UNIVERSIDAD DE BELGRANO

Las tesinas de Belgrano

Facultad de Ingeniería y Tecnología Informática
Licenciatura en Sistemas de Información

Arbol de decisión para digitalizar documentos

Nº 200 **Alejandro Pérez Scianca**

Tutor: Carlos Gerardo Said

Departamento de Investigación
Julio 2007

Contenidos

1. Introducción 5
   1.1 Hipótesis 6
   1.2 Objetivos 6
   1.3 Metodología 6
2. Tecnologías de Digitalización 6
   2.1 Introducción 6
   2.2 La cadena de digitalización 6
   2.3 Primer Paso: Captura de imágenes 7
   2.4 Segundo paso: OCR 7
      2.4.1 Captura de imágenes y OCR 7
      2.4.2 Corrección de errores 7
2.5 La alternativa: Ingreso Manual 7
   2.6 OCR + Imágenes: Una combinación ganadora 8
      2.6.1 Wiki + OCR + Imágenes 8
3. Las imágenes digitales 8
   3.1 Archivos de Imágenes 8
      3.1.1 Imágenes Vectoriales 8
      3.1.2 Imágenes Raster 9
      3.1.3 Características de las imágenes 9
      3.1.4 Profundidad de color 9
      3.1.5 Resolución 12
4. Formatos de archivo 12
   4.1 Kodak PhotoCD 12
   4.2 TIFF 13
   4.3 JPEG 13
      4.3.1 JPEG2000 14
   4.4 JBIG / JBIG2 14
5. Formatos de Archivos de Texto 14
   5.1 El marcado en los textos 14
      5.1.1 Marcado procedural 15
      5.1.2 Marcado estructural 15
   5.2 Formatos de archivo de texto 16
      5.2.1 Texto plano 16
      5.2.2 Rich Text Format – RTF 16
      5.2.3 SGML – El padre de todos 16
      5.2.4 HTML – El hijo exitoso 17
      5.2.5 XML – La nueva tendencia 18
      5.2.6 DocBook 18
      5.2.7 Tex / LaTex 18
   5.3 Formatos mixtos 19
      5.3.1 MRC – Mixed Raster Content 19
   5.4 Comparación entre formatos 21
6. Metadatos 21
   6.1 Metadatos descriptivos 22
   6.2 Metadatos estructurales 22
   6.3 Metadatos administrativos 22
   6.4 Encabezado TEI - TEI Header 22
      6.4.1 La estructura de un texto TEI 22
   6.5 MARC 23
   6.6 La Iniciativa Dublin Core 23
7. Software para OCR 24
   7.1 Como funciona el OCR 24
      7.1.1 Comparación de patrones 24
      7.1.2 Extracción de características 25
      7.1.3 Comparación con diccionario 25
      7.1.4 OmniPage 26

7.1.5 SimpleOCR 26
7.1.6 Kooka – OCRad 26
7.1.7 Libre vs. Comercial 27
8. Herramientas de decisión 27
8.1 La matriz de decisión 27
8.2 El árbol de decisión 27
8.3 ¿Cual usar? Ambos! 29
9. Hardware de digitalización 29
9.1 Escáner 29
9.1.1 Funcionamiento de un escáner 29
9.1.2 Escáner Flatbed 29
9.1.3 Escáner de Alimentación automática 29
9.1.4 Escáner de tambor (Drum Scanner) 30
9.1.5 Fotocopiadoras 30
9.2 Cámara Digital 30
9.3 Tiempo vs. Dinero 31
9.4 Comparación entre dispositivos 31
10. Análisis del Documento 32
10.1 Definiendo el documento 32
10.1.1 Requerimientos según el tipo de documento 32
11. El factor económico 33
11.1 ¿Vale la pena la inversión? 33
12. Factores de decisión 33
12.1 Tipo de documento 33
12.1.1 Análisis visual y estructural 34
12.2 Volumen de la obra 34
12.3 Tipo de encuadernación 34
12.4 Mano de obra 34
12.5 Accesibilidad 34
12.6 Búsqueda de la información 34
12.7 Presupuesto disponible 34
12.8 Estándares de facto 34
13. Aprendiendo de otros proyectos 35
13.1 Project Gutenberg 35
13.2 Proyecto Crecer 35
13.3 Proyecto Biblioteca Digital Argentina – Clarín 35
13.4 BibliotecasVirtuales.com 36
13.5 Escuela Superior de Comercio Carlos Pellegrini 36
13.6 University of Virginia Library 37
13.7 Historietas - Comic Books 37
14. Tres Casos posibles 38
14.1 Estudio Jurídico 38
14.2 PYME que digitaliza documentacion 38
14.3 Biblioteca barrial 38
15. Conclusión 38
Anexo A: Glosario 39
Anexo B: Bibliografía 41
Anexo C – Carta para la preservación del patrimonio digital 42

1. Introducción

Desde el comienzo de la escritura, la acumulación, catalogación y conservación de escritos han sido de particular importancia para las culturas alrededor del mundo. Antes de Gütenberg hacer una copia de una obra escrita requería la paciente labor de uno o varios copistas que luego de meses o años de trabajo lograban reproducirla. Cada una de éstas era algo único y muy valioso. El invento de la imprenta, en 1540, logró que se abarataran muchísimo los costos aumentando así las posibilidades de tener múltiples copias de obras a nivel popular y acercando la brecha entre el autor y el lector, en definitiva popularizando (en términos relativos) la lectura y con ello la diseminación del conocimiento codificado y su consecuente penetración cultural en el conocimiento tácito. Esta realidad era la única hasta hace unos años. Sin embargo la popularidad de la informática introdujo un nuevo paradigma: la creación y distribución del libro electrónico. Esta manera de publicar requiere menos esfuerzo e inversión por parte del autor y acorta aún más la distancia entre este y su público separando la obra de su sustento físico. Este nuevo paradigma para las obras literarias o trabajos escritos se aplica a los autores contemporáneos y futuros pero no nos resuelve automáticamente el problema de los libros escritos con anterioridad. Millones de volúmenes que ya no serán reimpresos inexorablemente se perderán con el tiempo. Muchísimos libros que son parte de la historia serán deteriorados con el uso. Convertir las páginas de estos libros condenados a desaparecer a un formato digital parece una solución posible para estas obras que pasarían a tener la misma flexibilidad que las que nacieron entre bits y bytes.

¿Digitalizar una biblioteca? Esto nos permite conservar libros antiguos ya que al no ser necesario consultarlos directamente pueden ser guardados en ambientes adecuados y evitar el deterioro del uso cotidiano. También permite a muchas personas consultar simultáneamente una misma obra sin ser necesario conservar varias copias de la misma disminuyendo los costos y los requerimientos de espacio. Además permiten la búsqueda de texto y su catalogación de manera más eficiente y veloz. Estas son tan solo algunas de las ventajas que permite la digitalización de libros para formar una biblioteca electrónica.

La tecnología para la digitalización de libros esta disponible para bibliotecas de millares de volúmenes con presupuestos millonarios desde hace años y son muchas las que se han digitalizado de esta manera; sin embargo es poco lo que existe o está normado para la conversión de bibliotecas con «bajos» presupuestos como es el caso de las rurales o en países emergentes que quieran conservar su patrimonio u ofrecer nuevos accesos a sus recursos culturales sin requerir una gran inversión inicial.

El aporte de este trabajo es el de comparar tecnologías de digitalización disponibles y necesarias para un proyecto de estas características.

Para realizar este análisis comparativo, tomaremos en cuenta las siguientes características:

- Costo inicial de proyecto: Este punto es importante ya que no es lo mismo proponer la utilización de un escáner hogareño y software gratuito o de bajo costo que basarlo en dispositivos de entrada especialmente diseñados, cuyos costos son mucho mayores.
- Costo de mantenimiento: Es importante el costo de mantenimiento para mantenerlo operativo.
- Presupuesto inicial disponible
- Velocidad de entrada.
- Cantidad y calidad de mano de obra que la llevará a cabo.
- Capacitación necesaria de los recursos
- Posibilidad de obtener un sponsor: Existen proyectos financiados por empresas privadas y organizaciones no gubernamentales (ONG) que en determinados casos pueden llegar a aportar la financiación necesaria para iniciar alguno de estos proyectos.
- Formato: El formato obtenido luego del proceso es importante ya que es el que permitirá articularlo con otros proyectos similares en distintos lugares con economías emergentes potenciándose la obra como proyecto de carácter social. Por esto se compararán los distintos formatos de salida diferentes y también las ventajas y desventajas de unos y otros basándonos en popularidad, costos de licencia, tamaño de archivo resultante, tecnología, aplicaciones y visión a futuro. Ejemplos de ello son el texto plano (ASCII), Word, RTF, PDF (Portable Document Format), PostScript, MRC (Mixed Raster Content), Digipaper, TIFF Group 4, JPEG2000, HTML, XML y otros propuestos por la ISO.
- Cantidad de volúmenes
- Tipo de contenido predominante
- Valor intrínseco de las obras objeto del presente trabajo

Debido a que la terminología a utilizar puede ser demasiado específica se aclaran los acrónimos y definiciones principales en la presentación de cada tema (ej: TIFF, JPG, OCR). El resto de las definiciones se aclaran en el A**nexo A** donde se incluye un glosario.

1.1 Hipótesis

Mediante un árbol de decisión el responsable de una biblioteca puede evaluar la viabilidad de una digitalización y los métodos a utilizar ubicando en los distintos nodos del mismo el volumen de los documentos, los recursos económicos disponibles, recursos humanos, tiempo destinado al proceso, tipo de material a ser digitalizado y el valor intrínseco de las obras entre otros.

1.2 Objetivos

Comparar las tecnologías líderes de digitalización en precio, disponibilidad y performance.

Esquematizar el árbol de decisión en base a los datos obtenidos.

1.3 Metodología

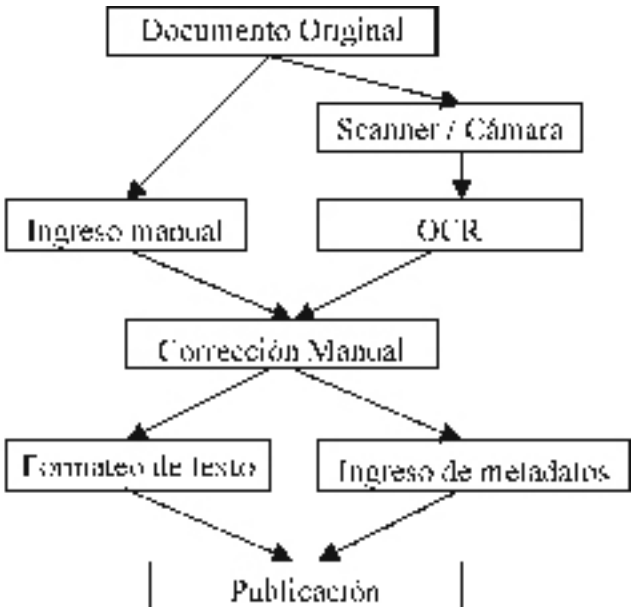
Se presentarán las opciones de hardware y software necesarios para la digitalización de una biblioteca y luego se hará una comparación entre ellos por medio de árboles de decisión.

2. Tecnologías de Digitalización

2.1 Introducción

Este trabajo se organizará en tres partes. En la primera se expondrán las tecnologías disponibles. La segunda tendrá como objetivo poner un marco al análisis del trabajo a digitalizar. Por último se expondrá un árbol de decisión que ayude en la elección de tecnologías a utilizar para cualquier proyecto de esta índole.

2.2 La cadena de digitalización



2.3 Primer Paso: Captura de imágenes

Lo primero que se debe hacer es capturar las páginas del libro que luego serán procesadas. Hay muchas variables (y por lo tanto decisiones) que hay que tomar a la hora de capturar una imagen y por lo tanto muchas preguntas a hacerse. ¿Qué tipo de texto estamos digitalizando? ¿Es importante el color? ¿Hay fotos o dibujos en el material? ¿Es necesario conservarlo como un facsímile (documentos históricos, firmas, escritura manual) o no?

En la digitalización siempre hay un compromiso entre la facilidad de distribución y uso y en la preservación de la apariencia original del documento. Por ejemplo un texto en formato plano es legible por prácticamente cualquier computadora en el mundo pero no mantiene la apariencia original. Por el contrario los formatos más sofisticados, que tienen información no sólo sobre el texto sino también sobre la distribución y apariencia, generalmente requieren programas especiales para ser leídos, achicando el porcentaje de equipos que pueden usarlos.

2.4 Segundo paso: OCR

Una vez digitalizadas las hojas, si se trata de textos, es necesario convertir esas imágenes de palabras en palabras reales.

El cerebro humano es capaz de reconocer una letra sin importar que esté en cientos de tamaños o estilos diferentes. El reconocimiento óptico de Caracteres (**OCR** en inglés, **O**ptical **C**haracter **R**ecognition) es un sistema que intenta emular esta habilidad en una computadora. Un software de OCR se encarga de traducir imágenes de un texto (normalmente capturadas por medio de un escáner o una cámara digital) y las convierte en algún formato de texto editable por medio de un procesador de texto.

El OCR se usa en los Estados Unidos desde el año 1965 para ordenar el correo postal y desde ese entonces su utilización ha aumentado en variadas aplicaciones. Existen muchos programas hechos para éste fin, cada uno con sus debilidades y fortalezas haciéndolos convenientes para distintos trabajos.

El reconocimiento con programas de OCR suele tener una efectividad superior al 99%[1] en casos ideales: captura nítida de la imagen, letras de imprenta bien marcadas. Sin embargo, aún con el 99% de aciertos, es necesaria la supervisión de un corrector. Si tenemos en cuenta que 1 de cada 100 letras puede estar equivocada entonces tenemos que suponer que en una página que tiene 80 caracteres de ancho vamos a encontrar prácticamente un error por línea. Lo que aparentemente es una tasa de error despreciable, un 1%, convierte a un libro de cientos de páginas en un arduo trabajo de corrección, con decenas de errores por página y miles al finalizar la corrección.

Con respecto a la letra manuscrita es mucho más difícil de reconocer que la de una imprenta. Usualmente es más fácil re tipear[2] el texto que intentar digitalizarlo por medio de OCR.

2.4.1 Captura de imágenes y OCR

Para poder reconocer adecuadamente las letras de una imagen es necesario que la digitalización se haya hecho en forma adecuada. Una imagen con poca resolución, manchas, poco contraste o algún defecto similar va a ser fuente de varios errores induciendo al software de reconocimiento a confundir letras o directamente a no poder reconocer nada. Tampoco es aconsejable usar una resolución excesiva: 300 DPI es una buena resolución para texto mayor a los 10 pts. En cambio para texto de 10 pts o menos, 400 DPI es más adecuado.

2.4.2 Corrección de errores

Una vez efectuado el OCR se debe hacer una corrección del texto. Dependiendo de la fiabilidad que se haya podido obtener del reconocimiento de los caracteres, la tarea puede ser sencilla o ardua. En determinados casos puede ser más rápido y eficaz re tipear todo el texto que perder tiempo intentando reparar algo irreparable.

2.5 La alternativa: Ingreso Manual

El re-tipeado es la única opción para los casos en que el OCR es poco aplicable o implica más problemas que soluciones. Si el texto a digitalizar es de pobre calidad por tener hojas manchadas o sucias, poco contraste o problemas similares entonces es posible que el tiempo transcurrido arreglando los errores que cometió la aplicación de OCR sea mayor al que se emplearía en pasarlo a mano desde cero.

---

[1] http://www.scansoft.com/omnipage/features.asp

[2] Aunque tipiar es la forma que incluye el Diccionario de la lengua, en la Argentina es corriente y aceptable la grafía tipear, que la Academia Española registra como neológica y con nota de americanismo en la última edición de su Diccionario manual.

2.6 OCR + Imágenes: Una combinación ganadora

Cuando se tiene que procesar un texto con dificultades para reconocer o cuando no se puede hacer el trabajo de corrección posterior existe la alternativa de usar la imagen de las páginas para la lectura e impresión y el texto de OCR sin corregir para la búsqueda de datos.

En pantalla se tendría una foto de la página original, que se usaría para leerla en la computadora y para generar copias impresas. El texto producido por el programa de OCR, a pesar de los errores, sería muy útil para la búsqueda de palabras dentro del contenido.

2.6.1 Wiki + OCR + Imágenes

Una manera de hacer correcciones sobre un texto no corregido es a través de un *Wiki*[3]. De esta manera cuando un usuario necesita la información en forma de texto (para citarla, copiarla, editarla) tiene la posibilidad de cotejarla con el texto original y corregir los errores. El sistema de *Wiki* permite que cualquier lector pueda agregar correcciones y al mismo tiempo mantener las versiones anteriores como resguardo.

3. Las imágenes digitales

En esta sección se analizará el estado actual de las tecnologías implicadas en el proceso de digitalización que complementan el trabajo realizado por el dispositivo físico o hardware.

