DIEPER

LB-5632

Deliverable 15

General report on image capturing and document management

Date	30/06/99
Produced by	Marco Köttstorfer
Distribution list	All DIEPER partners
Contact person	Marco Köttstorfer

✉ Universität Linz
Abteilung „Informatik für Blinde„
Altenbergerstr. 69
A-4040 Linz
✆ + 43 699 100 53803
🖷 + 43 732 400 53803
✍ marco@mvblind.uni-linz.ac.at

Bruno Sperl
✉ Karl-Franzens-Universität
Hauptbibliothek
Universitätsplatz 3
A-8010 Graz
✆ + 43 316 380 1415
🖷 + 43 316 380 691415
✍ bruno.sperl@kfunigraz.ac.at

Document history

Versions			
Version	**Date**	**Author**	**Comments**
0.1	20/12/98	R. Ecker	TOC for D15
1	30/06/99	M. Köttstorfer	Final Version of D15

Updates		
Chapter	**Description of modifications**	**Version**

Table of contents

1. Introduction

This document describes the current status of image capturing and document management methods, especially with respect of library documents.

Some years ago several libraries have started to retrodigitize printed materials, as e.g. books or periodicals, and to distribute these digital documents via Internet or other networks to users over the world. We are now at the beginning of a development which will, as we hope, make all important information make available immediately from anywhere and at any time.

This new kind of information access meets already some enthusiasm from their users. This will work like a catalyst for starting additional projects. It is expected, that the information behaviour will be influenced considerably by the direct access to digitised documents.

Libraries – in our tradition one of the significant groups of conventional information providers – will identify and use this chance to overtake also a leading role in the digital information society.

The goal of the DIEPER project is to enhance these developments with respect to the digitisation, indexing and presentation of scientific periodicals.

2. Document management

Electronic document management allows us to manage all our documents - fast and efficiently. All your documents are available within seconds on your PC, and at a much lower cost level compared to a paper-based system to boot!

An electronic document management system is based on these four basic procedures:
- Capturing
- Indexing
- Storing
- Retrieving

Capturing
When capturing a document, it becomes available to the document management system.

- *Digitising of a paper document using a scanner*
  Paper documents such as books or periodicals are converted into a digital format with the help of a scanner. This underlying technology is similar to the methods used by a fax machine, but the scanning process is much faster. The size of the documents can vary considerably as scanners can handle documents from quite any size.

- *Importing computer-generated documents*
  PC-generated documents from any Windows application can be imported directly as well as a print file from host, UNIX or DOS systems.

Indexing
After capturing, the documents are available in the document management system. For later retrieval, the documents are now assigned with keywords such as authors, date, publisher name and document type. The type and number of keywords can be defined very individual. Which type and which number of keywords are applicable is dependent on the amount of documents which have to be handled and on the document type.

- Manual indexing and select lists
  Manual entry of keywords with a keyboard or choosing keywords from a select list are options which are always available. If possible, documents can be indexed fully or semi-automatically with OCR (Optical Character Recognition).

- Optical Character Recognition (OCR)
  On documents which have a standardised format, the index criteria is read fully automatically from each page and stored.
  For all other documents, the indexing time can be reduced significantly by double-clicking with the mouse on the appropriate keywords (Point-and-Shoot).

- Automatic import
  Keywords can be transferred fully automatically from application programs.

Storing

After keywords have been assigned to a document, the documents are stored in a digital file cabinet. The keywords are automatically added to the database.
All stored documents are available immediately over the network.
On demand, older and less requested documents can be moved to slower, but more affordable storage media (e.g. CD-ROM's). This process is called migration

Retrieving

In the retrieval process, the strengths of an electronic document management system can be experienced in full force. You can search for all keywords which you have entered during the storage process and for the storage date and name of the person who stored the document.

The retrieval of documents can be performed in seconds. Retrieved documents can

- be viewed on the screen - even by multiple colleagues at once
- be printed on paper
- be processed with other application programs
- be mailed by fax
- etc.

3. Capturing

To make a printed document available via the Internet it has to be converted into an electronic format.

There are 2 different ways how the capturing can be done:
- digitisation of printed documents by using a scanner
- importing of electronic documents
  all PC-documents can be used

Only the first point is relevant for the project, because there are no electronic documents available that can be used. We only deal with printed documents.

3.1. Scanning

One of the first difficult issues that must be addressed in any digital conversion project concerns the selection of appropriate formats and technologies for storage, display and distribution of the material.

Another difficult question is what file format (images, PDF, SGML, HTML, etc.) should be used to deliver the content.

Another question is whether to store and deliver the materials as images or as text. Given the technology available to web browsers, the most accurate way to replicate completely the originally published material, which is full of special characters, foreign languages, mathematical symbols, charts and pictures, is with scanned images.

In addition, by the use of Optical Character Recognition software a corresponding text file can be built that would allow the user to search the full-text of the journals in the database. But these (uncorrected) OCR-text files should not be made available to users.

General Recommendation:

It is recommended to use the images for display and for printing, and the text files for searching.

Coded information vs. non-coded information

Coded information	**Non-coded information**
Text files. Can be directly retrieved. Text processing editor. OCR/ICR. Desktop publishing. Electronic publishing.	Image files. Can not be retrieved directly. Image scanning. Pixel editor.

Scanning can be divided in 4 main categories:

- Line Art Scanning (1bit)
- Halftone Scanning (1bit)
- Greyscale Scanning (8bit)
- Colour Scanning (24bit)

In each category there are problems regarding how to reach the best scan quality.
Refer to [2] for further details.

Basic scanning parameters

- Kind of the original document (printed text on paper, printed image, photograph, colour, microfilm, microfiche, ...)
- Size of the original document (micro form, <A 4, >A 4 <A3, <A 3, ..., >A0)
- Scanning resolution (100 dpi, 300 dpi, 400 dpi, 600 dpi, ...)
- Image depth (pixel information: 1 bit, 8 bit, 12 bit, 3 x 12 bit, ...)
- Intended exploitation of the digital materials
- File size of the digital materials

Criteria for the definition of scanning parameters

The definition of scanning parameters depends on the

- kind of the printed materials
- kind of intended use
- kind of intended access

Kind of printed materials

Paper based materials

- Bounded volumes
- Single sheets of paper
- Maps
- Library catalogue cards
- etc.

- One side – double sided

- Usual book format (~ A 4, A 5 ..)
- Small size
- Large size (A 0 or larger)

- Text
- Graphics
- Halftone
- Colour

Microfilm based materials

- Microfilm
- Microfiche
- Slides
- Professional reprofilm

Other materials

- 3-D objects
- etc.

Kind of intended use and further processing of the digitised materials

Presentation on the screen

Reproduction by a printer

- Local print of a small number of document pages
- Reprint of the complete document
- Reprint in professional quality
- Reprint of coloured posters in an optimum true colour quality
- etc.

Further processing of documents

- Automatically OCR-conversion to full text
- Automatically vectorisation of graphical information
- Automatically analysis of the document type
- Production of a CD-ROM
- etc.

Kind of intended access to the digitised materials

- Via Internet/Intranet
- Local access within the premises of the library

Categories of scanners

- Flat bed scanners
- Flat bed scanners with automatic feeders
- Camera scanners
- Specialised book scanners
- Microfilm scanners
- Other specialised Scanners (x-ray, 3 D objects, ...)

- Black/white scanners
- Greyscale scanners
- Colour scanners

- Digital resolution (dpi)

Image processing

- Separation of double pages to single pages
- Clipping to remove the border or any black back ground
- Orientation of the image to vertical
- Purification from dirt (e.g. from marks caused by mould)
- Optimisation of contrast
- Scaling to original size
- etc.

3.2. Storage of the image files

Here the question is, what do I want to do with the scanned Images?
Do I use them as

- a digital master file
- an archive file
- for screen presentation
- for local print
- a download format
- Other

In general we have to divide the above into two groups: Digital master file and archive file into the first group and the rest into the second group.
The first group is for a one to one saving of the documents, for this reason the file format for storing these documents has to be lossless.
The second group has to bear in mind the technical realisation. Slow download rates and small amounts of disk space forces us to compress the data to reduce the size of the image files. Because of that limitation, it may not be possible to offer very high resolution colour or greyscale images.
And in general high compression means loosing Information, due to lossy compression.

