ponerlo y vamos a poner una persona a escanear. Trabajar de esta manera, como precisamente no te has planteado toda esta serie de cosas previamente, pues a menudo lo que se convierte es en haber tirado el dinero y en no obtener unos resultados correctos. También el hecho de no haber planificado todo este tipo de cosas no te permite iniciar proyectos de cooperación, más que nada porque, qué le vas a decir a otro centro al que le estas pidiendo cooperación: oye ayúdame a escanear, pero qué, cómo, para qué, qué esfuerzos nos vamos a dividir, cómo nos los vamos a dividir. Por tanto digamos que, como conclusión, estas dos primeras fases que no son informáticas del proyecto son las más importantes y las que no se pueden eliminar o dejar de lado. Después lo que queda ya es un tema puramente informático. A lo mejor cuando hemos acabado estas dos fases nos damos cuenta de que el proyecto no es viable, pues entonces lo mejor es no iniciarlo, o que el proyecto es viable con una serie de recursos, pues vamos a buscar los recursos, también nos permitirá planificarlo en el tiempo. Bueno, siento que me haya desviado un poco del plan inicial. ....

NOTA DE LA TRANSCRIPCIÓN: Hemos optado por incluir la transcripción de esta conferencia, a pesar de que haya algún vacío en el discurso así como la parte final, que debido a problemas técnicos, no quedó registrada