Recomendaciones para la digitalización de Documentación manuscrita. Creación, conservación y difusión de Archivos digitales

Bergara, 27 de marzo de 2007

0. Introducción

Los objetivos de estas Recomendaciones son:

a) normalizar la terminologia
b) proponer una serie de parámetros técnicos
c) presentar las normas tanto a un público conocedor como al no experto

d) describir un protocolo de actuación que abarque el proceso de creacion de Registros/Archivos digitales, incluyendo de manera diferenciada la creación de un registro o imagen estandarizada original, el Archivo Digital Original, y la creación de un Registro de segunda generación a partir de este original, para su difusión (en copias facilmente manejables o en web).

Las Recomendaciones siguen este esquema:

1. descripción de los aspectos técnicos básicos del proceso de digitalización
2. proceso de digitalización
3. difusión

1. Conceptos digitales básicos

1.1. Imagen digital

La diferencia entre una reproducción fotográfica ordinaria (analógica) y una reproducción fotografica digital estriba en que la primera obtiene la imagen sobre una emulsión química fotosensible (película fotográfica), mientras que la reproducción digital captura una imagen de formato electronico por medio de un sensor y construye su representación mediante una cadena de bits. Esta cadena de bits, interpretada por un ordenador, presenta una reproducción de la imagen en pantalla.

El Sensor es un dispositivo electrónico compuesto por un soporte cubierto de Fotodios[1] sensibles a la luz, sobre el que se proyecta la imagen. La imagen analógica, obtenida por escaneo, es dividida en una matriz de puntos a modo de cuadricula, tantos como Fotodiodos tenga el Sensor. Cada uno de estos puntos recibe el nombre de Píxel[2], que toma el valor binario 1 ó 0 dependiendo de la luminosidad y el tono lumínico leido por el escaner.

Esta cadena de código binario es enviada al ordenador, donde la imagen queda almacenada en formato digital.

1.2. Resolución digital

Es la capacidad de reproducir los detalles con precisión. Se mide por el número de pixels que lee el escaner por pulgada. Cuantos más pixels pueda

[1] Fotodiodo es un dispositivo electrónico que cuando recibe luz genera una corriente eléctrica de una magnitud acorde a la cantidad de luz recibida.

[2] Término formado de la mezcla de los términos ingleses picture+element. Es el elemento más pequeño de que se compone una imagen digital.

leer en una pulgada (dpi=ppi=ppp= puntos por pulgada), el escaner tendrá mayor capacidad para capturar los detalles del original manuscrito y, en consecuencia, la imagen final tendrá mayor resolución.

1.3. Profundidad de bits

Como se ha visto en el apartado 1.1, el resultado del escaneo se transforma en una matriz (cuadrícula) de datos. A cada cuadrícula le corresponde un píxel, y a cada pixel un bit, bien de valor 0 bien de valor 1.

Este proceso es el estandar cuando trabajamos con un documento original en blanco y negro. Cambia sustancialmente cuando trabajamos con originales con escala de grises o a pleno color. En este caso, basta con asignar más bits a cada píxel, de forma que sea posible reproducir cualquier color.

De este modo, para una escala de grises aplicaremos 8 bits a cada píxel, de modo que la imagen digital resultante sea capaz de representar 256 valores o tonos de grises.

En el caso de las imágenes en color, utilizando 24 bits se obtendrán 16,7 millones de colores.

Esta capacidad de reproducción de colores se conoce como **profundidad de bits**.

1.4. Registro y Archivo digital

El proceso de escaneado tiene como primer objetivo la creación de un Registro o Archivo de copia en soporte digital, que está compuesto por la secuencia de las paginas que incluye el documento o expediente original.

Los registros digitales tienen la consideración de Archivo Digital Original a todos los efectos y se prescribe su conservación indefinida como tales con el objeto de integrar la Memoria Digital.

Si bien su objeto no es sustituir al documento original, poseen varias ventajas añadidas. Entre las principales: minimizan el deterioro por el uso de la copia y facilitan la difusión a través de copias digitales, bien in situ o bien en web

1.5. Formatos

Existen diferentes formatos en el mercado. Cada uno con sus características específicas en relación a su capacidad para soportar distintas profundidades de bits, de colores, de aceptar o no distintas técnicas de compresión, etc.

Nos interesan estos dos:

TIFF : Es un formato de registro/archivo digital que acepta la compresión sin pérdidas, por lo que conservamos la información original completa. Se trata de un formato estándar y de uso libre, y es el más utilizado como archivo de conservación. Su principal inconvenitne es el alto volumen de almacenamiento motivado por el tamaño del archivo que genera

JPEG y PDF: estos formatos permiten distintos grados de compresión, pero ésta siempre es con pérdidas. Son muy adecuados para distribuir información, es un estándar muy extendido y mantienen una calidad muy buena. El PDf es Perfecto para imprimir y crear documentos secuenciales con páginas múltiples.