Here are the most important file formats:

with lossless compression

- TIFF
- GIF
- PNG

with lossy compression

- JPEG
- FlashPix
- Wavelet
- Fractal
- STiNG

3.2.1 Tag(ged) Image File Format TIFF

TIFF is a widely supported format within the libraries community. The latest TIFF version is 6.0.

TIFF limitations: There are no provisions in TIFF for storing vector graphics and text annotation (although such items could be easily constructed using TIFF extensions). TIFF uses 4-byte integer file offsets to store image data, with the consequence that a TIFF file cannot have more than 4 Gigabytes of compressed raster data. This is not a big deal for DIEPER since this limit is far from being reached within a single document. It is considered that an average document is a 10-page document with each page having 100 KB compressed size. This makes the average size of a requested article roughly 1 MB.

TIFF strengths: TIFF is primarily designed for raster data interchange. Its main strengths are a highly flexible and platform-independent format which is supported by numerous image processing applications. Supported compression algorithms are: raw uncompressed, PackBits, LZW (Lempel-Ziv-Welch), CCITT Group 3 & 4 and JPEG compression.

Regarding time transfer for an average document: Suppose that we have an end-user connected through a dial-up connection (28,800 BPS). A 1 MB document requires then roughly 5 to 10 minutes to download. This seems to be accepted by end-users compared to the classical postal delivery.

Colour depth 24-bit for full colour and 8-bit for grayscale.
Typical compression rate 2:1

The 4 major TIFF formats and compressions are:
1.Bitmap:

- uncompressed
- word-boundaries uncompressed
- Huffman CCITT-1D compression
- Packbits compression
- LZW compression (LZW Compression under the licence of the Unisys Corporation, U.S. Patent Nr. 4,558,302)

2.Grayscale:

- 4-bit uncompressed
- 8-bit uncompressed
- 4-bit LZW compressed
- 8-bit LZW compressed
- 4-bit Thunder Scan compression

3.Paletten Colours:

- 4-bit uncompressed
- 8-bit uncompressed
- 4-bit LZW compressed
- 8-bit LZW compressed

4.RGB Colours:
- 24-bit uncompressed
- 24-bit LZW compressed
- 6-5-5 16-bit uncompressed

and as further formats:

CMYK Colours:
- 8-8-8-8 CMYK uncompressed
- 8-8-8-8 CMYK LZW compressed

Advantages :
- It is possible to store additional information like metadata in the TIFF Header, to identify the image, the descent, publisher ... The TIFF Header will be explained later in the deliverable.

Disadvantages:
- Not a standard on the World Wide Web. Plug-ins for Web-browsers are desired to display the images.
- The LZW Compression Algorithm is licensed by the Unisys Corporation. That means that professional users have to pay a licence-fee if they want to use this algorithm.

3.2.2 GIF

GIF ™ is CompuServe's standard for defining generalised colour raster images. This *Graphics Interchange Format* ™ allows high-quality, high-resolution graphics to be displayed on a variety of graphics hardware and is intended as an exchange and display mechanism for graphics images.
The Graphics Interchange Format is defined in terms of blocks and sub-blocks which contain relevant parameters and data used in the reproduction of a graphic. A GIF Data Stream is a sequence of protocol blocks and sub-blocks representing a collection of graphics. In general, the graphics in a Data Stream are assumed to be related to some degree, and to share some control information; it is recommended that encoders attempt to group together related graphics in order to minimise hardware changes during processing and to minimise control information overhead. For the same reason, unrelated graphics or graphics which require resetting hardware parameters should be encoded separately to the extent possible.

GIF can handle colour depths from 1 to 8 bit – a maximum of 256 colours – and uses the LZW- compression-algorithm. Several images can be stored in a GIF File and also Slideshows can be presented.

Advantages:
- standard in the World Wide Web

Disadvantages:
- can only display 256 colours
- a licence-fee has to bee paid because of the usage of the LZW-Algorithm

3.2.3 PNG

The *Portable Network Graphics* (PNG) format was designed to replace the older and simpler GIF format and, to some extent, the much more complex TIFF format. For the Web, PNG

really has three main advantages over GIF: alpha channels (variable transparency), gamma correction (cross-platform control of image brightness), and two-dimensional interlacing (a method of progressive display). PNG also compresses better than GIF in almost every case, but the difference is generally only around 5% to 25%. One GIF feature that PNG does not try to reproduce is multiple-image support, especially animations; PNG was and is intended to be a single-image format only.
PNG's compression is fully lossless - and since it supports up to 48-bit truecolor or 16-bit grayscale--saving, restoring and re-saving an image will not degrade its quality, unlike standard JPEG (even at its highest quality settings). And unlike TIFF, the PNG specification leaves no room for implementors to pick and choose what features they'll support; the result is that a PNG image saved in one application is readable in any other PNG-supporting app.
The *Portable Network Graphics* was developed to replace the older GIF format and to extend the TIFF format. Compression is up to 30% better than the one used in a GIF file.
It can also handle 8-bit palettes and was developed by the IETF.

Advantages:
- lossless compression
- no plug-in for Web-browsers necessary

Disadvantages:
- if a PNG plug-in is used some problems in displaying the image can appear
- unclear if the format will establish as a standard on the WWW

3.2.4 JPEG

JPEG stands for *Joint Photographic Experts Group*, the original name of the committee that wrote the standard.
JPEG is designed for compressing either full-colour or gray-scale images of natural, real-world scenes. It works well on photographs, naturalistic artwork, and similar material; not so well on lettering, simple cartoons, or line drawings.
JPEG is designed to exploit known limitations of the human eye, notably the fact that small colour changes are perceived less accurately than small changes in brightness. Thus, JPEG is intended for compressing images that will be looked at by humans.
JPEG can typically achieve 10:1 to 20:1 compression without visible loss, 30:1 to 50:1 compression is possible with small to moderate defects, while for very-low-quality purposes such as previews or archive indexes, 100:1 compression is quite feasible. [16]
Colour depth 24-bit.

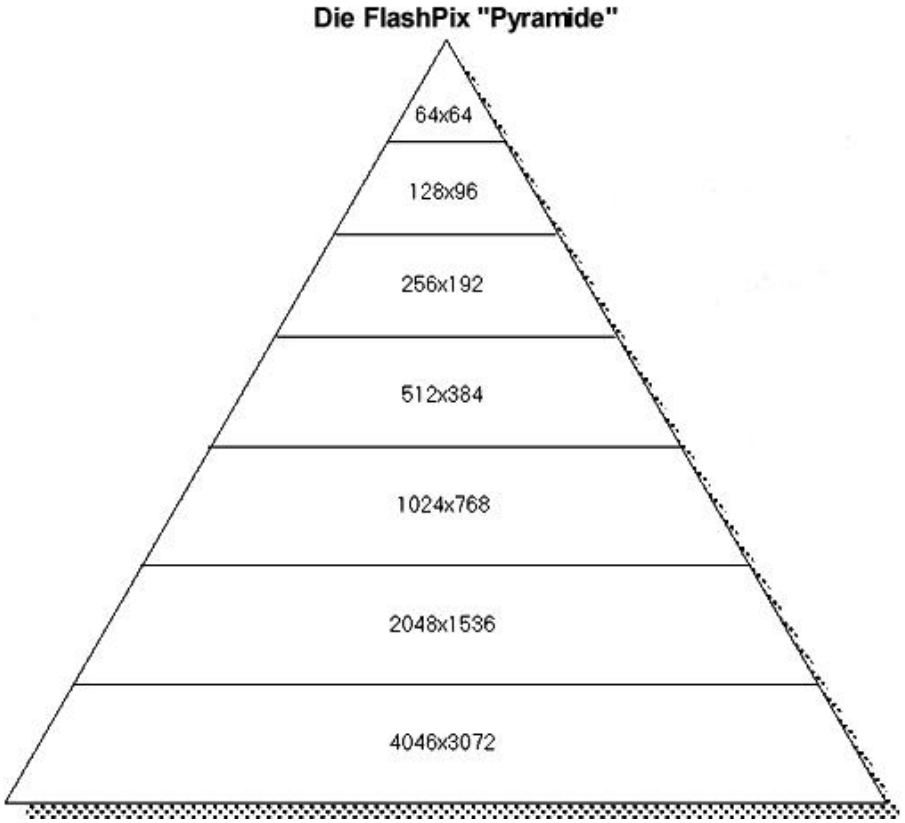
Advantages:
- compression can be adjusted
- files can be very high reduced

Disadvantages:
- high compression will produce an image with several plains on the image. This can be disturbing.
- lossy compression

3.2.5 FlashPix

The FlashPix format can store the original picture and some copies of this picture with a lower resolution. There is no limit to the size, the colour depth and the resolution of an image. There is also the possibility to display several variants of the same image without storing it

twice. Therefore filters are written in a script language and stored in the same file. When it comes to display the images, these “corrections” will be calculated in real-time to display the changed picture. Metadata can also be stored in the image file.