Estas tecnologías son:

- Las relacionadas con las imágenes y la compresión utilizada
- Las relacionaras con el formato de texto y su metainformación
- La de reconocimiento de caracteres u OCR

Los documentos digitalizados son almacenados en archivos de computadora. Hay varios tipos de archivos que catalogaremos en tres grupos principales: **archivos de imágenes, archivos de texto y mixtos**. La elección del tipo adecuado para el trabajo va a depender de muchos factores entre los cuales se incluye el resultado del análisis del documento sobre el que se va a trabajar, el uso que se le va a dar a lo digitalizado, el espacio destinado al almacenamiento y las herramientas de las que disponga el usuario para leerlo.

3.1 Archivos de Imágenes

Los archivos de imagen en dos dimensiones se pueden clasificar en dos grupos: raster o mapa de bits y vectorial. Esta clasificación no es rígida ya que las imágenes vectoriales pueden tener imágenes raster en el interior y viceversa.

Sin embargo, a los efectos de la digitalización de documentos solo se van a usar las raster, debido a que son estas las producidas por escáner y cámaras digitales.

3.1.1 Imágenes Vectoriales

Aunque estas imágenes no se tienen en cuenta en proyecto, se dará una explicación básica.

Una forma de generar una imagen es la de hacerlo mediante operaciones matemáticas. A diferencia de los raster, para trazar una línea se determinan unas coordenadas x1 e y1 y se traza una línea hasta otras coordenadas x2 e y2.

Así se pueden dibujar círculos, cuadrados, triángulos y miles de formas. Esta es la base de los llamados dibujos vectoriales.

Los trazados (líneas curvas o rectas propias de un dibujo vectorial) se pueden modificar fácilmente, ocupan muy poco espacio y además son independientes de la resolución, ya que no dependen de una retícula dada y basándose en que cualquier operación geométrica es multiplicable o divisible en su conjunto sin que eso afecte al aspecto del resultado, sino sólo a su tamaño final.

Las imágenes vectoriales de dos dimensiones suelen componerse de varias partes. Sólo el contorno y el relleno serán visibles al imprimir. Lo demás son instrumentos de trabajo. La base de estas operaciones son las llamadas «Curvas Bezier»:

El principal inconveniente de este tipo de imágenes es que son ideales para crear un dibujo esquemático, pero resulta difícil usarlo para fotos o dibujos complejos. Por eso la utilidad en la digitalización de un documento es muy reducida y no será tema de discusión en esta tesina.

La imagen 3-1 muestra una curva Bezier.

[3] Ver Glosario

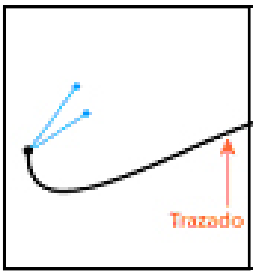


Ilustración 3.1

3.1.1 Imágenes Raster

Los archivos de salida de un escáner o de una cámara digital son raster y son un facsímile del documento existente. Esta característica los convierte en la alternativa mas propicia cuando se quiere preservar documentos en los que es tan importante el formato del documento como el contenido (por ejemplo en proyectos de preservación de documentos antiguos). Sin embargo al hacerse de esta manera se pierden muchos de los beneficios de un documento digital: la posibilidad de la búsqueda de texto dentro del documento, el rearmado automático para otros tamaños de página, la adaptación a la pantalla o a un dispositivo de lectura, la posibilidad de traducirlos automáticamente con el software adecuado, la lectura fonética para ciegos entre otros.

Los archivos de imagen, por lo tanto, son ideales para documentos fotográficos o pictóricos, manuscritos que se quieren preservar por su valor histórico, ejemplares firmados, formularios etc.

Éste tipo de archivos puede ser el final de la cadena de digitalización o tan solo un paso previo al del reconocimiento del texto por medio de OCR.

3.1.2 Características de las imágenes

Básicamente los tres aspectos que tenemos que considerar son: **la profundidad de color**, la **resolución** y el **formato del archivo** en el que se va a guardar. Estos tres elementos se explicarán a continuación.

3.1.3 Profundidad de color

Es uno de los tres aspectos que definen las características de una imagen.

La profundidad de color tiene que ver con la cantidad distinta de colores que se pueden mostrar por cada píxel.

Podemos clasificar a las imágenes en cuatro tipos básicos según este parámetro: **Imágenes de 1-bit, imágenes de 8-bit en escala de grises, imágenes de 8-bit en color e imágenes de 24-bit en color**. Existen también imágenes de 32-bit y de 48-bit por píxel, pero no son la norma y actualmente solo se usan para proyectos muy específicos.

3.1.3.1 1-bit

En las imágenes de 1 bit cada píxel puede ser blanco o negro. La pobre calidad asociada a esta poca versatilidad hace que este formato no sea muy utilizado, con la excepción de imágenes en las que se muestra texto o gráficos de líneas definidas. A pesar de podríamos pensar que para digitalizar un texto desde un libro con hojas blancas y texto negro sería viable utilizar este formato y ahorrar muchos bytes, esto no es así. Las hojas de los libros suelen estar manchadas, obscurecidas, rayadas etc. En una imagen en 1-bit las manchas no se pueden diferenciar de las letras, confundiendo al programa de OCR. En color, en cambio, este problema no se presenta en la misma medida. Esto nos indica que es un formato poco conveniente a pesar de lo atractivo de su tamaño en KB.

El abaratamiento del almacenamiento y las técnicas de compresión están volviendo obsoletas a las imágenes de 1-bit, al menos para la digitalización de documentos. Sin embargo sigue siendo muy útil para la producción de documentos directamente en formato digital.

Las imágenes de 1 bit pueden ser de dos modalidades distintas: ***Line art*** o ***Halftone*** (medios tonos).

3.1.4.1.1 Line art

En modo *line art* las figuras son bien definidas en blanco o en negro, transformando las áreas obscuras en manchones negros. Esto lo hace completamente inadecuado para fotografías o imágenes en tonos de grises y solo es útil para dibujos bien definidos o texto.

3.1.4.1.2 Halftone

Las imágenes en medios tonos o *halftone* en cambio intentan representar las gamas de grises con patrones de negros y blancos que «engañan al ojo» haciéndolo pensar que es un gris continuo. Este es el tipo de gris que suele encontrarse en los diarios para imprimir las fotos.

En la ilustración 3-2 tenemos el ejemplo de tres cuadros, uno totalmente en blanco y los otros dos en tonos de grises, formados por patrones de blancos y negros en un cuadrado de 8x8, es decir, 64 píxeles.

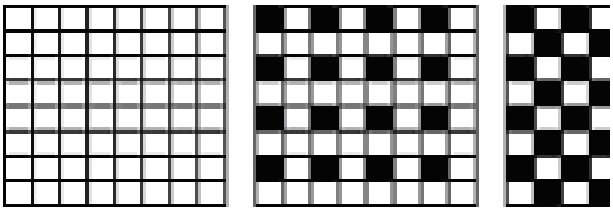


Ilustración 3 2

El cuadro de la izquierda es de referencia y es todo blanco. El del medio tiene solo 1 de cada 4 puntos de color negro formando un gris del 25%. Esto significa que visto de lejos el cuadro del centro aparenta ser de un gris uniforme claro. El de la derecha tiene la mitad de los cuadros en negro. Esto hace que represente un gris medio o gris al 50%.

3.1.4.2 8-bit en escala de grises

Una imagen en escala de grises es el equivalente a una fotografía en blanco y negro.

Las imágenes con una profundidad de 8-bit en escala de grises constituyen una gran mejora con respecto a las de 1-bit ya que utilizan 256 $(2)^8$ tonos de gris resultando de una exactitud mayor. Se le asigna un número del 0 al 255 a cada píxel de la imagen. El número 0 representa el negro y el 255 el blanco. Los números entre el 1 y el 254 representan los distintos tonos de gris.

Esta profundidad puede usarse para prácticamente todos los trabajos en blanco y negro con la excepción de los que deben ser preservados con la calidad necesaria para usarse como reemplazo del original en cuyo caso suele usarse 24-bit.

3.1.4.3 8-bit en color

El formato es similar al de 8-bit en escala de grises pero en cada píxel en lugar de 256 tonos de gris hay 256 colores. Esta menor variedad cromática hace que las imágenes sean de menor calidad en comparación con las de gris y definitivamente de mucho menor que las de 24-bit.

Para llegar a los 8 bits se parte de una imagen de 24-bit y luego, por medio de un proceso llamado *cuantificación,* se elige una paleta de 256 colores, los cuales representan mejor la imagen, el resto son descartados.

Las imágenes de 8-bit se utilizan principalmente la WEB o en casos en que la economía (entendiendo economía por ahorro de espacio) es factor decisivo y se necesitan imágenes a color.

3.1.4.4 24-bit en color

Las imágenes de 24-bit utilizan 16.8 millones de colores $(2)^{24}$ por cada píxel para representar una imagen. Estos 24 bits están repartidos en

Ilustración 3 3

3.1.4.4 24-bit en color

Las imágenes de 24-bit utilizan 16.8 millones de colores (2)24 por cada píxel para representar una imagen. Estos 24 bits están repartidos en tres canales, con 8 bits para cada uno: Rojo, Verde y Azul o Red-Green-Blue (RGB). Cada píxel está compuesto por tres puntos, uno de cada color base. Al igual en que la escala de gris, cada canal usa un número para la intensidad de su color, siendo el 0 para el negro y el 255 para el color puro. Un píxel formado por (0,0,255) será entonces perfectamente azul. Variando la intensidad de estos colores es que pueden formarse los millones antes mencionados.

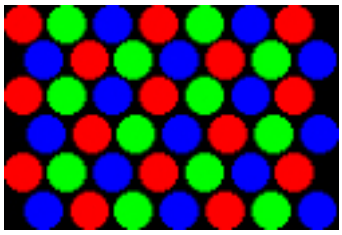


Ilustración 3.4

Cada combinación aparece como un color distinto y cuando los canales tienen el mismo valor entre sí, el color representado es el gris. El píxel (128,128,128) será entonces un gris medio.

24-bit suele ser la norma de referencia para cualquier trabajo gráfico de calidad y la única desventaja que puede encontrarse es su tamaño de almacenamiento. Las imágenes en 24-bit son las más fieles al original aún cuando la fuente es en blanco y negro.

3.1.4.5 Comparación de archivos

A continuación se presenta una misma imagen (Ilustración 3-5) en line art, halftone, escala de grises de 8bit y 24-bit en color.

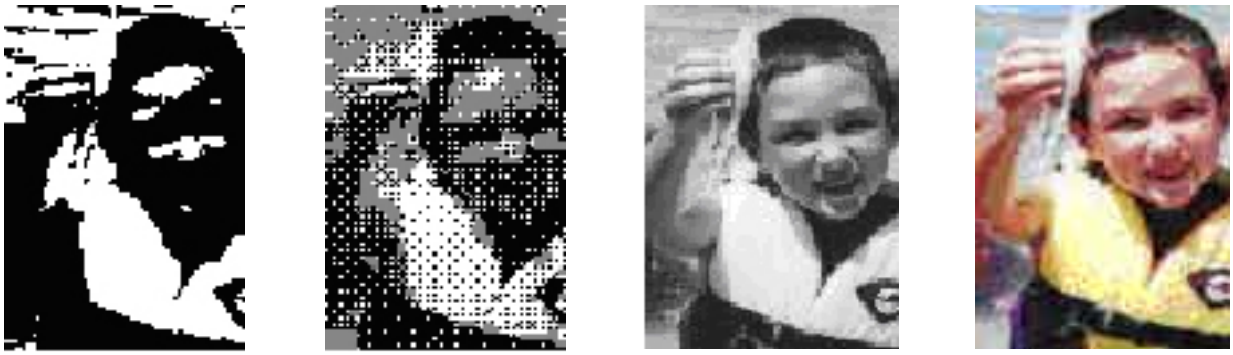


Ilustración 3.5

En la ilustración 3-6 se puede ver la ventaja de digitalizar con más colores a pesar de que el texto original es en blanco y negro. Al usar los colores se puede separar la mancha del texto, permitiendo al programa de OCR hacer un mejor trabajo.

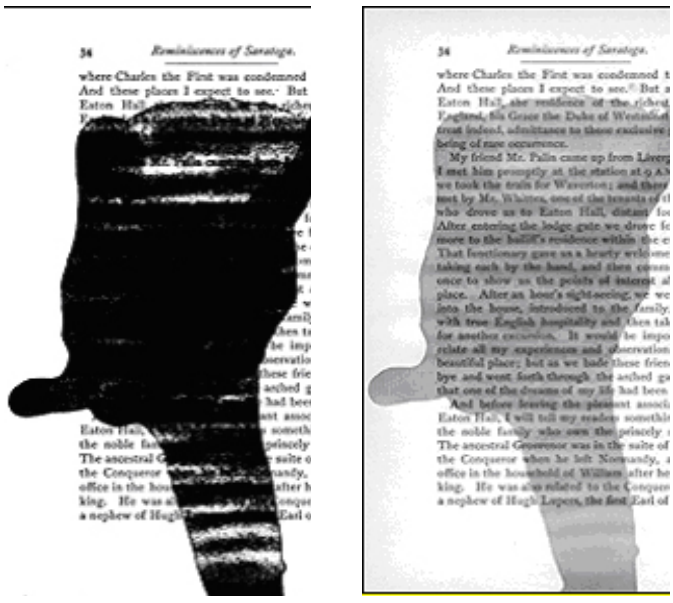


Ilustración 3.6

3.1.5 Resolución

La resolución está dada por la cantidad de puntos por unidad de medida. La medida usada generalmente es la pulgada dando origen a la sigla DPI, del inglés dots per inch, o puntos por pulgada. Cuanto mayor resolución se utilice mayor cantidad de información será registrada en el archivo haciéndolo mas preciso y, consecuentemente, más grande. La cantidad de DPI a utilizar estará dada, por lo tanto, por el compromiso entre la necesidad de resolución y la capacidad de almacenamiento de nuestro equipamiento. Sin embargo, en algún punto, una mayor resolución no tendrá como resultado una ganancia evidente en la calidad de la imagen, sino un mayor tamaño de archivo. El punto clave es determinar la resolución necesaria para capturar todos los detalles importantes que están presentes en el documento original.

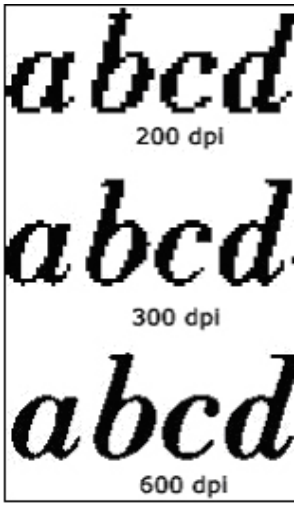


Ilustración 3.7 - Un mismo texto en distintas resoluciones

Por ejemplo, si la imagen debe ser considerada para un archivo histórico o debe agrandarse para su estudio entonces la resolución debe ser elevada Si en cambio va a utilizarse para un graficar el layout original de un libro ahora digitalizado o para un acceso Web entonces la resolución debe ser necesariamente menor.

La tabla 3-1 muestra ejemplos de resoluciones y sus respectivos tamaños en bytes para un cuadrado de imagen de 1x1 en pulgadas.

Resolución (DPI)	100 x 100	200 x 200	300 x 300	400 x 400
8-bit	9 KB	39 KB	89 KB	158 KB
24-bit	29 KB	118 KB	267 KB	475 KB

Tabla 3.1

4. FORMATOS DE ARCHIVO

4.1 KODAK PHOTOCD

Kodak es uno de los pioneros en la digitalización de fotografías a nivel masivo. Desde 1992, la tecnología Kodak PhotoCD se estuvo usando (principalmente en los Estados Unidos) para digitalizar fotos desde negativos y almacenarlas en Compact Disc. Este proceso se hace directamente en los Mini-Lab simultáneamente con el revelado. Éstos CD se pueden reproducir en cualquier PC o en un reproductor de DVD.

Se habla de tecnología PhotoCD porque ésta combina una técnica de escaneo, un medio de soporte, un formato de imagen y compresión y el sistema de administración de color para mantener la fidelidad cromática.

A pesar de ser una tecnología no tan expandida ni usar los formatos de imagen más populares se la incluye en este trabajo porque es la que se utiliza en el proyecto de la Universidad Cornell, un referente en la digitalización de documentos, con una de las biliotecas más grandes del mundo.

4.2 TIFF

El formato TIFF (del inglés Tagged Image File Format) es el formato más ampliamente aceptado para la conservación de imágenes. Fue desarrollado por Aldus y Microsoft. La especificación pertenecía originalmente a Aldus y luego fue comprada por Adobe quien actualmente tiene los derechos sobre la especificación de TIFF.