1.6. Tamaño del Registro digital original y niveles de compresión

El tamaño del Registro Digital depende del tamaño del original, de los dpi y de la profundidad de bits que se elijan.

A partir de una calidad de imagen suficiente, no es necesario, y consideramos, además, desaconsejable, crear registros de mayor tamaño. Este mayor tamaño, tradicionalmente considerado garantía de calidad, es el origen de la creación de registros digitales inmanejables desde el punto de vista informático.

Se han desarrollado, por este motivo, técnicas de compresión, basadas todas ellas en abreviar matemáticamente, por medio del uso de algoritmos muy elaborados, la cadena de código binario de la imagen.

La compresión puede ser:

1. *Sin perdida*: los sistemas sin perdidas abrevian el codigo binario sin desechar información; cuando se descomprime la imagen se reproduce bit a bit, como la original.
   No se suelen conseguir grandes reducciones de tamaño de archivo en imágenes tonales, pero sí es muy utilizada en escaneados de imágenes bitonales, ya que en esas circunstancias una orden de secuencia de bits iguales es mucho més corta que la secuencia de los mismos.
2. *Con pérdida*: las técnicas de compresión con pérdidas (como JPEG) utilizan un sistema para eliminar la información menos relevante basándose en la percepción visual del ojo humano.
   Esta técnica ha sido optimizada hasta tal grado que, dependiendo del porcentaje de compresión, es muy difícil detectar sus, de modo que puede considerarse una imagen “sin pérdida visual”.
   Un documento tamaño DIN A4, escaneado a 300 dpi, en formato TIF sin comprimir ocupa 25 MB, y en formato JPEG comprimido ocupa 1,3 MB. Sin embargo, es difícil distinguirlos visualmente. La percepción visual es

la misma, pero cuando se trata informáticamente el archivo, el formato JPEG acusará la perdida de calidad que ha sufrido.

2. Creación del Registro/Archivo digital

La creación de un registro/archivo digital original es la base del proceso.

Se conservará como **archivo de seguridad**, por lo que se debe definir un escaneado con el nivel de calidad suficiente para los procesos posteriores a los que se vaya a someter. A partir de este archivo se puede rebajar la calidad (resolución, profundidad de bits, etc.), pero será imposible mejorarla.

2.1. Definición de los parámetros de calidad de escaneado.

2.1.1. Clase de documento y profundidad de bits.

A) Letra impresa/dibujos: compuestos por lineas o masas de blanco y negro sin tonos intermedios y con bordes definidos.

B) Imágenes tramadas: son imágenes reproducidas sobre todo en publicaciones económicas (periodicos, boletines, anuncios, etc.), en las que la reproducción se realiza por medio de cuadriculas o tramas de distintas formas y tamaños que simulan la gama tonal de la imagen.

C) Tono contínuo: Reproducciones fotográficas, pinturas, dibujos de lineas que contienen trazos degradados y manuscritos. Se incluyen en este apartado los manuscritos porque, como las pinturas, presentan:

- una amplia variedad tonal de colores, tintas y degradados.
- trazos de borde suaves y poco definidos.
- decoloraciones de tinta y papel a causa de su antigüedad o de las inadecuadas condiciones de conservación.

D) Mezclado: documentos que contengan dos o más de los casos precedentes (por ejemplo un libro con fotografías).

A priori, la decisión parece facil. Los documentos del tipo **A** (letra impresa y dibujos en blanco y negro), y tipo **B,** (imágenes tramadas, que solamente tienen dos tonos), se escanean en blanco/negro (: bitonal 1 bit), un bit para el blanco y otro para el negro.

En circunstancias ideales (libro impreso con clara tipografía y papel sin manchas ni trama), se puede realizar un escaneo en blanco y negro bitonal,

que permite una fuerte compresión sin pérdida de calidad, obteniéndose un archivo de tamaño pequeño que reproducirá el documento con toda fidelidad.

Sería así mismo válido para un manuscrito en estas mismas condiciones (papel limpio, letra clara y de trazo bien definida, densidad de tinta constante a lo largo de todo el trazo del carácter). Ahora bien, ¿qué sucede si la tipografía no es de calidad suficiente, o hay falta de tinta en algunos tramos de trazo, o mala calidad en los tipos de imprenta (o la trama del papel no la define perfectamente, o existen manchas de diferente densidad en el documento) y, además, como sucede en el caso de los manuscritos, la densidad del trazo no es homogénea?.