Picture 1: The FlashPix „pyramid„ [4]

As you can see on the picture, the picture will be „broken down“ to smaller resolutions.

Advantages:
- only the desired resolution will be transmitted over the net and helps saving bandwidth.
- metadata can be stored in the file
- only 1 file necessary
- can be produced from JPEG, GIF, TIFF and also be taken direct from a digital camera or a scanner

Disadvantages:
- the World Wide Web Consortium (W3C) will not support this image format as a standard.
- no browser support at this time

3.2.6 Wavelet Compression

Wavelet compression is a method of mathematical modelling of images, which breaks the image down into small waves that represent the frequency analysis of a function. The shapes and patterns in an image are identified, and then described using mathematical functions or formulas. The function that models or describes the image is contained within the compression and decompression software. The image file contains only the coefficients or numbers used by the function, and compression is achieved by averaging the values of these coefficients so that an image is represented by fewer numbers.

An advantage of wavelet compression is that image processing can be incorporated into the wavelet transformation, including sharpening, contrast enhancement, and noise reduction for images. Also, images can be enlarged or reduced via embedded interpolation, using common interpolation algorithms, such as bicubic, bilinear, or nearest neighbour, as found in Adobe Photoshop and in other pixel-based image editing software.

Maximum compression rate 150:1
Typical compression rate 15:1 up to 100:1
time for decompression comparable to the one JPEG needs

Advantages:
- high compression
- time for decompression comparable to the one JPEG needs
- one effect of the compression is, that the image will be blurred. This effect is enjoyable for the human eye

Disadvantages:
- compression rate can not be adjusted
- long compression times
- not a standard
- plug-in for browsers necessary
- proprietary format

3.2.7 Fractal Compression

Mathematically, a fractal describes a structure that has many repeated forms regardless of scale. Real world images have properties that allow them to be described using fractals, such as repeated shapes and patterns. Fractal compression works by using a variety of methods to identify features within an image and then breaking down the image into a mathematically modelled series of repeating shapes and patterns.
Images can be magnified or reduced, because the compression process allows the modelled images to be resolution-independent. When a fractally-encoded image is converted to a pixel image, it can be enlarged or reduced to any desired size with minimal loss of image quality.
Fractal compression can take a very long time to compress images - significantly longer than wavelet compression in test trials, but the decompression is relatively quick.

Maximum compression rate 250:1
Typical compression rate 20:1 to 100:1

Advantages:
- high compression
- time for decompression comparable to the one JPEG needs
- resolution independent
- one effect of the compression is, that the image will be blurred. This effect is enjoyable for the human eye

Disadvantages:
- compression rate can not be adjusted
- long compression times
- not a standard
- plug-in for browsers necessary
- proprietary format

Comparison of JPEG, Wavelet, and Fractal Compression [1]

	JPEG	Wavelet	Fractal
Typical compression	10:1 to 20:1	15:1 to 100:1	20:1 to 100:1
Compression time*	8 seconds	45 seconds	13 minutes
Decompression time*	6 seconds	20 seconds	10 seconds
Image quality	excellent to poor	excellent to poor	excellent to poor
Resolution features	none	interpolation	scalable
Zoom	no in general; yes with FlashPix	yes	yes
Dedicated browser or plug-in required	no	yes	yes

* Times measured on a 23MB RGB image scanned from a 35mm slide. All files were compressed using Adobe Photoshop 4.0 running on an Apple Power Macintosh 9500 with 244Mhz G3 upgrade. JPEG compression (at 60:1) was native Photoshop JPEG. Wavelet compression plug-in used was AccuPress from Aware Inc. (image compressed at 120:1). Fractal compression plug-in used was Genuine Fractals from Altamira Group, Inc. (image compressed 120:1).

In conclusion, wavelet and fractal compressions offer the advantages of higher compression ratios, comparable to higher levels of image quality for the compressed files, and similar decompression times compared to JPEG or LZW compression. Fractal compression has the advantage of being resolution-independent, which allows end-users to open a file at any desired resolution or size (only up to a point). Wavelet compression can incorporate traditional interpolation methods to enlarge or reduce images on-the-fly and, therefore, can function very similarly to fractally-encoded images. The disadvantages for both wavelet and fractal compression include long compression times, non-standardised and proprietary compression algorithms, and the need for a dedicated viewer or plug-in to use the files.

3.2.8 STiNG

Iterated Systems' STiNG technology offers the advantages of both types of compression combined with lossless encoding, all of which should provide high compression ratios, good image quality, and scalability. I suspect that the advantages of both wavelet and fractal compression will end up outweighing the disadvantages. [1]
This new technology was not available when the deliverable was written.

3.3. Digital master file

The image file which results from the scanning will be the so called Digital master file of the highest quality. This is the archive version which should be stored in a standardised format under lossless compression on a long life data carrier. At present, for the permanent storage of scanned text documents the TIFF format should be preferred in combination with the CCITT G4 compression.

Alternative formats may be PNG and (only for grey scale and colour) the GIF format.

3.4. Application file formats

From the digital master file a copy will be stored in the archive system on an magnetic disc array (e.g. for scanned text documents: TIFF, 300 dpi).

In addition application formats will be prepared for the screen presentation (e.g. GIF or JPEG, 75 – 100 dpi), for the download and for local printing (e.g. Postscript or PDF, 300 dpi). These application formats may also be stored in the archive system or can be prepared on-the-fly.

3.5. Recommendation:

It is recommended to apply the parameters and formats shown in the following tables.

Text, Line-graphics

Scanning	(300)/400/600 dpi		1 Bit
Storage		TIFF/CCITT G4	1 Bit
Viewing	70-120 dpi	GIF	1-4 Bit
Gallery/Thumbnails	15 dpi	GIF	1 Bit
Download	300/400/600 dpi	PDF	1 Bit

If an OCR conversion is intended, scanning should be done with a resolution at least of 400 dpi. If it is further planned to provide a print-on-demand service, scanning should be done with a resolution of 600 dpi.

Grey-scale graphics, Photographs

Scanning	300dpi		8 Bit
Storage		TIFF uncompressed	8 Bit
Viewing	512x768 to 1024x1536	JPEG	4 Bit
Gallery/Thumbnails	~ 100x150	JPEG	4 Bit
Download	2048x3072	JPEG	8 Bit

Manuscripts

Scanning	300dpi		8 Bit
Storage		TIFF uncompressed	8 Bit
Viewing	512x768 to 1024x1536	JPEG	1-4 Bit
Gallery/Thumbnails	~ 100x150	JPEG	< 8 Bit
Download	2048x3072	JPEG	8 Bit

Colour graphics

Scanning	200-300dpi		3x8 Bit
Storage		TIFF uncompressed	3x8 Bit
Viewing	512x768 to 1024x1536	JPEG	3x8 Bit
Gallery/Thumbnails	~ 100x150	JPEG	8 Bit
Download	2048x3072	JPEG	3x8 Bit

2-D Representation of 3-D Objects

Scanning	200-300dpi		3x8 Bit
Storage		TIFF uncompressed	3x8 Bit
Viewing	512x768 to 1024x1536	JPEG	3x8 Bit
Gallery/Thumbnails	~ 100x150	JPEG	8 Bit
Download	2048x3072	JPEG	3x8 Bit

Acceptable compression for JPEG-files

Grey scale: maximum 10:1
Colour: maximum 15:1

For an exact calculation which resolution should be used for scanning or which resolution is useful, the formula of Mr. Jeff Bones in his FAQ „The Scanning FAQ“ [2] can be used.