Los archivos TIFF tienen la ventaja de ser leídos en prácticamente cualquier plataforma. Esta característica los convierte en la alternativa más usada para trabajos de archivo.

La mayoría de los procesos de digitalización comienzan convirtiendo el documento original a TIFF ya que este el es formato que permite capturar más información y pasa a ser considerado el original.

Los archivos TIFF están diseñados para almacenar imágenes raster. No soportan gráficos vectoriales, anotaciones de texto etc.

La gran virtud del formato es su flexibilidad. Soportan un número arbitrario de bits por píxel, múltiples imágenes por archivo, imágenes de pre-visualización y distintos formatos de compresión.

Entre los espacios de color soportados se encuentran:

· Escala de Gris
· Pseudocolor
· RGB
· CMYK

Algunos de los formatos de compresión son los siguientes:

· Sin comprimir - raw
· PackBits
· Lempel-Ziv-Welch (LZW)
· CCITT Fax 3 & 4
· JPEG

4.3 JPEG

JPEG es un método de compresión de imágenes. Es el acrónimo de Joint Photographic Experts Group, nombre del comité que desarrolló el estándar.

JPEG está diseñado para comprimir archivos de color real (24-bit) o de escala de grises. Su mejor aplicación es en fotografías y no es buena en letras o dibujos de línea.

JPEG usa un formato de compresión lossy o con pérdida de información. Esto significa que el archivo una vez descomprimido nunca vuelve a ser exactamente igual al original. De esta manera logra radios de compresión mucho más grandes que con cualquier compresión lossless o sin pérdida de información. La manera de lograr esta técnica es explotando las limitaciones del ojo humano, principalmente el hecho de que las pequeñas variaciones en el color son menos perceptibles que las variaciones en la luminosidad. Entonces las imágenes comprimidas con este método tienen que ser destinadas a ser vistas por los humanos y no a ser reprocesadas por un programa.

La compresión puede variar según las necesidades, pudiendo ser más agresivo en caso de necesitar mayor ahorro en tamaño o menos agresivo si se necesita mantener la calidad. Una compresión 10 a 1 es prácticamente indistinguible del original, mientras que una 100 a 1 se verá borrosa y con bloques en la imagen que son típicos del método en cuestión. El nivel adecuado de compresión depende del uso destinado a la imagen.

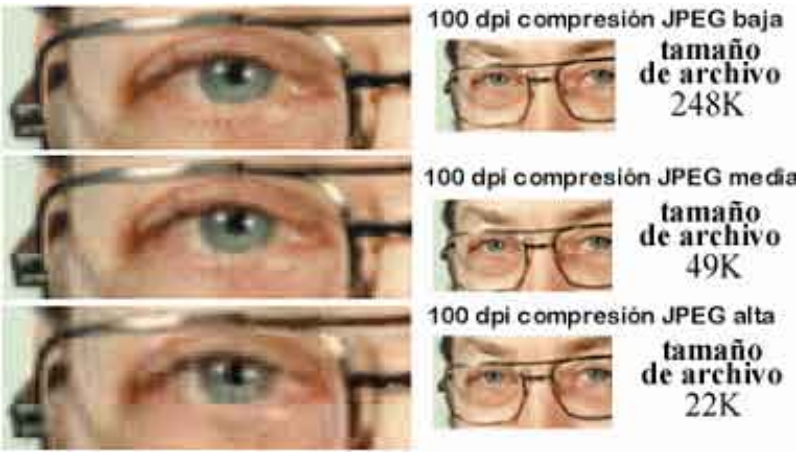


Ilustración 4.1 © Universidad de Cornell / Departamento de Investigación

En la imagen 4-1 podemos ver el deterioro de la calidad de una imagen a medida que se comprime.

4.3.1 JPEG2000

JPEG 2000 fue creado por el Joint Photographic Experts Group e ideado como el sucesor de JPEG. Se convirtió en un estándar ISO en el 2000 (ISO/IEC 15444-1:2000) pero no logró reemplazar al JPEG aun. La lentitud en la adopción del JPEG 2000 es consecuencia de su novedad en el mercado y de la falta de soporte en el Internet Explorer, que imposibilita su uso en la WEB a la mayoría de los usuarios de PC.

JPEG 2000 puede manejar factores de compresión más altos que JPEG sin las manchas y cuadros típicos de los JPEG originales.

La compresión alcanzada con este método es aproximadamente un 20% más eficiente que la del formato JPEG en una compresión media. En compresiones más agresivas el ahorro es aún mayor.

4.4 JBIG / JBIG2

JBIG es acrónimo de Joint Bi-level Image experts Group, un comité designado por los organismos de estándares y las grandes compañías del rubro para producir un estándar de codificación para imágenes de 1-bit.

5. FORMATOS DE ARCHIVOS DE TEXTO

Los archivos de texto pueden contener texto y también incluir imágenes. Estos archivos generalmente usan un lenguaje de marcado para dar formato al texto.

A continuación se explicará el significado de un lenguaje de marcado, los diferentes tipos de marcado y su importancia en un texto. Luego se desarrollarán los distintos formatos de archivo de texto.

5.1 EL MARCADO EN LOS TEXTOS

Históricamente el marcado fue utilizado para describir anotaciones que indicaban al compositor o tipista de que manera tenían que imprimirse o acomodarse los textos en la imprenta. A medida que la impresión de los textos se fue automatizando el termino marcación se extendió a todo tipo de códigos especiales que se le agregan a un texto para darle formato, la impresión o cualquier otro proceso.

Generalizando se puede definir como marcado a cualquier método para hacer explícita una interpretación de un texto. En el sentido más general se puede decir que los signos de puntuación, las mayúsculas y las convenciones de escritura son una forma básica de marcado que indican al lector la manera en que debe ser leído un texto. En informática el marcado se extiende al formato del texto, tipo de letra, información del texto etc.

El marcado necesita un lenguaje de marcado. Éste lenguaje es una convención utilizada que debe especificar cuáles son las marcas permitidas, cuáles son requeridas, cómo van a ser distinguidas del texto al que modifican y qué significan.

El lenguaje de marcado base es el SGML y de este derivan el HTML y el XML entre otros.

En la imagen 5-1 vemos una página con las anotaciones para el encargado de imprimirla. Los comentarios son acerca del tipo y tamaño de letra.

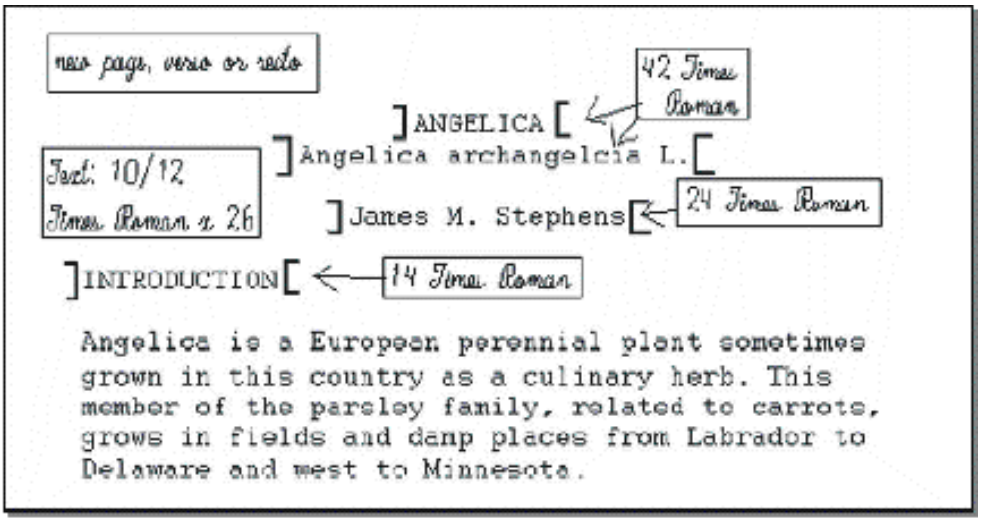


Ilustración 5.1 – Ejemplo de una página con marcado manual.

5.1.1 Marcado procedural

Como su nombre lo indica, el marcado procedural conlleva un procedimiento para el dispositivo de salida.

El usuario analiza de manera implícita la estructura de su documento e identifica por separado cada elemento significativo.

El usuario especifica la apariencia del texto (tipo de fuente, tamaño, posición), utilizando los comandos apropiados para producir el formato deseado para cada elemento.

Este tipo de marcado es muy similar al tradicional, la única diferencia es que el marcado se almacena electrónicamente.

El marcado procedural incluye opciones de marcado tipográfico en medio del documento perdiendo su estructura. El contenido verdaderamente importante del documento y sus elementos significativos se pierden entre el tipo de codificación utilizada.

Por sus características presenta las siguientes varias desventajas:

- No conserva la estructura del documento. Graba el resultado del proceso y pierde la estructura.
- No es flexible. Cualquier cambio a las normas de formato implica cambiar de manera manual el documento.
- Es lento y propenso a errores.

Ejemplo de un texto codificado con el lenguaje de marcado procedural RTF (Formato de Texto Enriquecido).

```
{\rtf
 Hola
 \par
 Texto de
 {\b prueba }.
 \par
 Fin
}
```

5.1.2 Marcado estructural

Quienes empezaban a participar en la creación de nuevas formatos para la escritura de documentación teniendo en cuenta las notables desventajas del marcado procedural, empezaron a emplear la codificación genérica o estructurada. Ésta se originó con la introducción de los macros o etiquetas. Cada etiqueta se agrega a cada elemento de texto y es a las etiquetas a las que se asocian normas de formato. Un formateador procesa el texto y produce un documento de salida.

El marcado estructural por sus características tiene las siguientes ventajas:

- El marcado describe únicamente la estructura lógica del documento.
- Es más flexible. Para cambiar la apariencia del texto basta con adaptar la etiqueta correspondiente.
- El autor se centra en la labor de escritura y creación del documento y no en su formato.
- La estructura del documento facilita el manejo, almacenamiento, consulta y proceso de la información que éste proporciona.

Ejemplo de un texto codificado con el lenguaje de marcado estructural LaTex[4]*.*

```
\documentclass {article}
\usepackage [spanish] {babel}
\begin{document}
\title{Un documento de Prueba}
\author{Alejandro Perez Scianca}
\maketitle
\section{Introducci\'on}
!`Hola mundo TeX !,
\section{Una f\'ormula}
Un ejemplo de la capacidad de tex, una f\'ormula:
$$ \alpha = \int_0^1 \frac{1}{1+x^2} dx $$
\end{document}
```

4. Ver 5.2.7

5.2 FORMATOS DE ARCHIVO DE TEXTO

A continuación se describirán los formatos de texto considerados para este trabajo: texto plano, Rich Text Format, SGML, HTML, HML, DocBook y Tex.

5.2.1 Texto plano

Este formato fue inicialmente desarrollado para sistemas usando terminales “bobas” con pantallas que muestran 80 caracteres por línea y 24 líneas por vez limitados al juego de caracteres estándar ASCII.

Un archivo de texto plano (plain text en inglés) es un archivo con caracteres ASCII generalmente compuesto de líneas de texto de 80 caracteres con un CR o CR/LF[5].

El formato de los párrafos es llevado a cabo a través de líneas en blanco para los párrafos o espacios en blanco para las sangrías.

Las opciones para el formato del texto son aún más limitadas ya que no hay posibilidad de elegir tamaños ni tipo de letra y suelen usarse caracteres especiales para el lenguaje de marcado aunque no hay una convención. Un ejemplo de esto son las marcas para *negritas*, /itálicas/ o _subrayados_. También suele enfatizarse el texto usando MAYUSCULAS.

Otro problema es que no soporta juegos de caracteres distintos al romano obligando al escritor a hacer trucos para poder mostrar letras no romanas.

5.2.2 Rich Text Format – RTF

La especificación de RTF (acrónimo del inglés Rich Text Format o Formato de Texto Enriquecido) define la estructura de un archivo texto con formato y gráficos que pretende ser común entre los procesadores de texto.

Fue creado por Microsoft para facilitar el traspaso de información entre programas y plataformas y es usado en la actualidad con ese fin. A pesar de ser diseñado para mantener la estructura y el formato del texto, no contiene la suficiente información para asegurar WYSIWYG[6] entre plataformas. Este formato está soportado por la mayoría de los procesadores de texto y en los principales sistemas operativos de PC de escritorio (Windows, Linux, MacOS X). Lamentablemente el estándar no es seguido adecuadamente ni siquiera por los productos de Microsoft y por lo tanto no se convirtió en el de intercambio como estaba planificado.

RTF usa palabras de control o control words como comandos que usan las aplicaciones para aplicar el formato adecuado y manejar los documentos. Estas palabras de control no pueden ser mayores a 32 caracteres y empiezan con un backslash (\).

Ejemplo de un texto RTF

Código fuente

```
{\rtf
 Hola
 \par
 Texto de
 {\b prueba }.
 \par
 Fin
}
```

Se mostrará

```
Hola
Texto de prueba.
Fin
```

5.2.3 SGML – El padre de todos

SGML son las siglas de Standard Generalized Markup Language o Lenguaje de Marcación Generalizado. La Organización Internacional de Estándares (ISO) normalizó este lenguaje en 1986. SGML es un estándar internacional de marcado para descripción de texto electrónico. Mas específicamente SGML es un metalenguaje, en este caso el lenguaje que describe es el de marcado.

5. Es uno de los caracteres de control en el código ASCII que le indica a la impresora que el cursor debe posicionarse en la primera posición de la línea. Suele usarse con la sentencia LF (line feed) para moverse a la línea inferior
6. Ver Glosario

El lenguaje SGML sirve para especificar las reglas de etiquetado de documentos y no impone en si ningún conjunto de etiquetas en especial. SGML por lo tanto no está restringido a ningún tipo de aplicación, usándose en documentaciones técnicas de la industria Aeroespacial o en lingüística.

5.2.3.2 Marcado descriptivo

El SGML usa códigos de marcado que provee nombres para categorizar partes de un documento. Por ejemplo: <para> o \end {list} identifican que comienza un párrafo o que es el final de la última lista empezada.

5.2.3.3 Tipos de documento

SGML introduce el concepto de tipo de documento y de definición de tipo de documento (en adelante DTD, Document Type Definition). El tipo de documento es formalmente definido por su estructura. La definición de un informe, por ejemplo, puede consistir de un título, un autor, un resumen y una secuencia de uno o más párrafos. Por lo tanto, un escrito que no tenga un título o alguno de los elementos requeridos no es formalmente un informe según esta definición, sin importar si el contenido es similar al de un informe.

Los programas de manejo de documentos hacen uso de esta estructura para trabajar más inteligentemente.

5.2.3.4 Independencia de datos

Un objetivo primordial en el diseño de SGML fue el de asegurarse que los documentos descriptos correctamente puedan ser intercambiados entre distintas plataformas de software y hardware sin ningún tipo de perdida de información.

5.2.3.5 Historia del SGML

En 1969 un equipo de IBM que estaba trabajando en un proyecto para integrar información de estudios de abogados desarrollo el GML[7] o Generalized Markup Lenguaje (lenguaje de marcado generalizado) como un método de edición, formateo y recuperación de información de los documentos compartidos. GML pasó a ser el lenguaje de marcado de IBM.

GML, en vez de ser sólo un sistema de marcado, introdujo el concepto de un documento formalmente definido con una estructura anidada.

SGML agregó al GML los conceptos de referencias cortas y vínculos, lo que lo hizo mas versátil.

El SGML fue desarrollado por un comité creado por la ANSI en 1978 y fue publicado como estándar en 1980. En 1985 se convirtió en un estándar internacional siendo aprobado por la ISO en 1986.

5.2.4 HTML – El hijo exitoso

HTML son las siglas de HyperText Markup Language, Lenguaje para el Formato de Documentos de Hipertexto.

El HTML fue desarrollado originalmente por Tim Berners-Lee mientras estaba en el CERN (Centre Européen de Recherche Nucléaire, Centro Europeo de Investigación Nuclear) en Ginebra, Suiza, en el año 1989 como un derivado del SGML. El objetivo era el de organizar los documentos internos del CERN valiéndose del concepto de hipertexto.

El HTML fue popularizado por el navegador Mosaic desarrollado en el NCSA (The National Center for Supercomputing Applications). Durante los años 90 se extendió su uso con el crecimiento explosivo de la Web. Durante este tiempo, el HTML se ha desarrollado de diferentes maneras. La Web depende de que los autores de páginas Web y los creadores de los programas para navegarlas compartan las mismas convenciones de HTML. Esto ha motivado el trabajo colectivo en las especificaciones del HTML que son mantenidas por el W3 Consortium, fundado en 1994 por el mismo Tim Berners-Lee.

El HTML es una implementación simple del SGML y permite crear documentos multimedia con imágenes que pueden ser publicados para un acceso fácil.