Debe tenerse en cuenta cómo opera técnicamente el sensor: en los parámetros de digitalización bitonal, existe un punto denominado umbral. Este define el punto en una escala de 0 a 256, en el que los valores grises capturados se convierten en pixeles blancos o negros. Por ello, toda la información de valor tonal intermedia distinta de 0 ó 256 desaparecerá añadida a ambos extremos.

En estos casos deberá realizarse una digitalización de tono continuo, sea en escala de grises o color ya que de otra forma se perderá gran parte de la información contenida en el documento y en algunos casos puede llegar a ser ilegible el texto.

Dada la insuperable calidad de lectura, así como la información adicional que ofrece el color (tonos y testura del papel, de las tintas, colores de manchas, etc.), **toda la documentación histórica se debe de considerar como documentos de tono continuo** y recomendamos realizar los escaneos a una profundidad de color de 24 bits.

2.1.2. Resolución

No existe una norma (ISO, UNI, ANSA/NMA, DOD, BSI, AFNOR, DIN, UNE) que regule los parámetros de resolución en la digitalización de imágenes.

Se han definido pautas centradas en los requisitos de resolución para textos impresos, en los que se relacionan la calidad con el tamaño de carácter, basándose en las normas del microfilm. En otros casos, se toman como referencia la calidad mínima necesaria para reproducir el ancho de trazo. Como los manuscritos, mapas, dibujos o grabados no ofrecen parámetros estables, se suele defender la mayor calidad para la reproducción de los trazos.

Irargi ha desarrollado un método para aplicar la resolución correcta a cada documento, basado en la utilización de una “Mira de legilibidad” [Descrita en la norma DIN 19 051, que desarrolla el método operativo completo.]. Mira de

legibilidad que es utilizada para ensayos de legibilidad de caracteres tipográficos en la microfilmación de documentos y cuyo funcionamiento elemental es el siguiente.

La Mira de legibilidad se compone de grupos de signos creados específicamente para esta función, en grados de disminución calibrada, que, una vez digitalizados, nos permiten confirmar el tamaño de carácter más pequeño legible a una resolución determinada (para dar por legible un grupo de signos, se debe poder identificar, por lo menos, las bandas de orientación de ocho signos de entre los que se compone el grupo, ya sea a simple vista o aumentando su tamaño en pantalla).

Por ejemplo, si digitalizamos la tarjeta a 200 dpi., y vemos que en el grupo marcado como 120 podemos identificar la orientación de las bandas de por lo menos ocho de los signos del grupo, se entiende que a esa resolución serán legibles todos los caracteres con ese tamaño y, en mayor medida, por supuesto, los tamaños mayores.

La letra manuscrita no posee la naturaleza de los signos de la Mira de legibilidad (trazo, borde y densidad muy definidos), el trazo puede tener diferente grosor, el borde puede estar diluido y la densidad puede variar dentro del mismo trazo; por tanto, los resultados no serán tan precisos.

Se aconseja realizar ensayos y utilizar la Mira de legibilidad como referencia, para decidir la resolución más idónea.

En cualquier caso, la resolución escogida debe ser capaz de posibilitar la lectura de la letra más pequeña del documento, ya sea manuscrito o impreso.

La **resolución** para archivos de conservación con la que opera Irargi y que recomendamos es de **300 dpi**. Si la naturaleza de algunos documentos concretos lo requiere se corrige al alza.

2.2. Registro Digital original. Definición de calidad, formato y soporte.

2.2.1. Calidad y formato

El Registro Digital Original, es decir, la copia hecha directamente mediante la digitalización del documento original, se considera un Archivo de Conservación indefinida. Tiene dos funciones:

- primaria: conservación permanente
- secundaria: soporte de las copias de difusión. De él se genera una copia Master, de la que se obtendrán todas las copias posteriores.

La calidad, formato y soporte de este Registro Original no está regulada por norma alguna. Al digitalizar el documento original decidiremos la calidad y el formato en que guardaremos el Registro/Archivo Digital que vamos a producir.

Se recomienda que este Registro se cree ajustándose a **los siguientes criterios:**

a) nivel de calidad alto, ya que siempre es factible generar a partir de éste otros archivos con calidades inferiores, mientras que el proceso a la inversa es imposible.
b) Concepción global de los usos: el criterio de archivo original, como ya hemos señalado, responde unicamente a una visión convencional centrada en la conservación y olvida el uso posterior de esta imagen, que hay que tener siempre presente.
c) Ajuste de calidades. Combinados ambos criterios (conservacion + difusión) asumiremos un matiz muy importante: se almacenará con la calidad adecuada.