4. Indexing - Metadata

The Internet is growing from minute to minute and is growing even faster every minute. So it gets quite hard to find the information wanted and needed. The big search-engines like Lycos, Yahoo, Altavista and so on can only handle a small amount of the data on the net.(about 35%)
But the amount of data available in electronic form on the net is still growing. The chance for outsiders to find data is rather difficult not to say even impossible. Therefore it is necessary to combine normal data with metadata to make indexing easier.
TEI header and MARC are two standards that would fulfil these needs. But the creation and maintenance of these data are time-consuming.

Metadata, what's that?
Metadata are data about data. They can consist of information about the author, links and so on. One form of metadata would be the index catalogue on these little cards in libraries.

Categories of indexing

Bibliographic indexing
Document identifier
Document structure (SGML or XML representation of the document structure)
Full text (complete full text or in part)

Bibliographic indexing
- Bibliographic data
- Catalogue data sets
- Dublin Core data sets
- Storage of bibliographic data in the TIFF-Header
- Storage of bibliographic data in the TEI-Header
- etc.

Document identifier
- DOI
- URL
- PURL
- etc.

Document structure
- SGML/XML representation
- Ebind format
- Hyperlinks text ↔ Image page

SGML is quite used for describing the structure of catalogue records. SGML is an international standard used for the formal definition of electronic text. SGML is thus a structure driven meta language. HTML for instance is an application of SGML. The structure of an SGML set of documents is described in a single definition document referred to as the DTD, the Document Type Definition. HTML corresponds to a specific DTD as well as

Netscape browser (HTML viewer) which has its own DTD called Mozilla, a superset of HTML 2.0.

Full text (complete full text or in part)
- Complete documents
- Special parts of the document
- Summary, abstract
- Tables of contents
- Indexes
- Key words

Full text formats
- HTML
- ASCII
- .DOC
- .XLS
- TEX
- etc.

Methods of full text + meta data capturing
- Manually input
- OCR/ICR
- Download of catalogue data
- etc.

4.1. Bibliographic indexing

4.1.1 TIFF Header

TIFF allows the storage of additional and private information in its header. The size of these data, called metadata, are only limited to the max size of the TIFF file, 2GB. Therefore the normal "standard" tags which are already provided by the TIFF header and to provide extended metadata the so called "private" tags can be used. Some programs seems to have their problems when private tags are used. So in the following only the standard tags will be discussed in detail.

For books and periodicals we can use the following tags for storing metadata:

Tag	Name	Type	Usage
269	DocumentName	ASCII	for saving the name of the document or the ISSN Number
270	ImageDescription	ASCII	for saving metadata
285	PageName	ASCII	physical page number (the same as the one in the filename)
305	Software	ASCII	software used for scanning
315	Artist	ASCII	the scanning department (University, firm, ...)

33432	Copyright	ASCII	Copyright, this tag can be ignored if the copyright was stored within the ImageDescription

4.1.2 TEI Header

For a detailed description on TEI please refer to chapter 5.9 Ebind and TEI.
Like TIFF also TEI can store a lot of metadata in its header.
The TEI Header consists of 4 parts:

- File Description
- Encoding Description
- Text Profile Description
- Revision Description

The *File Description* is used for a full bibliographic description like the title, author, publisher, creator of the electronic version, size and information on the electronic version and details about the printed source from which the electronic one was created.

The *Encoding Description* is used for a detailed description about how the text was treated when converting to the electronic version and so on.

The *Text Profile Description* is used for a detailed description of non-bibliographic aspects of the text like the language used or the historical situation in which the book was written

The *Revision History Description* is used to store a description of all changes made while developing the text in electronic form.

4.1.3 MARC

MARC (*Machine- Readable Cataloguing*) is a set of records to describe bibliographic data. There are several standards on national level e.g. USMARC, Can/MARC, InterMARC, UKMARC, CCF and so on. MARC is very rich on describing elements, not to say it is to rich. MARC records are, because of their extent very difficult to maintain.
Conversion from MARC to SGML/XML and vice versa is planned, and it is worked on that topic.
As Example you can look at the MARC DTD.

4.1.4 MARC DTD

The Term „*MARC DTD*“ (*Machine Readable Cataloguing Document Type Definition*) is used for the Implementation on the Standard Generalised Markup Language (SGML).
MARC DTD handles MARC Records like individual types of Documents.

Please refer to http://www.oasis-open.org/cover/marcdtdback.html

4.1.5 Dublin Core

Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of Web resources, it has attracted the

attention of formal resource description communities such as museums, libraries, government agencies, and commercial organisations.

The characteristics of the Dublin Core that distinguish it as a prominent candidate for description of electronic resources fall into several categories:

- **Simplicity**
  The Dublin Core is intended to be usable by non-cataloguers as well as resource description specialists. Most of the elements have a commonly understood semantics of roughly the complexity of a library catalogue card.

- **Semantic Interoperability**
  In the Internet Commons, disparate description models interfere with the ability to search across discipline boundaries. Promoting a commonly understood set of descriptors that helps to unify other data content standards increases the possibility of semantic interoperability across disciplines.

- **International Consensus**
  Recognition of the international scope of resource discovery on the Web is critical to the development of effective discovery infrastructure. The Dublin Core benefits from active participation and promotion in some 20 countries in North America, Europe, Australia, and Asia.

- **Extensibility**
  The Dublin Core provides an economical alternative to more elaborate description models such as the full MARC cataloguing of the library world. Additionally, it includes sufficient flexibility and extensibility to encode the structure and more elaborate semantics inherent in richer description standards

- **Metadata Modularity on the Web**
  The diversity of metadata needs on the Web requires an infrastructure that supports the coexistence of complementary, independently maintained metadata packages. The World Wide Web Consortium (W3C) has begun implementing an architecture for metadata for the Web. The Resource Description Framework, or RDF, is designed to support the many different metadata needs of vendors and information providers.

For further information see the Dublin Core Home Page http://purl.org/dc/ .

Description of the Elements [11]

All Elements are optional and can be repeated as often as needed.

1.Title	Label: TITLE The name given to the resource by the CREATOR or PUBLISHER.
2.Author or Creator	Label: CREATOR The person or organisation primarily responsible for creating the intellectual content of the resource. For example, authors in the case of written documents, artists, photographers, or illustrators in the case of visual resources.

3.Subject and Keywords	Label: SUBJECT The topic of the resource. Typically, subject will be expressed as keywords or phrases that describe the subject or content of the resource. The use of controlled vocabularies and formal classification schemas is encouraged.
4.Description	Label: DESCRIPTION A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.
5.Publisher	Label: PUBLISHER The entity responsible for making the resource available in its present form, such as a publishing house, a university department, or a corporate entity.
6.Other Contributor	Label: CONTRIBUTOR A person or organisation not specified in a CREATOR element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organisation specified in a CREATOR element (for example, editor, transcriber, and illustrator).
7.Date	Label: DATE The date the resource was made available in its present form. Recommended best practice is an 8 digit number in the form YYYY-MM-DD as defined in http://www.w3.org/TR/NOTE-datetime, a profile of ISO 8601. In this scheme, the date element 1994-11-05 corresponds to November 5, 1994. Many other schema are possible, but if used, they should be identified in an unambiguous manner.
8.Resource Type	Label: TYPE The category of the resource, such as home page, novel, poem, working paper, technical report, essay, dictionary. For the sake of interoperability, TYPE should be selected from an enumerated list that is under development in the workshop series at the time of publication of this document. See http://sunsite.berkeley.edu/Metadata/types.html for current thinking on the application of this element
9.Format	Label: FORMAT The data format of the resource, used to identify the software and possibly hardware that might be needed to display or operate the resource. For the sake of interoperability, FORMAT should be selected from an enumerated list that is under development in the workshop series at the time of publication of this document.
10.Resource Identifier	Label: IDENTIFIER String or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented). Other globally-unique identifiers, such as International Standard Book Numbers (ISBN) or other formal names would also be candidates for this element in the case of off-line resources.
11.Source	Label: SOURCE A string or number used to uniquely identify the work from which this resource was derived, if applicable. For example, a PDF