A diferencia de otros formatos no hace hincapié en mantener el la forma ya que un documento puede verse considerablemente distinto según la plataforma sobre la que se esté observando. Como contrapartida es un formato ideal para distribuir contenido debido a la universalidad que adquirió. Prácticamente no hay PC que no esté en condiciones de poder ver un documento HTML.

7. Los creadores del GML son Charles Goldfarb, Edgard Mosher y Raymond Lorie, cuyas iniciales también son GML.

5.2.5 XML – La nueva tendencia

SGML es extremadamente poderoso pero es difícil de implementar. Esta característica lo alejó de volverse un estándar popular. El HTML con su limitada cantidad de etiquetas y su simplicidad en cambio se volvió extremadamente popular. Es fácil de aprender y crear un documento HTML es posible con prácticamente cualquier herramienta de producción de texto. Sin embargo es limitado su uso.

Los creadores de contenido entonces se enfrentaban al siguiente dilema: Usar SGML con los requerimientos necesarios en cuanto a capacitación que necesita o usar HTML y encontrarse en seguida con sus límites. Para cubrir la brecha en 1998 se creo el XML.

XML es acrónimo de eXtensible Markup Lenguaje o Lenguaje de Marcado Extensible.

Las características de XML son:

- XML debe ser usable sobre Internet sin complicaciones
- Debe soportar una amplia gama de aplicaciones
- Debe ser compatible con SGML
- Debe ser fácil hacer un programa para manejar XML
- Los documentos XML deben ser legibles por un humano,

La idea entonces era usar la mayoría de las ventajas del SGML con la simplicidad del HTML. Se dice que cumple con el 80% de las especificaciones del SGML con sólo el 20% de las complicaciones.

5.2.6 DocBook

DocBook es una implementación de SGML que permite la escritura de documentación técnica y que se volvió muy popular dentro de la comunidad del software libre y Open Source, particularmente en los ambientes Linux y BSD. Prueba de ello es que la documentación de los proyectos principales de Linux: Gnome, KDE y LinuxDoc esté realizada con este formato.

```
<!DOCTYPE book PUBLIC “-//OASIS//DTD DocBook V4.1//EN”>
<book lang=”es”>
<chapter>
<title><acronym>DocBook</acronym></title>
<para></para>
<sect1>
<title>Historia</title>
<para></para>
</sect1>
<sect1>
<title>Marquillas</title>
<para></para>
</sect1>
</chapter>
<chapter>
<title><application>emacs</application></title>
<para></para>
<sect1>
<title>Invocación</title>
<para></para>
</sect1>
<sect1>
<title>Escribiendo y guardando un archivo</title>
</sect1>
</chapter>
</book>
```

5.2.7 Tex / LaTex

Desilusionado con las pruebas de imprenta de uno de sus libros, el profesor Donald Knuth, de la universidad de Standford, decidió crear, en 1977, un sistema de composición de textos científicos. El nombre que le puso fue TEX.

A principios de los ochenta se desarrolló un software que se basó en TEX pero con la idea de que los autores pudieran concentrarse más en la estructura del docu-

mento que en su formato. Este sistema se llamó LaTeX. Sobre este sistema se armó un kit de programas auxiliares para creación de índices, referencias, inserción de gráficos y tablas de contenido. Todas estas características no estaban incluidas en el TEX original.

Por su parte LaTex hace que el autor se concentre en la estructura lógica del texto encargándose el programa de la composición y dejando gran parte de las decisiones técnicas a los profesionales del diseño tipográfico. De esta forma cuando, por ejemplo, comenzamos un capítulo, debemos indicárselo a LaTex así como su título, pero nos olvidaremos de tener que tomar decisiones sobre la manera de escribir la cabecera del capítulo: el tipo y tamaño de la letra del título, los espacios, la justificación, etc. Todas esas especificaciones se dan en un fichero de estilo que basta modificarlo para cambiar automáticamente las cabeceras de todos los capítulos.

LaTex es un auténtico lenguaje: los archivos de entrada deben ser compilados para que se produzca como salida el documento final formateado, disponible para ser mostrado en pantalla o impreso. Esto presenta dos ventajas: por un parte los archivos se componen únicamente de caracteres ASCII y ocupan poco; y por otra, para producirlos se puede utilizar cualquier editor de texto.

TEX, la máquina de composición de LaTex, está disponible en todos los sistemas operativos de escritorio y es gratis. Por esto, el sistema funciona prácticamente en cualquier plataforma.

5.3 FORMATOS MIXTOS

5.3.1 MRC – Mixed Raster Content

El MRC (Mixed Raster Content, contenido mixto de raster) es un modelo de tres capas desarrollado inicialmente como una manera de comprimir imágenes en color para ser transmitidas por fax.

Los fax a color no prosperaron, pero el modelo MRC 3-layer fue implementado dentro de un número de formatos para almacenamiento y transmisión de imágenes. Entre estos formatos encontramos al TFX (TIFF-FX) y DjVu.

El MRC es llamado tambien ITU T.44 o color fax protocol.

Las tres capas del modelo MRC son: Frente, Fondo y la de máscara de imagen[8]. La capa de Fondo usualmente tiene una imagen y la de Frente gráficos y texto coloreado. La máscara de imagen en cambio tiene formas de texto o similares, es decir bien definidas y de alto contraste.

El texto entonces puede ser comprimido en menor cantidad de colores y a una mayor resolución, por ejemplo 1-bit por píxel a 600 DPI. Las imágenes, en cambio, requieren menor resolución pero mayor cantidad de bits por píxel.

Para comprimir una imagen se usa un algoritmo para imágenes bi-tonales sin pérdida de información para la capa de texto (JBIG o JBIG2 por ejemplo) mientras que para el frente y fondo, que contienen imágenes, se usa un algoritmo con pérdida de información que son los que mejor se adaptan, por ejemplo JPEG o JPEG2000.

El proceso entonces es el siguiente: el documento es analizado y descompuesto en imágenes y texto. Cada uno de estos son identificados, analizados y comprimidos independientemente eligiendo la forma óptima en cada caso.

La implementación mas avanzada del MRC es DjVu[9], que a la especificación original le agrega elementos de texto en una capa adicional con la posibilidad de búsqueda e hipervínculos. Otras implementaciones, como LuraDocument, no incluyen una capa de texto.

El modelo MRC 3-Layer es seguido también en la nueva versión de Adobe Acrobat para los PDF. Esta implementación, llamada Adaptive Capture, se posiciona favorablemente frente al DjVu porque obtiene mejores factores de compresión y fundamentalmente porque se monta sobre un formato tan usado como el PDF.

En la ilustración 5-2 vemos un gráfico descompuesto en tres capas compositoras.

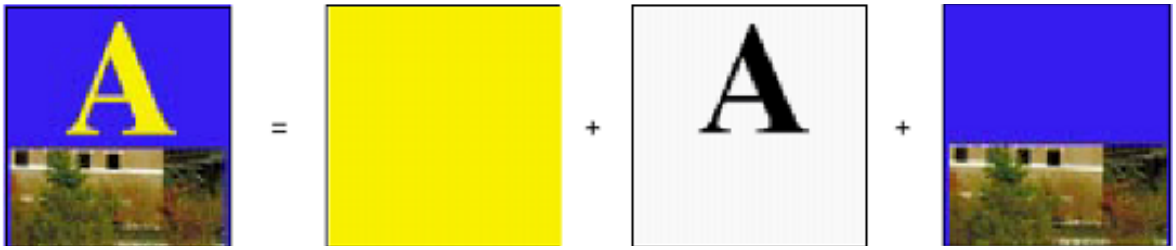


Ilustración 5.2

8. Los nombres en inglés son: Foreground, Background e ImageMask.
9. Ver 5.3.1.1

- Las secciones de texto son identificadas como símbolos que son guardados en una biblioteca de símbolos.
- Las imágenes line art (cuadros, tablas) son convertidas en vectores.
- Las imágenes blanco y negro (gráficos, dibujos) se comprimen con algoritmos adecuados para ese tipo de imágenes; G4, JBIG o JBIG2 (se explican más adelante).
- Las imágenes en escala de gris o color son comprimidas con JPEG o JPEG2000, que es el método más eficiente para este trabajo.

Éstas técnicas combinadas dan un factor de compresión de 100 a 1 o mayor.

5.3.1.1 DjVu

DjVu (pronunciado como déjà vu) fue desarrollado por AT&T Labs en 1996. Es un formato de documento que incluye varias diferentes técnicas de compresión.

DjVu fue diseñado para ser una tecnología que se aplique tanto a texto en blanco y negro, contenido en escala de grises y color.

Usa un esquema en capas que permite que los documentos combinen texto e imágenes y aplica a cada capa un algoritmo en forma separada para hacer mas eficiente el documento y mostrarlo en forma óptima. Esto significa que aplica algoritmos lossy (con pérdida de información) para las imágenes y lossless (sin pérdida de información) para las partes en blanco y negro.

DjVu maneja archivos de una y varias páginas. La forma de visualizarlo es por medio de un programa de distribución gratuita.

Un problema de DjVu es que es un formato propietario. LizardTech, que compró los derechos a AT&T, es la única fuente de soluciones comerciales que adhieren a este estándar. Por eso, a pesar de las ventajas que tiene, la adopción es muy lenta. Otro inconveniente es que las capacidades de almacenar meta datos son muy limitadas.

5.3.1.2 PostScript

En 1985 Adobe creó un lenguaje de programación de impresoras llamado PostScript. El lenguaje describe la apariencia de la página incluyendo elementos de texto, gráficos, colores e imágenes y de esta manera la página mantiene su integridad en su transmisión de la computadora a la impresora. Desde ese momento las impresoras que entienden el lenguaje PostScript se convirtieron en la norma para corporaciones y para el ambiente del diseño gráfico.

5.3.1.3 PDF

El formato PDF (PDF es acrónimo de Portable Document Format, en castellano: formato portable de documento) fue creado en el año 1993 por Adobe para complementar al formato PostScript. El PDF es una versión reducida del PostScript. Los documentos creados en este formato son usualmente hechos en formatos intermedios y se convierten a PDF como último paso en la cadena de creación para su distribución entre los usuarios finales.

La manera de leer los PDF es con el Acrobat Reader, software de lectura de Adobe y que se distribuye gratuitamente para prácticamente todos los sistemas operativos de escritorio pudiendo bajar la última versión desde la página de Adobe Systems. También existe un plug-in para los browsers de Internet más populares (MS Explorer, Netscape Navigator) que permite leer estos archivos de forma online.

El programa de Adobe para crear un PDF se llama Acrobat y, a diferencia del Acrobat Reader, no es gratuito. Esto limita un poco la producción de documentos en este formato. Afortunadamente existen alternativas al Acrobat que surgen del movimiento de software Open Source que cumplen la tarea adecuadamente con la ventaja de ser gratuitas. Entre ellas podemos encontrar al OpenOffice.org, al PDFCreator y al Ghostscript.

El formato PDF es una especificación abierta de formato de archivo que se encuentra a disposición de todas aquellas personas que deseen desarrollar herramientas para crear, ver o modificar documentos. Existen más de 1.800 proveedores que ofrecen soluciones basadas en el formato PDF, lo que garantiza que las organizaciones que adoptan el estándar PDF tengan a su disposición una gran variedad de herramientas para aprovechar el formato PDF y personalizar los procesos con documentos.

Entre las características del formato PDF se destaca que se permiten asignar derechos de acceso especiales y pueden firmarse digitalmente asegurando que la copia distribuida es la original y no se puede alterar.

Debido a estas características gobiernos y empresas de todo el mundo adoptaron el formato PDF. Por ejemplo, el formato PDF es el estándar utilizado para la entrega electrónica de solicitudes para la

aprobación de medicamentos a la FDA (del inglés Food and Drug Administration, administración de alimentos y medicamentos) de los Estados Unidos. También lo utilizan los gobiernos de Reino Unido y Alemania para el intercambio electrónico de documentos. Asimismo, la especificación PDF/X de ISO es el formato de archivo estándar utilizado para la distribución digital de anuncios para su posterior publicación. Es el formato utilizado mayoritariamente por los fabricantes para la distribución de manuales tanto de hardware como de software. Muchos de los e-books que existen en el mercado son comercializados con este formato.

El PDF en si mismo no es ni un formato de imagen ni un método de compresión. A pesar de eso, PDF puede funcionar como contenedor de imágenes digitales comprimidas con distintos esquemas de compresión, incluyendo G4 y JBIG2 (se verán mas adelante).

PDF soporta documentos simples o de múltiples páginas.

A pesar de ser un formato propietario, Adobe mantiene la especificación PDF abierta, dando como resultado una gran cantidad de aplicaciones de terceros que lo soportan.

Algunas de sus ventajas también son desventajas. La poca amabilidad para permitir modificaciones en este tipo de documento los hace sólo utilizables en el caso de originales que no deben ser editados por los usuarios.

Otra desventaja es la lectura ya que, si bien es un formato excelente para la impresión gracias a la fidelidad que guardan con la apariencia del original, en la pantalla de una computadora debido a su geometría apaisada en contraposición con las hojas de impresión que suelen ser verticales. Tampoco es posible acomodar el tamaño de las letras haciéndolos a veces difíciles de leer.

Otras desventajas: tiene una pobre capacidad de almacenar meta datos internamente y no es legible con herramientas estándar haciendo necesario un plug-in o una aplicación especial para su lectura.

5.3.1.4 Microsoft LIT / Microsoft Reader

Microsoft saco su Microsoft Reader y junto con él su formato propietario, el .LIT.

El formato .Lit es derivado del estándar Open E-Book, basado en XML y luego compilado. Una vez compilado no puede volverse al archivo abierto. Debido a que es un formato cerrado y solo legible a través del software de Microsoft no se considera su uso en una biblioteca.

Existen herramientas para convertirlo desde los formatos más populares de edición de texto - .rtf, Word, txt etc- a .lit.

5.4 COMPARACIÓN ENTRE FORMATOS

En base a lo visto anteriormente se puede organizar una matriz dando valores a cada formato de archivo en los distintos campos evaluados. Los valores van del 1 al 5 siendo 5 el máximo grado y 1 el mínimo.

	Facilidad	Herramientas	Estilo	Control sobre impresion	Facilidad de uso	Licencia
Texto	5	4	3	2	3	4
RTF	5	5	4	2	3	3
HTML	1	3	2	5	5	5
SGML	1	3	2	3	2	5
XML	4	4	3	3	4	5
PDF	3	2	5	5	4	2

6. METADATOS

Se podría definir a los metadatos como “datos estructurados acerca de otros datos”. Estos datos pueden ser ingresados a mano o generados automáticamente y pueden estar incluidos dentro del archivo o en un archivo adjunto. En el contexto de los libros en formato digital, los datos típicos que se utilizan son el origen del documento, su creador, el formato en que está creado, ubicación, fecha de creación, versión, categoría etc. Cuando surgieron los primeros documentos digitales los esfuerzos estaban concentrados en el proceso de creación sin preocuparse demasiado en el proceso de búsqueda y recuperación. Sin embargo, a medida que aumenta la cantidad de información se vuelve crucial la metainformación para poder catalogar y buscar adecuadamente lo creado y hacer uso efectivo del conocimiento.

Podemos clasificar a los metadatos en tres grandes categorías: descriptivos, estructurales y administrativos. Estas categorías suelen superponerse y el límite entre unos y otros es bastante difuso. Por ejemplo, entre los metadatos administrativos hay muchos que también pueden ser considerados descriptivos.

6.1 METADATOS DESCRIPTIVOS

Se usan para la descripción e identificación de recursos de información. Permiten la búsqueda y recuperación de los documentos en una base de datos local.

Los campos que incluyen son los de atributos bibliográficos: título, autor, creador, idioma, traductor, editorial, palabras clave etc.

Entre las implementaciones de este tipo de metadatos podemos encontrar a:

· Dublín Core
· Tei Header
· MARC

6.2 METADATOS ESTRUCTURALES

Son usados para facilitar la navegación y presentación de los documentos electrónicos. Proporcionan información sobre la estructura interna de los archivos, incluyendo paginados, sección, capítulos, numeración, índices, árboles de contenido, comentarios, fe de erratas etc. También la relación con los sub-objetos como los gráficos dentro de los archivos.

Las implementaciones de estos metadatos pueden realizarse con varios formatos, por ejemplo:

· SGML
· XML
· MOA2, Structural Metadata Elements (Elementos de Metadatos Estructurales)

6.3 METADATOS ADMINISTRATIVOS

Son usados para la gestión y el procesamiento de las colecciones digitales. Contienen información sobre datos técnicos de la creación y el control de calidad (tipo y modelo de escáner, resolución y profundidad de color, software de OCR usado, compresión) y control de acceso y derechos (propietario, derechos de autor, limitaciones en cuanto a copiado).

A pesar del consenso que existe en dar importancia vital a la metainformación no se ha impuesto ningún modelo que satisfaga las necesidades de todos. Actualmente existen tres modelos en uso bastante generalizados: el Dublin Core Element Set, MARC y el TEI Header.