Por ello, si el proyecto incluye una o varias calidades de archivo se debe optar por la mejor de las utilizadas en el mismo. Todo lo que se realice por encima no tendrá utilidad alguna y será una pérdida de esfuerzo y medios.

El espacio requerido para guardar el archivo con una imagen digitalizada de un documento manuscrito de tamaño estándar (A4/folio), escaneado a 300 dpi de resolución y en color con profundidad de 24 bits, puede llegar a ser 25/30 MB. Se trata de un tamaño de archivo muy voluminoso[3] y no apto para distribución masiva de información. Los equipos de lectura deberían tener muy altas prestaciones para manejar con fluidez estos tamaños de archivo y el DVD [a fecha de hoy] solo tendría capacidad para 140 imágenes de documentos.

Es necesario utilizar un formato de archivo que permita tasas de comprensión altas. Si es preciso, habrá que utilizar una compresión con pérdidas, pero que, en cualquier caso, no comprometa la legibilidad y capacidad de impresión de la información, y que aproveche las capacidades de los soportes y permita una gestión informática agil.

Tenemos dos posibilidades:
Archivos de conservación sin compresión.
Se guarda el Registro original completo generado por el sensor. Es aconsejable un formato estandar, de uso libre y sin compresión. Si se comprime algo, debe de ser del tipo de compresión sin pérdidas de información, de modo que se asegure la calidad de reproducción del original. El estándar de libre uso para archivo de conservación más extendido es el formato TIFF, sin comprimir.

Archivos de conservación con compresión.

[3] Si se quiere guardar un archivo con 70.000 páginas [que correspondería a un fondo de aprox. 35 cajas de documentación] en tamaño DIN A4, escaneado a 300 ppp en color (archivos de tamaño +/- 25 Mb por página), con un formato sin compresión, serían necesarios 500 DVD. Aparte del costo económico, debe tenerse en cuenta el tiempo empleado en la grabación y etiquetado de los DVD, así como el volumen de material originado por éstos, (el espacio que ocupan los DVD puede llegar a superar el de la documentación original). Este mismo archivo guardado con un formato comprimido cabría en 20 DVD sin aparente pérdida de calidad visual.

El Registro original se guarda con un formato que puede soportar distintos grados de compresión. Una vez decidida la compresión adecuada a la calidad final que se desea obtener, se guarda el archivo para conservación.

Se recomienda un formato de archivo estándar y de libre uso. Los más extendidos son el JPEG y PDF. Se trata de formatos muy optimizados y, de no utilizarse compresiones muy altas, visualmente es difícil diferenciarlos de los no comprimidos.

2.2.2. Conservación y Difusión

El objetivo de un proceso de digitalización de documentación de Archivo y creación de Registros digitales es doble:

1.- Generar un archivo digital original de **conservación** de modo que no sea necesario volver a manipular los documentos originales.

2.- Generar una copia de **difusion**, con buena calidad de lectura e impresión y posibilidad de su difusión en web.

Irargi.Centro de Patrimonio Documental de Euskadi, después de varios ensayos, ha decidido generar un **Archivo digital original** de estas caracteristicas:

a) en formato JPEG
b) con un tamaño entre 2 y 3 MB para una página estandar DIN A4[4].

Se trata de un archivo que:

1. permite una calidad óptima de lectura en pantalla
2. ofrece una buena capacidad de impresión del documento
3. sirve perfectamente para generar el archivo de conservación, con el consiguiente ahorro de tiempo, espacio, recursos y soportes.
4. ofrece un tamaño adecuado para su difusión en DVD.

Las copias de difusión se crearán a partir del master que se obtiene de este Registro original.

2.2.3. Soporte

Para la elección del soporte para conservar la información se tendrá en cuenta:

a. la garantía de preservación de la información
b. la facilidad de reproducción de la misma

A fecha de hoy (marzo de 2007) se aconseja el DVD, por las siguientes razones:

[4] Tamaño en el que asimilamos el tamaño de folio *prolongado* de la documentación de archivo clásica.

1. se trata de un soporte barato, fiable y estándar
2. puede ser leído en los sistemas operativos usuales: Apple Mac, Windows, MS Dos y Unix.
3. tiene una capacidad de almacenaje (estándar a fecha de hoy de 4,7 Gb) adecuada
4. no permite que la imagen sea manipulada o adulterada una vez grabado el DVD
5. la rápida y sencilla obtención de copias, acceso y lectura.

En cuanto a la perdurabilidad de la información grabada en DVD ésta permanecerá inalterable mientras el soporte no se destruya, ya sea por agresiones físicas o por degradación del mismo.