	version of a novel might have a SOURCE element containing an ISBN number for the physical book from which the PDF version was derived.
12.Language	Label: LANGUAGE Language(s) of the intellectual content of the resource. Where practical, the content of this field should coincide with RFC 1766. See: http://ds.internic.net/rfc/rfc1766.txt
13.Relation	Label: RELATION The relationship of this resource to other resources. The intent of this element is to provide a means to express relationships among resources that have formal relationships to others, but exist as discrete resources themselves. For example, images in a document, chapters in a book, or items in a collection. Formal specification of RELATION is currently under development. Users and developers should understand that use of this element is currently considered to be experimental.
14.Coverage	Label: COVERAGE The spatial and/or temporal characteristics of the resource. Formal specification of COVERAGE is currently under development. Users and developers should understand that use of this element is currently considered to be experimental.
15.Rights Management	Label: RIGHTS A link to a copyright notice, to a rights-management statement, or to a service that would provide information about terms of access to the resource. Formal specification of RIGHTS is currently under development. Users and developers should understand that use of this element is currently considered to be experimental.

Targets of the DC:

- to provide a small set of Elements and keep this set as small as possible.
- to be syntax independent
- all elements are optional
- all elements are repeatable
- each element can be extended by an optional qualifier

Dublin Core has not reached its final format. A new version of Dublin Core, Dublin Core 2 will be released in fall 1999 which will deal with less elements. This version will not only be orientated to the needs of the librarian community. Likely the DC will grow to an open standard and will be used in many different kinds where metadata are used.
About disadvantages of the Dublin Core and how to get around them, please read [6].

Recommendation:

The Dublin Core Set will change a little bit, but these changes can be easily altered in existing systems. The Dublin Core Set is an extreme small set, but all essential bibliographic metadata are stored.

5. Document Structure

Document storage formats

The document format is mainly important regarding three issues:
- size of a document which relates to the size of the required hard disk and transfer time
- recommended exchange format as a common format between partners
- need of a special viewer or not to display and print the document when retrieved

To use these formats the images have to be converted into text e.g. by OCR or the text has to be produced manually, by typing in.
One big advantage of these formats is that they can be read by blind and visually handicapped people although sometimes it is necessary to use a plug-in. (e.g. for the PDF Format from Adobe)
A big disadvantage of these formats is, that they can be altered very easily.

It is essential that the format of electronic documents be carefully considered, for there are many choices, ranging from proprietary formats, such as WordPerfect, to nonproprietary formats, such as HTML. Proprietary formats are inappropriate for the long-term storage and use of electronic texts because of changing hardware, operating systems, and application software. Documents created using MultiMate a decade ago can no longer be used, because the company has long since ceased to exist. Most contemporary word processors cannot convert Multimate files to their own proprietary format. Ten years is a long time for software companies, but a brief time in library terms.

In general, documents encoded in a program's proprietary format can only be used with that program for the particular uses which that software allows. For example, documents created with Adobe Acrobat can only be used with that software. A user cannot create a concordance of the text or use other textual analysis software on it.

Even when conversion is possible between formats, such as between WordPerfect and Microsoft Word, it is rarely simple or flawless. Librarians need to ensure that the representation format will allow the text to be used under a variety of operating systems and on a variety of hardware platforms. Even texts distributed on CD-ROM, bundled with their own search software, may only work in a single operating system/hardware environment.

Text formats that will be discussed later on:
- ASCII
- HTML
- SGML
- RDF
- PostScript
- PDF
- RealPage
- T_EX
- Ebind
- TEI

5.1. Portable Document Format (PDF)

PDF is a file format used to represent a document independent of the application software, hardware, and operating system that were used to create it. A PDF file contains a PDF document and other supporting data. A PDF document contains one or more pages. Each page in the document may contain any combination of text, graphics, and images in a device- and resolution-independent format. This is the page description. A PDF document may also contain information possible only in an electronic representation, such as hypertext links.

PDF limitations: Printing a PDF document requires installing the article embedded fonts on the end-user's machine and several steps in order to convert the file to a postscript format. Pages are not necessarily stored in sequential order in the PDF file.

PDF strengths: PDF is primarily a portable format. To reduce file size, PDF supports a number of industry-standard compression filters: JPEG compression, CCITT Group 3 & 4, LZW. PDF viewers are supported freely from Adobe and exist for all platforms: UNIX, Macintosh and PC environments. Supporting hyperlinks is also a helpful feature. Editing a PDF document requires professional tools which may be helpful to guarantee the authenticity of the original document. This compared to HTML for example which presents the content as a free text that may be easily modified by a novice end-user.

Advantages:
- able to handle hyperlinks

Disadvantages:
- proprietary format
- a separate program has to be used to view the text
- plug-in for blind users is required
- the text is not searchable

5.2. ASCII

The *American Standard Code for Information Interchange* is a 7-bit code and a widely embraced standard. Project Gutenberg and other similar projects insist that texts must be in ASCII form, without any tagging. However, the basic ASCII character set is limited to a narrow range of characters and numbers, and it provides for no standard way of representing any other characters. The extended ASCII character set provides for some frequently used Western accented characters, but this is not common to all platforms. Also, unencoded ASCII provides no standard way of documenting bibliographic sources, creating intra- or inter-textual links, or indicating the various structural, syntactic, or semantic elements of a text. For example, if one wished to find all occurrences of a particular word in a collection of unencoded ASCII electronic verse, the computer would be unlikely to be able to distinguish between bibliographic, other extra-textual information, and the text itself. It could not even tell in a standard way where a poem begins and ends. Consequently, simple unencoded ASCII is not a good representation format for most electronic texts.

5.3. HTML

To publish information for global distribution, one needs a universally understood language, a kind of publishing mother tongue that all computers may potentially understand. The publishing language used by the World Wide Web is HTML (*Hyper Text Markup Language*).

HTML gives authors the means to:

- Publish online documents with headings, text, tables, lists, photos, etc.
- Retrieve online information via hypertext links, at the click of a button.
- Design forms for conducting transactions with remote services, for use in searching for information, making reservations, ordering products, etc.
- Include spread-sheets, video clips, sound clips, and other applications directly in their documents.

5.4. SGML

SGML is a *Standard Generalized Markup Language* defined in ISO standard 8879:1986 and is the best and the most widely accepted representation format for electronic texts. SGML takes the concept of descriptive markup beyond the level of other markup languages. By defining the role of each piece of text in a formal model, users of programs based on the SGML can check that each element of text is used in the correct place. SGML allows computers to check, for example, that users do not accidentally enter a third-level heading without first having entered a second-level heading.
SGML guidelines for markup languages require the use of angle brackets to enclose the name of elements, which are commonly called "tags." Entity references are used to replace files, character strings or words, or more commonly, accented or foreign characters (e.g., an accented "e" is represented by é). Attributes are used to further refine and define elements (e.g., within the <A> tag in HTML, HREF and NAME are commonly used attributes). All of these elements are defined within a Document Type Definition (DTD). Along with many other markup languages, both HTML and the TEI Guidelines are forms of SGML, and they have the features described above. (HTML has a Document Type Definition called HTML.DTD, but many users are not aware of its existence because it is hidden by Web browsers.)

A SGML file consists of 3 parts:

1. SGML Definition (syntax)
   The SGML declaration, made up of character sets and codes particular to the version of SGML being used, is typically kept in compiled tables by the SGML processor and isn't normally seen by the user.
2. Document Type definition (DTD)
   The DTD contains the list of elements, attributes, entities and other declarations by which the tags that mark up the text are distinguished. The DTD may be exist separately from the texts that refer to it, or may be contained within the texts.
3. The document Instance
   The Document Instance is the text itself. An important distinction between Document Instance and DTD is that the Document Instance can not contain any tag definitions; any tag within a text not defined by the text's DTD will not by recognised as such.