6.4 ENCABEZADO TEI - TEI HEADER

TEI viene de las siglas Text Encoding Initiative o Iniciativa para la codificación de textos. El encabezamiento TEI provee un mecanismo para documentar todos los aspectos de un texto electrónico pero no se limita a esto, también sirve para documentar la fuente de información, el método de digitalización y su proceso de creación. Por lo tanto esta cabecera se convierte en un recurso esencial para los lectores del texto, el software que procesa la información y los catalogadores de las bibliotecas, museos y archivos.

Las Normas TEI usan el Standard Generalized Markup Language (SGML) para definir su esquema de codificación. El uso que hace el TEI del SGML no es diferente del de cualquier otro esquema de marcado en SGML. Por lo tanto, cualquier programa preparado para SGML puede procesar los textos que cumplan el TEI.

El TEI está patrocinado por la Association for Computers and the Humanities, la Association for Computational Linguistics, y la Association for Literary and Linguistic Computing y sus Normas fueron publicadas en mayo de 1994, tras seis años de desarrollo donde participaron cientos de estudiosos de diferentes disciplinas académicas de todo el mundo.

6.4.1 La estructura de un texto TEI

Todos los textos que cumplan el TEI contienen como mínimo:

· un encabezado TEI (marcado con el elemento <teiHeader>)
· la trascripción del propio texto (marcado con el elemento <text>).

El encabezado TEI provee información similar a la de la contratapa de un texto impreso:

· una descripción bibliográfica del texto electrónico
· una descripción de cómo ha sido etiquetado
· una descripción no bibliográfica del texto (un perfil del texto)
· una revisión de su historia (su creación).

Un texto TEI puede ser individual (una única obra) o compuesto (una colección de obras, como por ejemplo una antología). En cualquier caso, el texto puede tener un front o back opcional. En medio está el body, cuerpo de la obra, que, en el caso de un texto compuesto, puede estar formado por groups, cada uno conteniendo a su vez más grupos o textos.

Un texto individual se etiquetará siguiendo una estructura genérica como esta:

```
<TEI.2>
   <teiHeader> [ Información del encabezado TEI ]  </teiHeader>
   <text>
      <front> [ materia del front ... ]   </front>
      <body>  [ cuerpo del texto ... ]   </body>
      <back>  [ materia back ...  ]   </back>
   </text>
</TEI.2>
```

6.5 MARC

MARC es el acrónimo de MAchine-Readable Cataloging. Define un formato que fue desarrollado por la librería del congreso de los Estados Unidos en los últimos treinta años y provee el mecanismo para catalogar y usar información bibliográfica por medio de computadoras. MARC se convirtió en USMARC en 1980 y en MARC 21 a fines de los ´90.

La estructura de un encabezado MARC está compuesta por tres componentes:

- Leader: Campo de longitud fija de 24 caracteres. Contiene información codificada para el catálogo.
- Directory: El Directorio es un índice de los datos dentro de un registro. Contiene información sobre el marcado, largo de los campos y localización de cada campo dentro del archivo. El directorio termina con un caracter de cierre de campo.
- Campos variables: Los datos de un registro MARC están organizados en campos variables los cuales varían en longitud y contenido, dependiendo del campo. Cada uno está identificado con un número de tres dígitos, que es almacenado en el directorio. Cada campo termina con un carácter de cierre de campo. El último de los campos termina con un cierre de campo y un cierre de registro. (ASCII 1D hex).

Los desarrolladores de MARC crearon un framework para trabajar con MARC en entornos XML y SGML.

También existen herramientas de conversión para convertir el MARC a otros formatos y viceversa, por ejemplo de Dublín Core a MARC y de MARC a Dublín Core llamada marc2dc.

6.6 LA INICIATIVA DUBLIN CORE

La iniciativa de metadatos Dublín Core o DCMI (Dublín Core Metadata Initiative) es una organización dedicada a promover la adopción de estándares de metadatos y a desarrollar vocabularios para los descriptores.

El primer juego de meta datos desarrollados por el Dublín Core se hizo en 1995 y es el Dublin Core Metadata Element Set (DCMES) y describe elementos básicos de información de un documento, como Descripción, Creador o Fecha.

Las características que lo distinguen de otros estándares son diversas:

- Simplicidad
- Consenso internacional
- Extensibilidad

La tabla 2 muestra los campos de una cabecera Dublín Core y el tipo de dato de cada uno.

Tabla 2

Elemento	Tipo de campo
Title	String
Identifier	String
Latest version	Url
Creador	String
Contributor	String
Date issued	Date
Date created	Date
Date modified	Date
Supersedes	Urls
Is superseded by	Urls
File format	Enumerated list
Language	Enumerated list
Description	Memo
Status	String

Fuente: http://dublincore.org/

7. SOFTWARE PARA OCR

En esta sección se explicará la finalidad y el funcionamiento de un programa de OCR.

7.1 COMO FUNCIONA EL OCR

Para reconocer las letras los programas de OCR tienen tres métodos principales:

· Comparación de patrones
· Extracción de características
· Comparación con diccionario

7.1.1 Comparación de patrones

La mayoría del texto está impreso en tamaño de entre 10 y 14 puntos y en tipografías Times new roman, Courier o Helvética. Los programas de OCR, entonces, se aprovechan de esta situación y guardan todas las letras de las tipografías más comunes y en los tamaños más comunes para poder compararlas con los símbolos encontrados y de esta manera intentan reconocerlos. La limitación de este método es que sólo está acotada a los símbolos guardados previamente.

La ilustración 7-1 muestra una letra a en una grilla.

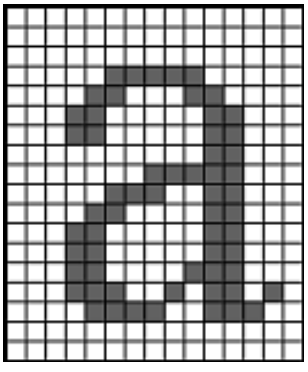


Ilustración 7.1

7.1.2 Extracción de características

La comparación de patrones no siempre da buenos resultados. Ésta técnica intenta extraer las características básicas de los símbolos encontrados y las compara con las características de las letras que tiene en su base de datos.

En el ejemplo de la letra a, las principales características son las de tener un circulo, una línea vertical en el lado derecho y un semi arco en la parte superior. El arco superior sin embargo es opcional. Entonces, si un símbolo escaneado tiene estas características será identificado por el programa de OCR como una letra a.

La ilustración 7-2 es una letra a formada por vectores.

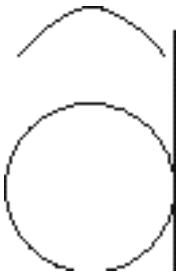


Ilustración 7.2

7.1.3 Comparación con diccionario

No importa cuan bien aplicados estén los métodos anteriores, no siempre es posible reconocer el 100% de las letras de manera correcta. Una vez que el reconocimiento inicial se termina, el software de OCR intenta reconocer a las letras no reconocidas mirando a las letras que las rodean. Si encuentra la palabra BIBLIO#ECA, y no puede reconocer el símbolo #, usando un diccionario puede darse cuenta que la palabra a buscar es BIBLIOTECA y que la letra faltante es la T.

7.1.3.1 Reconocimiento de escritura manual

Actualmente el reconocimiento de escritura manual no está en estado de ser usado masivamente. Existen soluciones ad hoc que pueden servir en algunos casos especiales pero no es posible digitalizar un documento manuscrito de la misma manera que uno impreso.

Existe un proyecto patrocinado por Google Inc. y la Universidad de Oxford para digitalizar el catalogo de la universidad pero este aún se encuentra en una etapa de análisis preliminar y se estima que para agosto de 2005 se comenzará con el desarrollo, con una duración estimada de tres años[10].

En la ilustración 7-3 se aprecia como se seccionan las palabras de un manuscrito para ser analizadas individualmente.

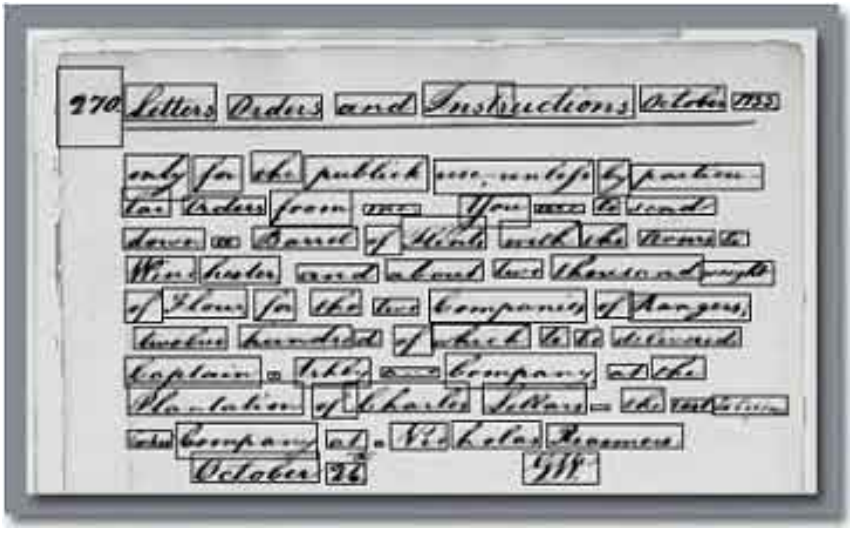


Ilustración 7.3

10. http://www.bodley.ox.ac.uk/google/
http://www.google.com/intl/en/press/pressrel/print_library.html
http://print.google.com/googleprint/screenshots.html

7.1.4 OmniPage

OmniPage es uno de los programas líderes en el mercado. La última versión, a marzo de 2005, es la número 14.

OmniPage cuenta con las siguientes características:[11]

· Módulo quitamanchas

El módulo quitamanchas reduce la distorsión del fondo, lo cual permite una fácil conversión de documentos sombreados y en colores que antes eran irreconocibles.

· Impresión en PDF

La función de impresión en formato PDF sirve para convertir archivos de cualquier aplicación en archivos PDF con capacidad de búsqueda, que conservan el formato exacto del original.

· Administrador de proceso en lotes (batch)

El procesamiento de documentos en lotes sin intervención del usuario permite abocarse a otras tareas mientras OmniPage realiza el trabajo de OCR. Puede buscar carpetas continuamente en la red, realizar el procesamiento de OCR en todos sus contenidos y enviar los resultados a otras carpetas locales o en la red. Es la solución perfecta para organizaciones que necesitan convertir grandes volúmenes de documentos en archivos electrónicos de fácil recuperación.

· Reproducción en voz alta con RealSpeak™

OmniPage no sólo reconoce documentos, también los “lee” en voz alta. La tecnología galardonada de reproducción en voz alta RealSpeak permite verificar con los oídos, no sólo con los ojos.

· Salida XML

Utilice OmniPage para transformar documentos impresos e imágenes en datos “inteligentes” que, una vez vinculados al esquema, se pueden utilizar en un sinnúmero de formas distintas. La nueva versión agrega compatibilidad con Word ML.

7.1.5 SimpleOCR

Simple OCR es un programa con licencia gratuita para usos no comerciales.

Sus principales características son:

- Diccionario de 120.000 palabras. Éste está en ingles, pero es posible agregar un diccionario en castellano.
- Limpieza de imágenes. Para evitar errores con los originales que tienen manchas o suciedad.
- Retención de formato. Intenta mantener el formato del original escaneado.
- Retención de imágenes. Importa directamente las imágenes encontradas dentro del nuevo documento ahorrando tiempo.
- Extracción a texto plano. Permite exportar sólo el contenido, sin formato ni imágenes. Ideal para adaptarla a un formato propio.
- Batch OCR. Procesa varias hojas a la vez sin intervención del usuario; de esta manera se automatiza el proceso.

7.1.6 Kooka – OCRad

Kooka[12] es una interfaz para un motor (engine) de OCR. El motor de OCR que usa es Ocrad[13] y juntos forman un paquete que funciona como un programa completo para reconocimiento de caracteres.

A diferencia de los anteriores mencionados, Kooka-Ocrad funciona bajo el sistema operativo Linux.

A pesar de la dificultad que puede llegar a representar la instalación y configuración de una terminal Linux para un usuario no familiarizado con el sistema operativo existe la alternativa de usar un CDVivo o LiveCD[14], que es un sistema operativo que se ejecuta directamente desde un CD-ROM sin la necesidad de instalarlo en el disco rígido de la computadora.

11. Información recolectada del sitio oficial de Omnipage: http://spain.scansoft.com/omnipage/features.asp
12. Más información en: http://www.kde.org/apps/kooka/
13. Más información en: http://www.gnu.org/software/ocrad/ocrad.html
14. Ver glosario

Ilustración 7.4 – Pantalla de reconocimiento de Kooka

7.1.7 Libre vs. Comercial

A diferencia de otras categorías de software (editores, por ejemplo) en OCR no hay alternativas gratuitas con el mismo nivel de servicio que los programas comerciales.

8. HERRAMIENTAS DE DECISIÓN

Los elementos del proceso decisorio se formalizan en dos herramientas básicas: árboles de decisión y matrices de decisión.

Respecto de los primeros, es conveniente utilizarlos cuando se trata de situaciones de decisión de tipo secuencial o cíclica. Las matrices, en cambio, se aplican ante situaciones que implican una decisión única.

Teóricamente se utiliza una matriz para cada decisión única en un momento dado, de modo que una serie de decisiones entrelazadas en futuro implican una matriz para cada una. En tal caso es conveniente utilizar un árbol de decisión que sustituye la serie de matrices individuales.

Un árbol puede sustituir a un conjunto de matrices y viceversa. La elección entre uno y otra se basa en la comodidad para el análisis

8.1 LA MATRIZ DE DECISIÓN

Las matrices de decisión son una herramienta de formalización del proceso decisorio que ayuda a ordenar los elementos actuantes. Constituyen el punto final del proceso. Los datos que exhiben son el producto de la búsqueda y análisis de la información y de la definición de todos los elementos de este proceso. Al lado de la matriz de decisión se debe mostrar como se ha llegado a las cifras del final.

La matriz de decisión esta constituida por filas donde se expresan las alternativas y por columnas que representan los estados, niveles o grados de las variables inciertas consideradas en el problema de la decisión. En las celdas formadas por la intersección de filas y columnas se expresan los resultados de la elección de una alternativa y la ocurrencia de un determinado estado en las variables inciertas consideradas. Se debe tener presente que tales resultados son la medida de la consecución de los objetivos.

8.2 EL ÁRBOL DE DECISIÓN

El árbol de decisión constituye una de las principales herramientas de la teoría de la decisión y se los utiliza cuando la situación es analizada como una serie de decisiones concatenadas.

La representación gráfica de la situación de decisión con el ordenamiento de un árbol es factible y provechosa en las decisiones secuenciales.

Los árboles de decisión pueden representarse como un conjunto de reglas Si – Entonces, es decir, si sucede tal condición entonces se debe tomar tal camino.

Un árbol típico de decisión se compone de dos elementos: los nodos y los arcos que los conectan. Existen dos tipos de nodos: de decisión, que marcan la toma de una decisión, y los de incertidumbre, que marcan el acontecimiento de estados de las variables inciertas.

De los primeros parten tantos arcos como alternativas existan. Estos arcos son denominados por la alternativa a la cual correspondan. De los nodos de incertidumbre parten tantos arcos como eventos inciertos se esperen.

Un conjunto de arcos y los nodos que unen en forma correlativa constituyen una rama, un conjunto de ramas, un camino. Un subárbol es el árbol que se inicia en algunos de los nodos del árbol original.

En los arcos de decisión, se registran las alternativas imaginadas. En los de sucesos inciertos estos son registrados pero, además, se consignan las probabilidades de cada uno (si es que se conocen) y los resultados esperados.

En cada nodo solo puede entrar un arco. Los arcos de salida pueden ser, en cambio, uno o varios.

En las ilustraciones 8-1 y 8-2 vemos ejemplos de árboles de decisión.