¿Qué garantía de conservación ofrecen estos soportes?. Aún cuando hay fabricantes de DVD que certifican la duración de los mismos durante 30, 50 e incluso 100 años, se entiende que se trata de un dato que no es posible certificar hasta que pase ese periodo de tiempo. Por ello es aconsejable programar un control periódico de calidad y conservación.

El archivo digital original es una copia de seguridad que no se vuelve a manipular. Se debe de conservar, por ello, en lugar seguro, de acuerdo con las indicaciones del fabricante.

Antes de guardar el archivo digital original se hará una copia que denominaremos **Master**. El master es la copia de uso de la que se generan las copias, de calidades y caracteristicas variables, para difusión directa y en web.

3. Difusión

Se entiende por difusión tanto la que se hace por medio de copias en DVD del master, como la que se efectua a través de internet. Estas Recomendaciones no abordan esta segunda opción.

Para las copias de uso, Irargi recomienda la aplicación **Acrobat**, de la casa Adobe, que utiliza el formato de archivo PDF, por las razones siguientes:

1. el formato PDF (Portable File Document) es el estándar más extendido, puede ser leido en cualquier plataforma informática, (PC, Mac Apple, MS Dos, Unix), es gratuito y se ha impuesto como el más utilizado por la comunidad de usuarios.
2. permite crear documentos que reproducen fielmente la secuencia de páginas del original
3. permite organizar los documentos contenidos en el DVD, incluir un índice de los mismos y crear “linkajes” para acceder directamente al documento seleccionado.
4. Permite visionar la documentación en distintos grados de ampliación, facilitando la lectura del documento.

5. permite imprimir los documentos, página a página o en su totalidad, con la calidad y tamaño que suministre la impresora del puesto donde se consulte.

Una vez que hemos concluido completamente el proceso de creación de un Archivo Digital, mediante la digitalización de los originales y su almacenaje, obtendremos como resultado la creación de las siguientes colecciones de DVD:

1º.- Una colección de DVD con los archivos originales tales cuales salen escaneados, y en formato JPEG comprimido. Se trata del almacenaje masivo de lo que denominaremos el **archivo digital original**, de conservación definitivo. Es la colección de copia de seguridad y no se vuelve a manipular.

2º.- A partir de estos DVD creamos un Master, que será la copia de uso para la obtención de copias sucesivas. De este Master se obtienen dos productos diferentes:

- la copia de difusión: el DVD, que a su vez hace el papel de copia original de la que se obtiene las demás copias para distribución entre instituciones, particulares, etc.
- la copia para distribución por internet: como ya hemos explicado, se trata, más que de una copia, de los archivos informáticos de menos volumen que se colocarán en el servidor web correspondiente.

4. Recomendaciones.

Creación del Registro/Archivo de Conservación

1. Todos los documentos históricos, tanto impresos como manuscritos, serán considerados como imágenes de tono continuo, y con el fin de reprodicur con la mayor fidelidad tanto su contenido como su estado físico, se digitalizarán a color real con una profundidad de color de 24 bits.
2. La resolución de escaneado será de 300 dpi, que puede aumentarse en caso de documentos con caracteristicas especiales.
3. El formato de archivo digital para conservación será JPEG comprimido, con un tamaño entre 2 y 3 MB (se entiende para documentación antigua asimilable al formato A4).
4. El Archivo digital original se grabará en soporte DVD o en un disco duro.
5. Una vez grabados los DVD se verificará que se abren correctamente y se recomienda testar su contenido antes de proceder a su almacenaje definitivo.
6. De estos DVD se generará una copia denominada Master, que servirá para generar todas las copias de difusión posteriores.
7. Se procurará no manipular ni tocar la superficie del DVD y no se escribirá ni se colocará etiqueta alguna sobre él. Se guardará en la caja

suministrada por el fabricante, que debe estar totalmente vacía. Se eliminarán las carátulas que pueda tener y solamente se le añadirá la etiqueta con el código de identificación del DVD.

Producción de una copia de difusión

8. El soporte utilizado será el DVD.
9. Se utilizará el formato de Archivo PDF para la creación y consulta de la información contenida en el DVD.
10. En cada DVD se incluirá un índice interactivo, con el inventario de la documentación contenida, que permitirá, por medio de linkajes, el acceso directo al documento seleccionado.
11. Se grabará el DVD en formato de grabación híbrido, es decir, legible tanto por sistemas operativos Mac Apple como Windows.
12. El primer DVD generado con estas especificaciones se denominará Master de difusión, y será el utilizado para generar las copias posteriores.