SGML is very good to describe documents but it is to extensive to be used. It is better to use one of its subsets like XML, TEI or Ebind.

These formats will be described later on.

5.5. XML

Extensible Markup Language (XML) is descriptively identified as an extremely simple dialect of SGML. The goal of which is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML, for which reason "XML has been designed for ease of implementation, and for interoperability with both SGML and HTML." The next version of 'HTML' is expected to be reformulated as an XML application, so that it will be based upon XML rather than upon SGML.

XML is fully internationalised for both European and Asian languages, with all conforming processors required to support the Unicode character set in both its UTF-8 and UTF-16 encoding. The language is designed for the quickest possible client-side processing consistent with its primary purpose as an electronic publishing and data interchange format.

XML documents are made up of storage units called entities, which contain either parsed or unparsed data. Parsed data is made up of characters, some of which form the character data in the document, and some of which form markup. Markup encodes a description of the document's storage layout and logical structure. XML provides a mechanism to impose constraints on the storage layout and logical structure. A software module called an XML processor is used to read XML documents and provide access to their content and structure. It is assumed that an XML processor is doing its work on behalf of another module, called the application. This specification describes the required behaviour of an XML processor in terms of how it must read XML data and the information it must provide to the application.

For more information have a look at [17]

5.6. PostScript

PostScript is a programming language optimised for printing graphics and text (whether on paper, film, or CRT is immaterial). In the jargon of the day, it is a page description language. It was introduced by Adobe in 1985 and first appeared in the Apple LaserWriter. The main purpose of PostScript was to provide a convenient language in which to describe images in a device independent manner. This device independence means that the image is described without reference to any specific device features (e.g. printer resolution) so that the same description could be used on any PostScript printer (say, a LaserWriter or a Linotron) without modification. In practice, some PostScript files do make assumptions about the target device (such as its resolution or the number of paper trays it has), but this is bad practice and limits portability.
Viewing PostScript documents is not that easy because a separate program like Ghostscript that can handle postscript has to be installed on the system. For printing a Postscript printer or a program that can convert postscript to a printer image is necessary.

5.7. Realpage

This format is rather unknown in Europe but very popular in America. 20% of the American libraries use or even support this format. The structure is equal to that PDF provides. A separate Browser has to be installed to view the pages. The big difference to PDF is, that the Browser used for Realpage is an online-browser. This means that the browser uses a TCP/IP

connection to download the pages needed from a Server over the Internet. The browser is also able to download temporary fonts necessary to display a page. One advantage is that the texts are searchable.

Advantages:
- text is searchable
- able to handle hyperlinks

Disadvantages:
- proprietary format
- a separate program has to be used to view the text
- permanent connection to the net is required

5.8. T_EX / LaT_EX

LaTeX is a document preparation system, a special version of Donald Knuth's TeX program. TeX is a sophisticated program designed to produce high-quality typesetting, especially for mathematical text. It takes a computer file, prepared according to the "rules" of TeX, and converts it to a form that may be printed on a high-quality printer, such as a laser writer, to produce a printed document of publication quality.
In former days TEX was mainly used on UNIX machines but in recent years it spread out the PC sector. The big advantage of TEX is that it can be used to display mathematical text and that the TeX text is searchable, also the parts with mathematical stuff. Unfortunately there are no WYSIWYG editors.

Advantages:
- can be used to display mathematical text
- can produce HTML source code
- can be used to produce Braille code

Disadvantages:
- no WYSIWIG editors
- viewer required

5.9. Ebind and TEI

The *Electronic Binding Project*, or Ebind, is a method for binding together digital page images using an SGML document type definition (DTD) It was developed at UC Berkeley in 1996. The Ebind SGML file records the bibliographic information associated with the document in an Ebind header, the structural hierarchy of the document (e.g, parts, chapters, sections), its native pagination, textual transcriptions of the pages themselves, as well as optional meta-information such as controlled access points (subjects, personal, corporate, and geographic names) and abstracts which can be provided all the way down to the level of the individual page.
The structure of the Ebind DTD is based loosely on the Core tag set of the Text Encoding Initiative (TEI) DTDs. Like TEI, Ebind is divided into a bibliographic header, front matter, a body, and back matter. The front, body and back elements can themselves be divided into generic textual divisions called divs. A type attribute on the div element may specify the type of division more precisely, e.g., type = "chapter".
For a description of the bibliographic header please refer to 4.1.2 TEI Header.
The Ebind DTD can be found at http://sunsite.berkeley.edu/Ebind/ebind.dtd

Difference between Ebind and TEI:

Two fundamental concepts separate Ebind from TEI. First, Ebind privileges the physical structure of a document while TEI privileges the intellectual structure. In Ebind, the atomic unit is the page while in TEI it can be down to the individual character. In TEI there is no element which can contain a page. The reason for this is that two distinct structural hierarchies cannot exist within the same document, at least not in current implementations of SGML. If a chapter ends in the middle of a page and a new chapter begins on that same page, one cannot explicitly describe both the hierarchy of the page and the hierarchy of the chapter. TEI favors the chapter by enclosing it within a div tag and describes the hierarchy of the page implicitly through the use of the pb (page break) empty tag, one of TEI's so-called "milestone" elements. In Ebind, all pages are enclosed within a <page> element. This allows one to gather together a variety of information associated with individual pages, such as textual transcription ("raw" OCR or keyed), page abstracts, even controlled access points for individual pages if desired.

The second fundamental difference between TEI and Ebind is that Ebind is more simple to use. It was recognized early on that Ebind would be used in a high-volume production environment and would be applied to a wide variety of documents. The same DTD can be used to encode books, manuscripts, diaries, newspapers, or magazines. For this reason, many of the requirements imposed by TEI were "loosened up" in Ebind. The DTD is far less restrictive. Page elements can occur just about anywhere, for example. They may occur between divs and in fact needn't be enclosed in divs of any kind. This greatly simplifies the task of automated markup.

5.10. Other formats

Some other formats like DOC and RTF have not been reviewed. They are proprietary formats and sometimes their appearance change. (Like Microsoft decides) It is not guaranteed that these formats will exist in the future nor if they will exist in this way. Therefor they are not usable for the bibliographic sector.

5.11. Media for long-time archiving

There are 3 possibilities for this purpose.

- DAT streamer tapes up to 40GB and backup systems with multiple cartridges.
- the good known CD, with a storing capacity of 640MB, 700MB for some overlength CD’s
- the “new” CD, the so called DVD (**D**igital **V**ersatile **D**isk) with a storing capacity from 2,6 GB (Single sided Rewriteable) up to 17 GB (Double Sided Read Only Memory)

For exchanging documents or data we should only have a look at the CD and the DVD because CD-Readers are standard and there should be one in every computer. DVD-Readers will become a standard soon. DAT streamer are quite costly and not built in every computer, so exchanging data is not easy.
The DVD, known as the Digital Video Disc, should soon make the CD obsolete. This new storage medium establishes high standards for digital recording of audio and video, and represents a major advance in storage capacity over CD-ROMs. It will also impose these standards on the personal computer (PC), facilitating the merger of PCs with televisions and VCRs.
If the DVD is written in-house than the storing capacity is up to 12 times higher than the one CD provides.

Regarding to experts in long terms DVD's will take over the place of CD's. But this was said a long time ago and it is not to foresee when this will happen. At this time the demand for DVD-readers and writers is quite low. DVD-readers cost twice as much as normal CD-readers. DVD-Writers and DVD-Media will cost up to 20 times more than CD-writers and media
But as usual, as the market grows these devices will get cheaper.

Supporting platform's:
Windows98
Windows NT 5.0
Mac
Unix

Storage Capacity
DVD readers and writers work exactly the same way that CD readers and writers work. For computers, the DVD is a transparent replacement for the CD-ROM, with the added benefit of a much larger storage capacity - ten to thirty times larger than CD-ROMs. The DVD comes in three versions:

- DVD ROM (Read Only Memory)
  for publishing music, videos, and software
- DVD WORM (Write Once Read Many)
  for storing documents and making single copies of music, videos, and software
- DVD RW (Re-Writeable)
  for making working copies of all types of documents that will be changed over time and for making temporary backups.