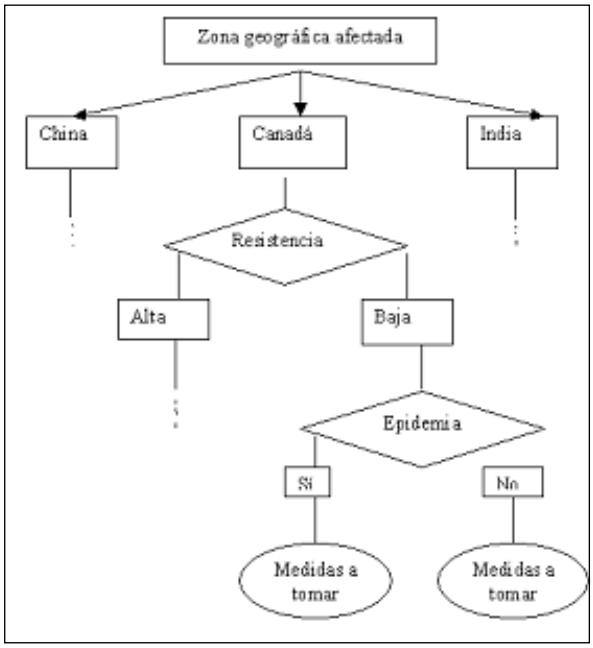


Ilustración 8.1

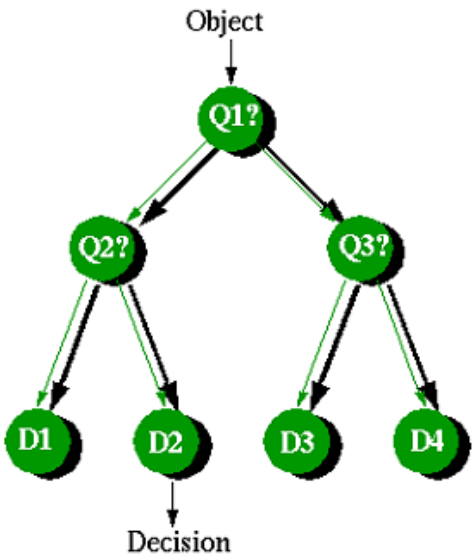


Ilustración 8.2 – Ejemplo de árbol de decisión.
Fuente: www.mindtools.com

8.3 ¿CUAL USAR? AMBOS!

En este trabajo se van a utilizar matrices de decisión para poder comparar los diferentes productos y tecnologías entre sí. Luego se desarrollará un árbol que las combine adecuadamente.

9. HARDWARE DE DIGITALIZACIÓN

Hay varios métodos de digitalización de imágenes comúnmente usados. El equipamiento varía desde escáneres de todo tipo hasta cámaras digitales de alta resolución. Sin embargo los más comunes son los escáneres flatbed y las cámaras digitales.

9.1 ESCÁNER[15]

El escáner es el dispositivo más adoptado para la digitalización de documentos por su bajo costo en relación con otras tecnologías y a su versatilidad para adaptarse a distintos trabajos.

9.1.1 Funcionamiento de un escáner

Cuando comienza el proceso de digitalización el dispositivo empieza a leer la primera línea del documento. Una vez que los datos de la primera línea son recolectados entonces un motor dentro del escáner mueve el sensor a la segunda línea comenzando el barrido de la página. Durante el barrido se mueve una luz fluorescente emitida por una lámpara ubicada sobre la cabeza del sensor. Esta luz atraviesa el vidrio de apoyo, llega al documento y es reflejada en la imagen. En este momento entra en juego el sensor (CCD[16] o CIS) que se encarga de recolectar la luz. Entonces el sensor analiza la luz entrante y la transforma en información digital dándole valores a los niveles de luz para ser representados como una imagen digital. Usualmente estos valores están entre 0 y 255 (28, 8-bit) para escala de gris o en caso del color se usan 8 bits por cada canal de color RGB dando como resultado una imagen de 24-bit. Algunos escáneres usan 12, 14 o 16 bits por canal para obtener una imagen más fiel al documento original.

9.1.2 Escáner Flatbed

Los escáneres flatbed se convirtieron en el método más común de capturar imágenes o texto. El nombre (traducido literalmente significa cama plana) deriva de su construcción, ya que el equipo tiene una superficie plana de vidrio sobre la cual se apoyan los originales boca debajo de manera similar a una fotocopiadora. La figura 9-1 es una descripción del funcionamiento de un escáner flatbed.

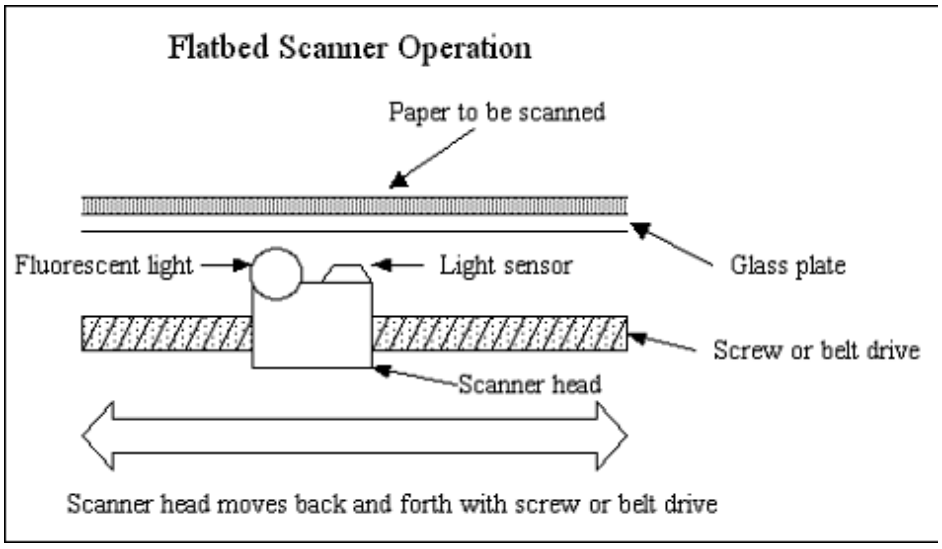


Ilustración 9.1

9.1.3 Escáner de Alimentación automática

Los escáneres de alimentación automática utilizan una tecnología similar a la de los Flatbed pero tienen adosado un alimentador automático de documentos que ingresa las hojas a medida que se necesitan. Estos alimentadores pueden, en algunos casos, ser agregados en forma de accesorio a un escáner flatbed. De todas maneras estos escáner pueden ser usados de a una página a la vez cuando en lugar de hojas sueltas se necesitan digitalizar hojas encuadernadas.

15. Se usa la palabra Escáner en vez de la voz inglesa Scanner.
16. Ver Glosario

Algunos escáneres sheet-fed (el término inglés para describir la alimentación automática) están pensados para cargas de trabajo más altas que las hogareñas y muchas veces suelen funcionar al mismo tiempo como fotocopiadora (o se puede usar la fotocopiadora como escáner).

Existen también escáneres muy pequeños en tamaño que en vez de usar una cama de vidrio (flatbet) para apoyar el documento usan un método en el que la hoja es la que se mueve de la misma manera que un FAX. Esta morfología los hace inútiles para su uso con un libro encuadernado.

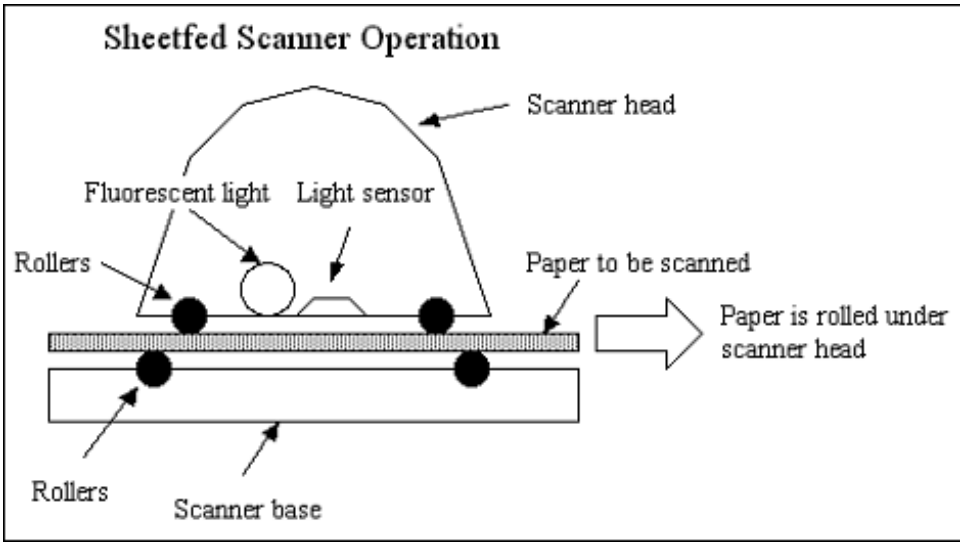


Ilustración 9.2

9.1.4 Escáner de tambor (Drum Scanner)

Este escáner presenta forma de tambor, el cual se rota para hacer la lectura de películas fotográficas, por lo tanto, sólo se usa con negativos o diapositivas. Su fuente de luz es un láser que se encuentra dentro del tambor, y la luz que se transmite a través de la película es medida por el detector del Tubo Foto Multiplicador el cual se encuentra en la parte exterior del tambor. Esta tecnología es más sensible que la tecnología del CCD y por lo tanto proporciona mejores detalles de color en las áreas de las sombras de la imagen. Una de las ventajas de este sistema es que se puede calibrar con precisión a los parámetros de la impresora que se utilizará.

9.1.5 Fotocopiadoras

Muchas fotocopiadoras funcionan también como scanner y suelen tener un alimentador de hojas incorporado. A pesar de no ser un dispositivo recomendado para su uso exclusivo en digitalización, dado que muchas empresas tienen un equipo de estos en su poder pueden aprovecharlo para la tarea.

9.2 CÁMARA DIGITAL

Uno de los problemas que existen con los escáneres flatbed es que para poder registrar la información hay que apoyar la superficie del original completamente plana sobre el vidrio del escáner. Con los libros esto es un problema porque la única manera de lograrlo es forzando el lomo. Esto puede ser aún peor cuando tratamos con libros frágiles de valor histórico. Una solución a este problema, muy usada en catálogos y algunos archivos digitales, es una cámara digital.

Las cámaras pueden tomar imágenes directamente del original sin interferir y pueden trabajar con objetos de cualquier forma y tamaño con muy alta resolución. Estas características son particularmente beneficiosas para libros antiguos o manuscritos delicados.

La resolución de las cámaras se mide en mega píxeles o MPx que representa la cantidad de píxeles que contiene una imagen. Las cámaras no especializadas tienen resoluciones de entre 2 y 8 mega píxeles.

En la tabla 4.1 se comparan las áreas que pueden abarcar las cámaras digitales obteniendo una imagen a 300 DPI. De esta tabla se puede deducir que para digitalizar una hoja A4 se debería utilizar una cámara de al menos 8 mega píxeles.

Tamaño en Mega píxeles	Dimensiones	Área aproximada a escanear a 300DPI
2	1600*1200	13 x 10 cm.
3	2048*1536	16 x 12 cm.
4	2272*1704	18 x 14 cm.
5	2592*1944	20 x 16 cm.
8	3264*2448	25 x 20 cm.

Tabla 9.1 – Área a digitalizar por una cámara digital

Para ser usada en un proyecto de digitalización se debe montar la cámara en un trípode ubicado convenientemente que permita tomar la totalidad de la página y ésta debe ser iluminada correctamente para poder captar los detalles.

9.3 TIEMPO VS. DINERO

Un scanner especializado –por ejemplo el Xerox WorkCentre™ M24[17], con valor de USD 1200 en Argentina- tiene un alimentador automático de 50 hojas y puede escanear a 20 páginas por minuto.

Como contrapartida un scanner hogareño – Vg. HP 2400 Scanjet, USD 85 en Argentina- puede escanear aproximadamente 1 página por minuto.

Dada la disparidad de las opciones antedichas se puede ver claramente que el volumen de páginas a digitalizar y la cantidad de mano de obra disponible son claves en el factor de decisión.

9.4 COMPARACIÓN ENTRE DISPOSITIVOS

Dispositivo	Ventajas	Desventajas
Escáner Flatbed	• Alta disponibilidad • Precio menor a USD 100 • Fácil de operar • Alta resolución – 1200 dpi o más.	• Productividad Baja • Alta atención del operador • Los escáneres accesibles suelen tener colores inexactos
Escáner de Alimentación Automática	• Alta productividad • Igual o mejor calidad que los Flatbed • Automatización de tareas	• No usable con hojas frágiles o libros. • Caros • Solo manejan tamaños de documentos predefinidos
Escáner de Negativos	• Muy productivos para rollos de negativos o diapositivas • Muy alta resolución – 2700 dpi a 4000 dpi.	• Muy sensibles a las ralladuras en la película. • Lentos para la alimentación de negativos o positivos enmarcados

17. http://www.office.xerox.com/perl-bin/product.pl?product=WCM24&page=spec

Escáner de Tambor	• Calidad extremadamente alta	• Muy caros
Cámaras Digitales	• Pueden digitalizar una mayor variedad de documentos (voluminosos, de tamaños grandes, frágiles, libros antiguos) • Captura sin contacto físico • Tamaño de captura ilimitado • Calidad aceptable	• Se requieren técnicas básicas de fotografía

Tabla 9.2

10. ANÁLISIS DEL DOCUMENTO

10.1 DEFINIENDO EL DOCUMENTO

Para poder clasificar el documento a digitalizar se tienen en cuenta sus características técnicas y el tipo de información que guarda. Por características técnicas se entiende el tipo de encuadernación, la cantidad de páginas a digitalizar, la tipografía y tamaño utilizados en caso de ser texto o el tipo de gráficos o imágenes y el valor del ejemplar a digitalizar (se puede estar digitalizando un incunable, un manuscrito valioso o un libro disponible en una librería)

10.1.1 Requerimientos según el tipo de documento

Las características del documento son las que definen la resolución y el formato de archivo a utilizar en la etapa de digitalización.

Si es texto que va a ser luego procesado con OCR se tienen los siguientes requerimientos:

- **Texto – Blanco y negro**
  - Resolución: 400 dpi
  - Profundidad: 8-bit
  - Formato: TIFF
  - Compresión: Sin compresión o compresión Lossless

- **Texto - Color**
  - Resolución: 400 dpi
  - Profundidad: 24-bit
  - Formato: TIFF
  - Compresión: Sin compresión o compresión Lossless

- **Textos Antiguos o especiales**
  - Resolución: 400 dpi
  - Profundidad: 8-bit o 24-bit
  - Formato: TIFF
  - Compresión: Sin compresión o compresión Lossless

Si se van a digitalizar imágenes, los requerimientos son:

- **Imágenes - Blanco y negro**
  - Resolución: 600 dpi
  - Profundidad: 1-bit
  - Formato: TIFF
  - Compresión: Sin compresión o compresión Lossless

- **Imágenes – Escala de gris**
  - Resolución: 300 dpi
  - Profundidad: 8-bit
  - Formato: TIFF o JPEG2000
  - Compresión: Sin compresión o compresión Lossless

- **Imágenes - Color**
  - Resolución: 300 dpi
  - Profundidad: 24-bit
  - Formato: TIFF o JPEG2000
  - Compresión: Sin compresión o compresión Lossless

En la práctica, al digitalizar se puede tener más de una versión del mismo documento. Esto es más frecuente en las digitalizaciones de documentos que requieren ser facsímiles de la obra. La idea detrás de estas múltiples versiones es que se puede tener una para archivo -que debido al gran tamaño puede ser poco adecuada para su distribución por la red- llamada master y otras de distribución optimizadas para los distintos usos: impresión, búsqueda, catalogo, publicación en Internet, uso en intranet etc.

11. EL FACTOR ECONÓMICO

11.1 ¿VALE LA PENA LA INVERSIÓN?

Un proyecto de este tipo implica un gasto de recursos proporcional a la magnitud del mismo. En caso de elegir hardware especializado y software con licencias es imprescindible contar con el capital para poner en marcha la digitalización. Si en cambio la elección es usar hardware preexistente o de propósitos más generales y programas gratuitos la erogación inicial puede ser menor, pero seguramente se deberán invertir mas recursos en capacitación y en personal que se dedique al trabajo, ya que el flujo de trabajo va a ser más lento y menos eficiente.

12. FACTORES DE DECISIÓN

12.1 TIPO DE DOCUMENTO

Los tipos de documentos a digitalizar pueden separarse en:

- Manuales o libros de texto sin imágenes
- Manuales o libros de texto con imágenes (Mixtos)
- Manuscritos o documentos firmados
- Libros de fotografías o imágenes

12.1.1 Análisis visual y estructural

Tipos de documentos

- Texto impreso / Dibujos de líneas simples — representación en base a bordes definidos, sin variación de tono, como un libro que contiene texto y gráficos de líneas simples.
- Manuscritos — representaciones en base a bordes suaves que se producen a mano o a máquina, pero no exhiben los bordes definidos típicos de los procesos a máquina, como el dibujo de una letra o una línea.
- Media Tinta — reproducción de materiales gráficos o fotográficos representados por una cuadrícula con un esquema de puntos o líneas de diferente tamaño y espaciadas regularmente que, habitualmente se encuentran en un ángulo. También incluye algunos tipos de arte gráfica, como por ejemplo, los grabados.
- Tono Continuo — elementos tales como fotografías, acuarelas y algunos dibujos de líneas finamente grabadas que exhiben tonos que varían suave o sutilmente.
- Combinado — documentos que contienen dos o más de las categorías mencionadas anteriormente, como por ejemplo, los libros ilustrados.