DVD's come in two sizes: the mini-CD size of about 3 1/2 inches in diameter, and the standard CD size of just over 4 inches in diameter. These DVD's look just like standard CDs, and have two usable sides. Each side can have two layers for a total of four layers per disc. Currently, there are no DVD drives that have two heads, so DVD's that have information recorded on two sides must be turned over. Two-headed DVD drives are technically possible, and will ultimately eliminate the need for disc flipping.

DVD ROM (Read Only Memory)	(maximum of two layers per side)
Top Layer	4.7 GB
Bottom Layer	3.7 GB
Single sided (two layers)	8.5 GB
Double sided (two layers per side)	17 GB
DVD WORM (Write Once, Read Many)	(maximum of one layer per side)
Single sided (one layer)	3.8 GB (lowest capacity of competing standards)
Double sided (one layer per side)	7.6 GB
DVD RW (Read Write, Rewriteable)	(maximum of one layer per side)
Single sided (one layer)	2.6 GB (lowest capacity of competing standards)
Double sided (one layer per side)	5.2 GB

DVD's Storage Capacity

State of the art

Servers with DVD-players and jukebox systems for storing up to 600 DVD‘s and a production street with CD- and DVD-writer and a printer for labelling the CD’s.

Recommendations:

In the long run DVD’s will replace the CD’s. For this reason we should stick to this new technology. Since DVD’s are able to read CD’s there is no problem in migrating CD’s to DVD’s.

6. Document Identifiers

Storing and Archiving

The amount of data a *Document Management System* has to store is quite large. Saving all this amount of data on one server is not always possible or useful. And sometimes documents are already stored in several places. Nowadays finding a document on the net, you have to use a *URL* (Uniform Resource Locator) like http://www.uni-linz.ac.at/icc (if you want to find something about the International Computer Camp for partially sighted and blind children and teens). But the Internet is growing and the structure is still changing. One big disadvantage of URL’s is for some reasons, that they have a very short lifetime. A domain names can change or the internal structure of a server can be updated. To avoid this problem of “broken links” the *Internet Engineering Task Force* (IETF) has developed the URN (*Uniform Resource Name*) Framework.

6.1. PURL

Persistent URLs developed by OCLC (Online Computer Library Centre) fulfils some of the requirements URN have to fulfil. As the name says PURL is not a real URN, it is just a URN with a long lifetime. If required URN’s can be easily migrated to real URN’s.
PURLs are, actually, URLs that instead of pointing directly to the location of a resource, point to an intermediate resolution service -the PURL Resolution Service. This resolution service associates each PURL with the actual URL and returns that URL to the client who proceeds to retrieve the intended resource in the ordinary fashion. Thus, PURL is in fact a standard HyperText Transfer Protocol (HTTP) redirection. A PURL Resolver can resolve a PURL and append the unresolved portion to the end of the resolved URL. This feature, called partial redirection, allows a PURL to be associated with a site’s root directory and still provide links to all the site’s subdirectories, saving considerable overhead for the publisher-maintainer of the site, who deploys PURL links to his site.

A PURL consists of three parts: (1) the protocol, (2) the Resolver Web address, and (3) the name of the resource. PURLs provide the means to assign a name for a network resource that is persistent, even if the item changes its actual location.
If the home page changes location, only a single change to the PURL database is required and instances of the PURL in documents will remain valid. PURL will always, if properly maintained, point to the current home page no matter where the exact location of the home page is.

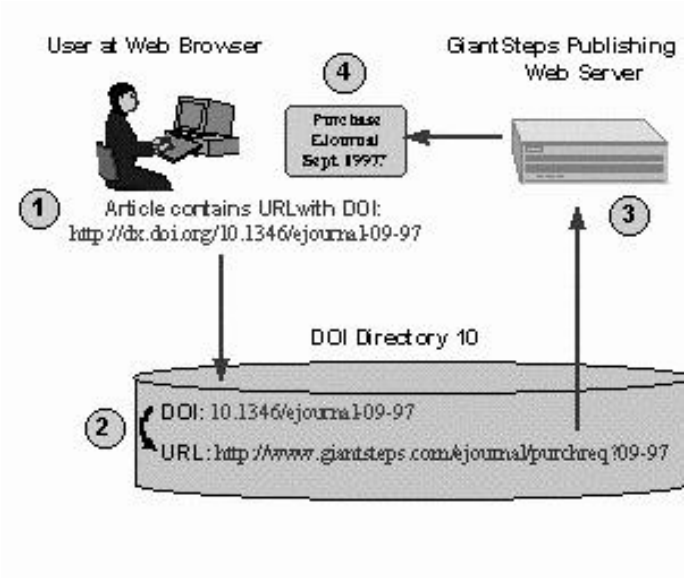
6.2. Handles

Another URN implementation, developed by the Corporation For National Research Initiatives (CNRI), is the *Handle System* which assigns, manages, and resolves persistent identifiers, known as 'handles', for Web resources. The Handle System introduces an innovative set of protocols, which enable a distributed computer system to store handles of digital resources and redirect them into the respective information in order to locate and access the resources. Additionally, the system provides the appropriate tools for this associated information to be changed, so as to reflect the current state of the identified resource without changing the handle, thus allowing the name of the item to persist over changes of location. These features are implemented in a way compliant the URN scheme, making the Handle System an efficient distributed global naming service for the management of Web resources.

A *Handle* consists of 2 parts: The „naming authority„ for identifying the host and the unique string that identifies the object. Handles can be resolved by a global so called Handle Service. There are 4 servers in America, 1 for Europe is planned.
Local Organisations can also set up a Handle Service that can be integrated in the global system. Software libraries for programmers are available, so that their systems can communicate directly with these servers.

6.3. DOI

The underlying technology for the *DOI* system was developed by the *Corporation For National Research Initiatives* (CNRI) and it is based on the Handle System protocols and specifications. The *Digital Object Identifier* is more than just an Identifier, it is an Initiative from AAP (*Association of American Publishers*) and could be considered as a commercial implementation of the Handle System which facilitates electronic commerce and supports copyright management systems. This is accomplished by including information -such as copyright, ownership, and instructions for the acquisition of the resource- in the DOI link associated with that particular resource. The publisher preserves all the rights for that resource and is responsible for giving restricted access to it.



The biggest disadvantage of the system is, that it is not free. For registering a fee has to be paid and there is also an annual fee.

6.4. Uniform Resource Name (URN)

A URN is a global unique, persistence identifier and will be used to find a resource or an amount of information.

Advantages of URN's :

- Global scope: A URN is a name with global scope which does not imply a location. It has the same meaning everywhere.
- Global uniqueness: The same URN will never be assigned to two different resources. *(The combination of naming-authority and opaque-string guarantees uniqueness.)*
- Persistence: It is intended that the lifetime of a URN be permanent. That is, the URN will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name.
- Scalability: URNs can be assigned to any conceivable resource as there are no restrictions with the opaque-string.

Identifiers that can be used as URN's for DIEPER

- ISSN
- SICI
- NBN (National Bibliographic Numbers)
- private Identifiers (e.g. DIEPER Article Number)

6.4.1 International Standard Serial Number (ISSN)

ISSN is a number that provides for the unique identification of a serial publication. It is assigned to a serial's title. The ISSN appears as two groups of four characters separated by a hyphen. A unique one-to-one correspondence exists between each assigned ISSN and the serial title to which it is registered; once registered, an ISSN is not reassigned (ANSI Z39.9-1992).
For a detailed description and how to obtain an ISSN please read [7]

Recommendation:

An ISSN can be requested for old publications which do not have an ISSN by now. Therefore ISSN is suitable for unique identifying serials.

6.4.2 SICI (Serial Item and Contribution Identifier)

This *Serial Item and Contribution Identifier* (SICI) standard defines a variable length code that will provide unique identifications of serial items (e.g., issues) and the contributions (e.g., articles) contained in a serial title. The standard is intended primarily for use by those involved in the use or management of serial titles and their contributions. While the SICI code

is intended to be applicable to both automated parsing and human-readable environments, it does not prescribe any specific machine-scannable symbology, nor does it prescribe a specific machine-readable interchange format for electronic transmission of the coded data.