12.2 VOLUMEN DE LA OBRA

Es importante el tamaño (en cantidad de páginas) de las obras a digitalizar. No es lo mismo digitalizar 20 libros que 1000.

También es importante la velocidad con la que se necesita hacer el trabajo.

Se pueden enmarcar las cantidades en:

- 0 a 2.000 paginas
- 2.000 a 20.000
- Mas de 20.000

12.3 TIPO DE ENCUADERNACIÓN

No es lo mismo digitalizar hojas sueltas o libros que pueden ser desarmados que libros de tapa dura que no pueden abrirse completamente

Los tipos de encuadernación se pueden reducir a:

- Hojas Sueltas
- Tapa blanda
- Tapa Dura
- Incunable

12.4 MANO DE OBRA

La cantidad y capacidad de la mano de obra es un punto clave en un proyecto.

Se la puede clasificar respecto a su capacitación en:

- Especializada:
  Personal con formación en bibliotecología.
- Idónea
  Personal con manejo de software de edición de imágenes y editores de texto.
- No calificada
  Personal sin experiencia ni capacitación previa.

12.5 ACCESIBILIDAD

La manera en que se accede a los datos es importante. Se debe decidir si es importante el acceso universal a la información, por ejemplo a través de la Web, o si se prefiere una aplicación más específica.

- Web – Browser Standard HTML
  - Internet
  - Intranet
- Software especializado – Acrobat Reader
- Software propietario
- Dispositivos Móviles
- Monolítico

12.6 BÚSQUEDA DE LA INFORMACIÓN

La manera en la que es necesaria la búsqueda de la información define también el tipo de tecnología a usar.

Las opciones son:

- Búsqueda en título
- Búsqueda en Metadatos
- Búsqueda en texto

12.7 PRESUPUESTO DISPONIBLE

- Menos de 1.000
- Entre 1.000 y 10.000
- Más de 10.000

12.8 ESTÁNDARES DE FACTO

Antes de tomar la decisión de los formatos a utilizar es necesario tener en cuenta que, a pesar de no ser los más adecuados en cuanto a las características, puede ser conveniente adherir a estándares que sean los más usados, para obtener el mayor soporte posible.

13. APRENDIENDO DE OTROS PROYECTOS

A continuación se presentan ejemplos de proyectos de digitalización de documentos de distinta índole.

13.1 PROJECT GUTENBERG

El proyecto Gutenberg es la primer y más grande colección de libros electrónicos (e-books) gratuitos. Lo comenzó en 1971 Michael Hart en la Universidad de Illinois.

La filosofía detrás del proyecto es que los libros deben ser accesibles por la mayor cantidad de computadoras posibles. Para esto se eligió el formato de Texto Plano que puede ser leído por más del 99% del hardware y software disponible.

En este caso la prioridad esta puesta sobre el contenido y no sobre el formato. Por eso se decidió que los archivos sean en formato de texto plano, sin imágenes ni formato.

El trabajo de digitalización, en vez de estar centralizado, es distribuido entre los participantes (miles de participantes) del proyecto, dejando la elección del programa y equipamiento a utilizar del lado del colaborador.

Como se verificó que la corrección del OCR es la etapa del proceso de digitalización que mas esfuerzo requiere, se lanzó un proyecto paralelo llamado Distributed Proofreaders, que se puede traducir como Correctores Distribuidos, y se encarga de corregir los textos del proyecto para asegurar la calidad de su publicación.

13.2 PROYECTO CRECER

El proyecto Crecer tiene como objetivo digitalizar obras en castellano de autores clásicos, con prioridad en los argentinos, luego en latinoamericanos y continuando por los españoles.

El proyecto data del año 1999 y esta dentro de las bibliotecas rurales argentinas.

En la página de Bibliotecas Rurales Argentinas se describe al proyecto de esta manera:

“El Proyecto Crecer está preparado para la población hispano parlante, es decir no solamente respecto de América Latina y España, sino también favoreciendo a comunidades de Estados Unidos de Norteamérica

Los resultados que se esperan lograr son la colocación en la red de Internet, de una serie de recursos educativos y culturales, hasta ahora inexistentes, a disposición de los interesados en forma totalmente gratuita.

Los alumnos que puedan acceder a esta metodología podrán tener las herramientas necesarias para estudiar, pudiendo imprimir y reproducir por cualquier medio temas de estudio, con la tranquilidad que los derechos de autor han sido cedidos a su favor.

Bibliotecas Rurales Argentinas es una modesta asociación que cuenta con escaso apoyo financiero y propio, realizándose todas las actividades solamente con el trabajo voluntario. De esta manera es que iniciamos este nuevo sueño, Biblioteca Virtual Universal, invitando a compartirlo a quienes coincidan con nosotros en la esperanza de un mundo mejor.”

13.3 PROYECTO BIBLIOTECA DIGITAL ARGENTINA – CLARÍN

CORPUS DE LA BIBLIOTECA DIGITAL ARGENTINA

Está integrado por las obras más representativas de nuestra literatura y también por aquellas de difícil acceso. Así, novelas, ensayos, relatos, biografías, obras teatrales, crónicas y poesías estarán al alcance de diversos usuarios: lectores corrientes y lectores especializados.

La biblioteca se actualizará, con el agregado de nuevas obras, en forma permanente.

OBRAS QUE CONTIENE EL PROYECTO

Reproduce aquellas obras que pertenecen al corpus de la literatura argentina y que, por imperio de la Ley 11.723, han pasado al dominio público. El mencionado texto legal impone tal dominio, ya que las

obras se encuentran protegidas por el derecho de propiedad intelectual hasta 70 años desde de la muerte de su autor (contados desde el 1 de enero del año siguiente de su fallecimiento).

Se publica on line el texto íntegro de la obra siguiendo, para ello, rigurosos procesos de control y edición. Se respetan, dentro de lo prudente, las características de las versiones originales. Si alguna forma ortográfica arcaica o en desuso se transformara para facilitar la lectura, se aclarará oportunamente.

ORGANIZACION DE LA BIBLIOTECA

La organización aprovecha los beneficios de catalogación de datos que ofrece Internet, lo que permite acceder a contenidos previamente clasificados y ordenados. Asimismo, el sitio cuenta con enlaces a bibliotecas electrónicas, bibliotecas nacionales y otros sitios de literatura.

El formato usado es HTML e involucra ciento sesenta obras, con proyección a llegar a trescientas antes de 2006..

13.4 BIBLIOTECASVIRTUALES.COM

Proyecto creado y liderado por Nidia Cobiella con los siguientes objetivos, según aparecen en la página.

Los objetivos que motivan la realización de esta Biblioteca Virtual, son:

- Dar acceso a creaciones literarias que ya han pasado a ser del dominio público, a quienes navegan por la red, e incentivarlos a la lectura de obras inmortales.
- Constituir un aporte para entidades educativas, y estudiantes que investigan en Internet, como fuente para sus trabajos escolares.
- Formar un depósito de obras que constituyen el acervo cultural de la humanidad, y en especial del mundo latino.
- Permitir engrandecer su contenido con aportes de escritores que quieran colocar en esta Biblioteca Virtual sus escritos.
- Promover el interés por la lectura de obras de la literatura de todos los tiempos, considerando que la cultura y la personalidad de una persona es en gran parte lo que esa persona ha leído con gusto y concentración.
- Dar oportunidad al visitante de conocer y valorar escritores y obras de distintos movimientos literarios, épocas y entornos socio-político-geográficos.
- Utilizar la Internet realmente como una herramienta, un importante recurso educativo en que se encuentre material bibliográfico para aplicar a la educación y formación del ser humano.
- Crear en este sitio un lugar en que juntos, padres e hijos, puedan seleccionar una buena lectura para sus ratos de ocio.
- En conclusión, este sitio tiene como única finalidad que el visitante encuentre material interesante para su formación, para la educación de él o de sus alumnos, y para acrecentar su acervo cultural, o el de sus hijos.

Éste proyecto, comenta la autora, “lo fui realizando con digitalizaciones de manera manual, artesanal, luego con OCR mediante el scanner .Los he pasado a word para corrección y luego a las páginas usando Front Page[18].”

13.5 ESCUELA SUPERIOR DE COMERCIO CARLOS PELLEGRINI

La biblioteca de la Escuela Superior de Comercio Carlos Pellegrini digitalizó mas de cien volúmenes ya no protegidos por las leyes de Copyright, junto con obras de alumnos y docentes. Estas obras están disponibles en su página Web.

El objetivo de la biblioteca digital, según lo expresado en la página institucional es el siguiente:

“El objetivo del proyecto de la Biblioteca Digital es poner a disposición de los alumnos, docentes, no docentes de la Escuela y público en general una colección de libros en formato digital para que puedan acceder a estos a través de cualquier computadora conectada a Internet.

Estos libros digitales pueden copiarse en una PC o laptop y ser leídos sin necesidad de estar conectados.

Inicialmente se hará hincapié en acumular la mayor cantidad de obras literarias clásicas, novelas, ensayos, relatos, biografías, obras teatrales, crónicas y poesías, para que puedan ser leídas en una PC después de descargarlas de este sitio sin costo alguno.

La biblioteca se actualizará, con el agregado de nuevas obras, en forma permanente. Estos textos provendrán de ediciones propias y de otros sitios Web, autorización mediante.

18. Front Page es un editor de HTML usado para crear páginas WEB.

Todas las obras que aquí se publicarán serán de las que se denominan de dominio público o sea aquellas cuyo autor ha fallecido hace mas de 70 años (para las obras argentinas) o menos ajustándonos a las leyes de derecho de autor de cada país en particular."

Los documentos están en formato .Lit o PDF mayoritariamente, siendo necesario un programa para leerlos o un plug-in para el navegador de Internet.

13.6 UNIVERSITY OF VIRGINIA LIBRARY

La Universidad de Virginia tiene una biblioteca digital de más de 70.000 volúmenes (a marzo de 2002).

Ésta biblioteca usa una serie de estándares, algunos de facto y otros *de jure*[19], para organizar su colección.

Para el uso de los metadatos de las obras, la biblioteca comenzó usando, en 1992, SGML y luego migró a XML.

http://etext.lib.virginia.edu/standards/

13.7 HISTORIETAS - COMIC BOOKS

Existe una gran actividad de digitalización de historietas, principalmente por la poca calidad de sus hojas que hacen difícil su conservación.

No hay un sitio aglutinador de todos estos trabajos sino que se comparten con herramientas peer to peer o p2p como el Kazaa o el e-donkey, formando una biblioteca distribuida.

Las historietas no requieren de OCR sino que solo se necesita un facsímile de las páginas.

Ej.

Ilustración 13.1 - http://comiclist.dnsalias.com/scanguide/guide.html

19. Estándar de jure es un estándar que esta avalado por una organización de estándares, por ejemplo ISO. Se contrapone a un estándar de facto, que es algo ampliamente usado pero no avalado por ninguna organización de estándares.

Las revistas son escaneadas a 300 DPI y, una vez procesados los colores y alineadas las imágenes, reducidas a 150 DPI para que ocupen menos espacio. Finalmente son guardadas en formato JPG para su distribución.

14. TRES CASOS POSIBLES

Se presentan ejemplos concretos, basándonos en tres casos de entidades que requieren una digitalización de su material impreso.

14.1 ESTUDIO JURÍDICO

El estudio jurídico posee montones de copias de fallos, sentencias, expedientes que carecen de valor legal pero que es conveniente guardar respetando su formato original.

El contenido no puede ser editable y es posible utilizar un facsímile o imagen como archivo.

- Solo texto
- El volumen de contenidos es de 40.000 hojas.
- El tipo de encuadernación es: Hojas sueltas
- Presupuesto mayor a $ 10.000

14.2 PYME QUE DIGITALIZA DOCUMENTACION

El problema de esta empresa es que los manuales de operación de sus productos están impresos y resulta imposible darle a cada empleado una copia de los mismos.

- Texto y gráficos simples
- El volumen es de 10.000 hojas
- Los libros son tapa blanda pero pueden cortarse
- Presupuesto de $ 5.000

14.3 BIBLIOTECA BARRIAL

La biblioteca tiene muchos manuales escolares que necesitan ser consultados simultáneamente por un gran número de escolares.

Tiene varias computadoras conectadas en red y 2 escáneres flatbed

- Texto y gráficos
- Volumen es de 5.000 hojas
- Libros encuadernados. Se pueden cortar y luego reencuadernar
- Presupuesto menor a $ 3000

15. Conclusión

Los recursos humanos, técnicos y económicos de las instituciones son diferentes entre sí; no son lo mismo una escuela rural, una biblioteca de barrio, una universidad o una empresa.

También son distintos los documentos a digitalizar: en lo físico se diferencian por la encuadernación o el valor histórico de los ejemplares; en el contenido se diferencian en si contienen texto o imágenes, o en la disposición de los mismos.

El final del proceso, es decir el archivo resultante, puede tener distintos formatos y, según las necesidades, puede requerir distintas tecnologías para el acceso al conocimiento. Sin embargo no se debe elegir el formato de los libros digitales únicamente por la conveniencia particular. Debe existir la conciencia de que el conocimiento debe ser accesible por la mayor cantidad de gente posible.

El motivo de este trabajo es el de dar una solución al problema de cómo implementar una digitalización. La solución propuesta es un árbol de decisión que, dadas las restricciones, nos permita elegir una opción adecuada. Y en todo caso las opciones propuestas son elegidas teniendo en cuenta la posible colaboración entre proyectos, la facilidad de distribución del conocimiento y su perdurabilidad ante posibles nuevas tecnologías.

Digitalizar libros puede ser o no ser una inversión económicamente rentable, pero siempre es un aporte social importante y es imprescindible que el esfuerzo no sea en vano.

Ojala este trabajo cumpla su cometido.

Anexo A: Glosario

A/D Converter

EL A/D Converter, o conversor Analógico Digital, es un componente que convierte las señales analógicas generadas por el sensor CCD o CIS en digitales.

ASCII

American Standard Code for Information Interchange; Código Estadounidense Estándar para el Intercambio de Información. Creado aproximadamente en 1963 por el Comité Estadounidense de Estándares (ASA) como una refundición o evolución de los conjuntos de códigos utilizados entonces en telegrafía.

Define 128 códigos posibles (7 bits de información por código), aunque utiliza menos de la mitad, para caracteres de control, alfabéticos (no incluye minúsculas), numéricos y signos de puntuación. Su principal ventaja, aparte de constituir un estándar, consiste en la ordenación alfabética de los códigos.

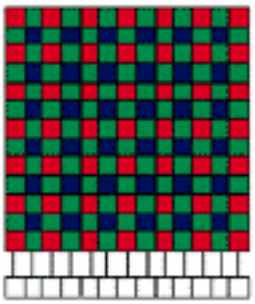
Más tarde, en 1967, se incluyen las minúsculas y se redefinen algunos códigos de control para formar el conocido US-ASCII.

Normalmente el código ASCII se extiende a 8 bits (1 byte) añadiendo un bit de control, llamado bit de paridad.

A menudo se llama incorrectamente ASCII a otros códigos de caracteres de 8 bits, como el estándar ISO-8859-1 que es una extensión que utiliza 8 bits para proporcionar caracteres adicionales usados en idiomas distintos al inglés, como el nuestro.

CCD

CCD es el acrónimo de Charged Coupled Device. Es el componente clave dentro de un escáner. Esta formado por un conjunto de celdas fotosensibles que reaccionan a la luz según sus propiedades de intensidad y color. El CCD captura la luz a través de un lente óptico para luego transformarla en señales analógicas que serán procesadas por un conversor Analógico/Digital – A/D.



CCITT Group III

Formato de compresión que promedia un índice de reducción de alrededor de 5:1. La implementación mas común de el CCITT grupo III es la compresión usada en los FAX.

CCITT Group IV

Formato de compresión que promedia una reducción de 25:1.

CDVivo

LiveCD o CDvivo es una característica para permitir ejecutar un sistema operativo desde un medio de almacenamiento normalmente CD-ROM o disquete de forma temporal a modo de demostración. Generalmente para permitirlo se descomprime una parte en la memoria RAM del ordenador, para usar esta memoria como disco duro virtual, sin necesidad de una instalación.

Uno de los mayores incovenientes de este sistema es el requerimiento de una cantidad generosa de RAM, una parte para su uso como RAM habitual y otra para funcionar como el disco duro virtual del sistema. Se le pueden pasar en el arranque distintos parámetros para adaptar el sistema al ordenador como la resolución de pantalla o activar/desactivar la búsqueda automática de determinado hardware.

La mayoría usan un sistema operativo basado en el kernel Linux, pero también se usan otros sistemas como FreeBSD o incluso Microsoft Windows (sin embargo, distribuir un LiveCD de este último es ilegal). El primer CD-ROM de esta forma fue DemoLinux en el 2000.

El auge de esta peculiaridad empezo alrededor del año 2003 con la distribución alemana de Knoppix, basada en Debian. Una de sus mejoras de este método fue la compresión cloop, esto permitió sobrepasar los 650-700 MB del CD (se usaba el driver loop) y lograr introducir hasta 2 GB. (explicación tomada de http://es.wikipedia.org/wiki/CDVivo)

CF - Carriage return.

Es uno de los caracteres de control en el código ASCII que indica a la impresora que el cursor debe posicionarse en la primera posición de la línea. Suele usarse con la sentencia LF (line feed) para moverse a la línea inferior

CIS

CIS es el acrónimo de Contact Image Sensor. El CIS tiene la misma funcion que el CCD pero es menor en tamaño y no necesita un lente óptico para poder funcionar.

CMYK

Acrónimo inglés de Cyan Magenta Yellow blacK, Cian, Magenta, Amarillo y Negro. Es un sistema de colores en el cual se pueden representar una gran gama poniendo diferentes valores a estos cuatro. El negro se nombra mediante K en lugar de B por que no haya confusión con Blue.

A diferencia del RGB, al que se denomina aditivo y en el que el color se genera con la emisión de luz, el CMYK es sustractivo y trabaja por absorción de luz.

CMYK es usado principalmente para las impresiones en cuatro colores a diferencia del RGB que es usado en los monitores y proyecciones.

Los valores de CMYK pueden transformarse en RGB y viceversa por medio de una operación aritmética.

DE JURE, estándar

Estándar de jure es un estándar que esta avalado por una organización de estándares, por ejemplo ISO. Se contrapone a un estándar de facto, que es algo ampliamente usado pero no avalado por ninguna organización de estándares.

DE FACTO, estándar

Ver de jure.

Píxel

Un píxel (contracción de picture element) es uno de los puntos que en conjunto forman una imagen digital. Usualmente son tan chicos y abundantes que al estar impresos o en un monitor dan la impresión de continuidad que tiene una fotografía analógica.

Plug-In

Plug-in es un programa que extiende las capacidades del browser de un modo específico, dando por ejemplo la capacidad de mostrar archivos de un determinado formato (PDF con el Agrobat, audio con el RealPlayer etc).

RGB

RGB es el acrónimo inglés Red, Green, Blue (Rojo, Verde, Azul). Simboliza un sistema de colores, en el cual es posible representar los colores mediante una combinación de tres valores, uno por el rojo, otro por el verde y un último por el azul.

En modo de 24-bit a cada color le corresponde un número entre 0 y 256. Ejemplos comunes de colores formados en RGB son:

- (0, 0, 0) es negro
- (255, 255, 255) es blanco
- (255, 0, 0) es rojo
- (0, 255, 0) es verde
- (0, 0, 255) es azul
- (255, 255, 0) es amarillo
- (0, 255, 255) es cyan
- (255, 0, 255) es magenta

WiKi

El término de WikiWiki («wiki wiki» significa «rápido» en la lengua hawaiana) se utiliza en la Wikipedia y en otros muchos sitios de Internet para nombrar una colección de páginas web de hipertexto, cada una de las cuales puede ser visitada y editada por cualquier persona. Una versión web de un wiki también se llama WikiWikiWeb. Se trata de un simple juego de palabras, ya que las iniciales son «WWW» como en la World Wide Web.

La forma abreviada wiki denomina a la aplicación de informática colaborativa que permite crear colectivamente documentos web usando un simple esquema de etiquetas y marcas, sin que la revisión del contenido sea necesaria antes de su aceptación para ser publicado en el sitio web en Internet.

Dada la gran rapidez con la que se actualizan los contenidos, la palabra «wiki» adopta todo su sen-

tido. El documento de hipertexto resultante, denominado también «wiki» o «WikiWikiWeb», lo produce típicamente una comunidad de usuarios. Muchos de estos lugares son inmediatamente identificables por su particular uso de palabras en mayúsculas, o texto capitalizado; uso que consiste en poner en mayúsculas las iniciales de las palabras de una frase y eliminar los espacios entre ellas, como por ejemplo en EsteEsUnEjemplo. Esto convierte automáticamente a la frase en un enlace. Este wiki, en sus orígenes, se comportaba de esa manera, pero actualmente se respetan los espacios y sólo hace falta encerrar el título del enlace entre dos corchetes.

El objetivo de un wiki es democratizar la creación y el mantenimiento de las páginas, al eliminar el «síndrome de un único webmaster o administrador».

El gran potencial del wiki radica en que no es necesario aprender a utilizar complicadas etiquetas para escribir de forma sencilla documentos y establecer enlaces desde el sitio web.

¿Para qué es un wiki?

En principio, un wiki se usa para cualquier cosa que sus usuarios deseen. El formato se presta a la colaboración, colaboración que involucra a cualquier persona, quien puede hacer lo que quiera con las páginas. Wikipedia es un wiki que tiene como misión específica ser una enciclopedia libre y actualizada por todo aquel que lo desee. Así, todas las ediciones que se hagan en este wiki deben ir orientadas a ese objetivo. No se debe abusar de la edición.

(Explicación tomada de Wikipedia – http://es.wikipedia.com)

WYSIWYG

WYSIWYG es el acrónimo de What You See Is What You Get. (lo que ves es lo que obtenés). Se aplica a los procesadores de texto y otros editores de texto con formato (como los editores de HTML) que permiten escribir un documento viendo directamente el resultado final, frecuentemente el resultado impreso. Se dice en contraposición a otros procesadores de texto, hoy en día poco frecuentes, en los que se escribía sobre una vista codificada del formato del texto. En el caso de editores de HTML este concepto se aplica a los que permiten escribir la página sobre una vista preliminar similar a la de un procesador de textos, ocupándose en este caso el programa de generar el código fuente en HTML

Anexo B: Bibliografía

- Wikipedia – La enciclopedia libre.
  - http://es.wikipedia.org/wiki/Portada
- Palo Alto Research Center
  - http://www.parc.com/
- TEI
  - http://www.tei-c.org.uk/Lite/
- DocBook: The Definitive Guide
  - http://www.docbook.org
- MSDN
  - http://msdn.microsoft.com
- XML.COM
  - http://xml.com
- Dublín Core Metadata Initiative
  - http://dublincore.org/
- La Decisión, Daniel Avenburg/Rubén Bortman, Grupo Editorial Norma, 2004.

- Digital Library Federation
  - http://www.diglib.org/standards.htm
- Comics
  - http://comiclist.dnsalias.com/scanguide/guide.html
  - http://www.comp.nus.edu.sg/~panjieyi/ScanGuideOokla/scanning.htm
- Library preservation at Harvard
  - http://preserve.harvard.edu/resources/digital.html

- MindTools
  - http://www.mindtools.com/dectree.html

Anexo C – Carta para la preservación del patrimonio digital
UNESCO

CARTA PARA LA PRESERVACIÓN DEL PATRIMONIO DIGITAL

PREÁMBULO

La Conferencia General, considerando que la desaparición de cualquier forma de patrimonio empobrece el acervo de todas las naciones, recordando que la Constitución de la UNESCO establece que la Organización "[debe ayudar] a la conservación, al progreso y a la difusión del saber, velando por la conservación y la protección del patrimonio universal de libros, obras de arte y monumentos de interés histórico o científico", que su programa Información para Todos ofrece una plataforma para el debate y la acción sobre políticas de información y sobre la salvaguardia de los conocimientos conservados en forma documental, y que su programa "Memoria del Mundo" tiene por objeto garantizar la preservación del patrimonio documental del mundo y un acceso universal al mismo, reconociendo que esos recursos de información y expresión creativa se elaboran, distribuyen, utilizan y conservan cada vez más en forma electrónica, y que ello da lugar a un nuevo tipo de legado: el patrimonio digital, consciente de que el acceso a dicho patrimonio brindará mayores oportunidades de creación, comunicación e intercambio de conocimientos entre todos los pueblos, entendiendo que este patrimonio digital se encuentra en peligro de desaparición, y que su preservación en beneficio de las generaciones actuales y futuras es una preocupación urgente en el mundo entero, proclama los siguientes principios y aprueba la presente Carta.

EL PATRIMONIO DIGITAL COMO HERENCIA COMÚN

Artículo 1 - Alcance

El patrimonio digital consiste en recursos únicos que son fruto del saber o la expresión de los seres humanos. Comprende recursos de carácter cultural, educativo, científico o administrativo e información técnica, jurídica, médica y de otras clases, que se generan directamente en formato digital o se convierten a éste a partir de material analógico ya existente. Los productos "de origen digital" no existen en otro formato que el electrónico.

Los objetos digitales pueden ser textos, bases de datos, imágenes fijas o en movimiento, grabaciones sonoras, material gráfico, programas informáticos o páginas Web, entre otros muchos formatos posibles dentro de un vasto repertorio de diversidad creciente. A menudo son efímeros, y su conservación requiere un trabajo específico en este sentido en los procesos de producción, mantenimiento y gestión.

Muchos de esos recursos revisten valor e importancia duraderos, y constituyen por ello un patrimonio digno de protección y conservación en beneficio de las generaciones actuales y futuras.

Este legado en constante aumento puede existir en cualquier lengua, cualquier lugar del mundo y cualquier campo de la expresión o el saber humanos.

Artículo 2 - Acceso al patrimonio digital

El objetivo de la conservación del patrimonio digital es que éste sea accesible para el público. Por consiguiente, el acceso a los elementos del patrimonio digital, especialmente los de dominio público, no debería estar sujeto a requisitos poco razonables. Al mismo tiempo, debería garantizarse la protección de la información delicada o de carácter privado contra cualquier forma de intrusión.

Los Estados Miembros tal vez deseen trabajar en colaboración con las organizaciones e instituciones pertinentes para propiciar un contexto jurídico y práctico que maximice la accesibilidad del patrimonio digital. Convendría reafirmar y promover un justo equilibrio entre los derechos legítimos de los creadores y otros derechohabientes y el interés del público por tener acceso a los elementos del patrimonio digital, de conformidad con las normas y los acuerdos internacionales.

VIGILANCIA CONTRA LA PÉRDIDA DE PATRIMONIO

Artículo 3 - El peligro de pérdida

El patrimonio digital del mundo corre el peligro de perderse para la posteridad. Contribuyen a ello, entre otros factores, la rápida obsolescencia de los equipos y programas informáticos que le dan vida, las incertidumbres existentes en torno a los recursos, la responsabilidad y los métodos para su mantenimiento y conservación y la falta de legislación que ampare estos procesos.

Los cambios en las conductas han ido a la zaga del progreso tecnológico. La evolución de la tecnología digital ha sido tan rápida y onerosa que los gobiernos e instituciones no han podido elaborar estrategias de conservación oportunas y bien fundamentadas. No se ha comprendido en toda su magnitud la amenaza que pesa sobre el potencial económico, social, intelectual y cultural que encierra el patrimonio, sobre el cual se edifica el porvenir.

Artículo 4 - Necesidad de pasar a la acción

A menos que se haga frente a los peligros actuales, el patrimonio digital desaparecerá rápida e ineluctablemente. El hecho de estimular la adopción de medidas jurídicas, económicas y técnicas para salvaguardar ese patrimonio redundará en beneficio de los propios Estados Miembros. Urge emprender actividades de divulgación y promoción, alertar a los responsables de formular políticas y sensibilizar al gran público tanto sobre el potencial de los productos digitales como sobre los problemas prácticos que plantea su preservación.

Artículo 5 - Continuidad del patrimonio digital

La continuidad del patrimonio digital es fundamental. Para preservarlo se requerirán diversas medidas que incidan en todo el ciclo vital de la información digital, desde su creación hasta su utilización. La preservación a largo plazo del patrimonio digital empieza por la concepción de sistemas y procedimientos fiables que generen objetos digitales auténticos y estables.

MEDIDAS NECESARIAS

Artículo 6 - Elaborar estrategias y políticas

Es preciso elaborar estrategias y políticas encaminadas a preservar el patrimonio digital, que tengan en cuenta el grado de urgencia, las circunstancias locales, los medios disponibles y las previsiones de futuro. La colaboración de los titulares de derechos de autor y derechos conexos y otras partes interesadas a la hora de definir formatos y compatibilidades comunes, así como el aprovechamiento compartido de recursos, pueden facilitar esa labor.

Artículo 7 - Seleccionar los elementos que deben conservarse

Al igual que ocurre con el conjunto del patrimonio documental, los principios de selección pueden diferir de un país a otro, aun cuando los principales criterios para determinar los elementos digitales dignos de conservación sean su significado y valor duraderos en términos culturales, científicos, testimoniales o de otra índole. Indudablemente, se deberá dar prioridad a los productos “de origen digital”. Los procesos de selección y de eventual revisión subsiguiente han de llevarse a cabo con toda transparencia y basarse en principios, políticas, procedimientos y normas bien definidos.

Artículo 8 - Proteger el patrimonio digital

Los Estados Miembros han de disponer de mecanismos jurídicos e institucionales adecuados para garantizar la protección de su patrimonio digital.

Hacer que la legislación sobre archivos, así como el depósito legal o voluntario en bibliotecas, archivos, museos u otras instituciones públicas de conservación, se aplique al patrimonio digital, ha de ser un elemento esencial de la política nacional de preservación.

Convendría velar por el acceso a los elementos del patrimonio digital legalmente depositados, dentro de límites razonables, sin que ese se haga en perjuicio de la explotación normal de esos elementos.

Para prevenir la manipulación o modificación deliberada del patrimonio digital, es de suma importancia disponer de un marco tanto jurídico como técnico en el que se proteja la autenticidad.

Esto exige, en ambos casos, mantener los contenidos, el funcionamiento de los ficheros y la documentación en la medida necesaria para garantizar que se conserva un objeto digital auténtico.

Artículo 9 - Preservar el patrimonio cultural

Por definición, el patrimonio digital no está sujeto a límites temporales, geográficos, culturales o de formato. Aunque sea específico de una cultura, cualquier persona del mundo es un usuario en potencia. Las minorías pueden dirigirse a las mayorías y los individuos a un público de dimensión mundial.

Hay que preservar y poner a disposición de cualquier persona el patrimonio digital de todas las regiones, naciones y comunidades a fin de propiciar, con el tiempo, una representación de todos los pueblos, naciones, culturas e idiomas.

ATRIBUCIONES

Artículo 10 - Funciones y atribuciones

Los Estados Miembros tal vez deseen designar a uno o más organismos que se encarguen de coordinar la preservación del patrimonio digital y poner a su disposición los recursos necesarios. La división de tareas y atribuciones puede basarse en las funciones y competencias existentes.

Convendría adoptar medidas para:

a) instar a los fabricantes de equipos y programas informáticos, creadores, editores y productores y

distribuidores de objetos digitales, así como otros interlocutores del sector privado, a colaborar con bibliotecas nacionales, archivos y museos, y otras instituciones que se ocupen del patrimonio público, en la labor de preservación del patrimonio digital;

b) fomentar la formación y la investigación, e impulsar el intercambio de experiencia y conocimientos entre las instituciones y las asociaciones profesionales relacionadas con el tema;

c) alentar a las universidades y otras instituciones de investigación, públicas y privadas, a velar por la preservación de los datos relativos a las investigaciones.

Artículo 11 - Alianzas y cooperación

La preservación del patrimonio digital exige un esfuerzo constante por parte de gobiernos, creadores, editoriales, industriales del sector e instituciones que se ocupan del patrimonio.

Ante la actual "brecha digital" es necesario reforzar la cooperación y la solidaridad internacionales para que todos los países puedan garantizar la creación, difusión y preservación de su patrimonio digital, así como un acceso constante al mismo.

Se insta a los fabricantes, las editoriales y los medios de comunicación de masas a que promuevan y compartan sus conocimientos teóricos y técnicos.

El hecho de favorecer programas de educación y formación, acuerdos de aprovechamiento compartido de recursos y mecanismos de difusión de los resultados de investigaciones y prácticas idóneas democratizará el conocimiento de las técnicas de preservación de objetos digitales.

Artículo 12 - La función de la UNESCO

En virtud de su mandato y funciones, incumbe a la UNESCO:

a) incorporar los principios establecidos en esta Carta al funcionamiento de sus programas y promover su aplicación tanto dentro del sistema de las Naciones Unidas como por las organizaciones internacionales, gubernamentales y no gubernamentales, relacionadas con la preservación del patrimonio digital;

b) ejercer de referente y de foro en el que los Estados Miembros, las organizaciones internacionales, gubernamentales y no gubernamentales, la sociedad civil y el sector privado puedan aunar esfuerzos para definir objetivos, políticas y proyectos que favorezcan la preservación del patrimonio digital;

c) impulsar la cooperación, sensibilización y creación de capacidades y proponer directrices éticas, jurídicas y técnicas normalizadas para apoyar la preservación del patrimonio digital;

d) basándose en la experiencia que adquirirá en los seis años venideros con la aplicación de la presente Carta y las directrices, determinar si se requieren nuevos instrumentos normativos para promover y preservar el patrimonio digital.