Goals
The goals that guided the work of the first committee as stated in the foreword to the 1991 version of the standard, were:

- to limit the scope of the standard to a code for unique identification of serial items and contributions
- to cover the broadest possible range of serials; for example, scholarly, trade, and popular, as well as domestic and foreign, regardless of physical form
- to allow independent derivation of the SICI code from the actual serial or citation to it, regardless of whether the serial is currently published and/or whether the publisher has placed the identifier on the serial
- to provide the briefest possible code consistent with unique identification
- to maintain consistency with and build upon other standards, such as the ISSN.

The first area addressed was the disambiguation of the SICI structure. ANSI/NISO Z39.56-1991 had established two levels of coding:

- Serial Item Identifier -- a unique code for the identification of an issue of a serial title
- Serial Contribution Identifier -- by adding data elements to the code that identifies the Serial Item, a unique code is created for each contribution that appears in the serial, even if more than one contribution begins on a given page (e.g., newspapers).

Principles and Guidelines
Implementation of this standard particularly by publishers and distributors of information about both the serial items and contributions will ensure that the coded information that uniquely describes these items and contributions is readily available.

The SICI uses the International Standard Serial Number (ISSN) to identify the serial title. Therefore, in order to use this standard in the construction of an item or contribution identifier for material published in the serial, the serial must have been assigned an ISSN.

In recognition of the large installed base of serial titles, contributions, and derived works (e.g., abstracting and indexing) databases, no data elements outside those normally associated with such works are introduced into this standard.

The Structural Model for Identifiers
The SICI is a combination of defined segments, all of which are required. These segments are:

- Item Segment, the data elements needed to describe the serial item (ISSN, Chronology, Enumeration).
- Contribution Segment, the data elements needed to identify contributions within an item (Location, Title Code, and other numbering schemes in a specific instance of the SICI).
- Control Segment, the data elements needed to record those administrative elements that determine the validity, version, and format of the code representation. This is the

most important segment of the SICI. Interpretation and processing are determined by the Control Segment.

For a detailed description of SICI have a look at
http://sunsite.Berkeley.EDU/SICI/version2.html

7. Exchange Formats

Formats for exchanging images have been described in chapter 3.2, and for exchanging documents in chapter 5. This chapter is dedicated to exchange, import and export the structure of a periodical, issue or book with all of it's text and images from one server to another.

7.1. RDF

The *Resource Description Framework* (RDF) [8] is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF emphasises facilities to enable automated processing of Web resources. RDF can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities, in cataloguing for describing the content and content relationships available at a particular Web site, page, or digital library, by intelligent software agents to facilitate knowledge sharing and exchange, in content rating, in describing collections of pages that represent a single logical "document", for describing intellectual property rights of Web pages, and for expressing the privacy preferences of a user as well as the privacy policies of a Web site. RDF with digital signatures will be key to building the "Web of Trust" for electronic commerce, collaboration, and other applications.

One of the goals of RDF is to make it possible to specify semantics for data based on XML in a standardised, interoperable manner. RDF and XML are complementary: RDF is a model of metadata and only addresses by reference many of the encoding issues that transportation and file storage require (such as internationalisation, character sets, etc.). For these issues, RDF relies on the support of XML. It is also important to understand that this XML syntax is only one possible syntax for RDF and that alternate ways to represent the same RDF data model may emerge.

The broad goal of RDF is to define a mechanism for describing resources that makes no assumptions about a particular application domain, nor defines (a priori) the semantics of any application domain. The definition of the mechanism should be domain neutral, yet the mechanism should be suitable for describing information about any domain.

This specification will be followed by other documents that will complete the framework. Most importantly, to facilitate the definition of metadata RDF will have a class system much like many object-oriented programming and modelling systems. A collection of classes (typically authored for a specific purpose or domain) is called a schema. Classes are organised in a hierarchy, and offer extensibility through subclass refinement. This way, in order to create a schema slightly different from an existing one it is not necessary to "reinvent the wheel" but one can just provide incremental modifications to the base schema. Through the sharability of schemas RDF will support the reusability of metadata definitions. Due to RDF's incremental extensibility, agents processing metadata will be able to trace the origins of schemata they are unfamiliar with back to known schemata and perform meaningful actions on metadata they weren't originally designed to process. The sharability and extensibility of

RDF also allows metadata authors to use multiple inheritance to "mix" definitions, to provide multiple views to their data, leveraging work done by others. In addition, it is possible to create RDF instance data based on multiple schemata from multiple sources (i.e., "interleaving" different types of metadata). Schemas may themselves be written in RDF; a companion document to this specification, describes one set of properties and classes for describing RDF schemas.

The foundation of RDF is a model for representing named properties and property values. The RDF model draws on well-established principles from various data representation communities. RDF properties may be thought of as attributes of resources and in this sense correspond to traditional attribute-value pairs. RDF properties also represent relationships between resources and a RDF model can therefore resemble an entity-relationship diagram. (More precisely, RDF Schemas — which are themselves instances of RDF data models — are ER diagrams.) In object-oriented design terminology, resources correspond to objects and properties correspond to instance variables.

The RDF data model is a syntax-neutral way of representing RDF expressions. The data model representation is used to evaluate equivalence in meaning. Two RDF expressions are equivalent if and only if their data model representations are the same. This definition of equivalence permits some syntactic variation in expression without altering the meaning.

8. Retrieving the Text

Text has an intricacy and complexity which places great demands on software. Text is filled with synonyms and variations in capitalisation, spelling, and word forms. The searching features in text management software are more suited to text than those found in other types of software.

Text management software can employ a variety of term searching techniques:

1. Word or exact phrase searching.
2. Truncation (right, left, and internal).
3. Case insensitivity (often with case sensitivity as an option in a particular search).
4. Proximity searching: specifying how close words are to each other.
5. Field specification: in software that divides information into fields, being able to specify which fields the search term should appear in.
6. Boolean operators (AND, OR, and NOT).
7. Parentheses and nesting of Boolean operators.
8. Usage of filters which can remove common words like articles

Several system capabilities can save the user time:

1. Building and manipulating multiple search statements.
2. Saving searches for later reuse.
3. Hedges or macros: storing multiple words which can be used in a search by entering the name of the hedge or macro.
4. Exploding sections of a hierarchical thesaurus.

A variety of methods can be used to increase searching consistency:

1. Use of a thesaurus for data entry, editing, searching.
2. Data validation when data is input.
3. Mapping from abbreviations or codes to full terms.

These searching features are familiar to users of the typical bibliographic and non bibliographic text databases commonly used in libraries. However, underlying these searching features are certain assumptions:

1. The user knows what words are used in the text.
2. The user knows how to spell.
3. The user knows how to type.

To help users find the text they want, some programs are
adding more flexible searching features, such as the following:

1. Spelling checkers.
2. Automatic plurals.
3. Sound-alike searching (useful for finding spelling variations, particularly in names).
4. Fuzzy searching: searching for variations in a word or phrase. For example, the search "full text database" could retrieve "full text data file," "free text data," and "full text searching."
5. Weighted searching: assigning weights to each search term to indicate its relative importance.
6. Ranked output: displaying search results in order of relevancy, rather than the typical alphabetical or last-in-first-out orders. There are various ways to determine relevancy,

such as the number of times the search terms appear in the text or the presence of the search terms in titles or section headings.

7. Profile: displaying a profile of the most common words in a document found using other searching techniques, thereby suggesting additional search terms to consider.
8. Similarity searching: this document is what I want - go find others like it

8.1. Text Retrieval Software

Text retrieval software searches files to find ones that match a search request. For example, text retrieval software can search the minutes of meetings that were created with a word processor, and identify all of the minutes which contain a particular word or phrase, such as "holiday hours" or "travel." Most text retrieval programs can then display the file(s) for browsing, highlighting the terms in the search request.

Text retrieval software comes in two general types: those that create indexes and those that don't. Programs that create indexes require additional time for indexing and additional disk space for the indexes, but search much more quickly. Non-indexing programs don't require the additional indexing time or space, but search more slowly because the program has to "read" each file every time it does a search. The most common type of index is the inverted index, although some programs use special proprietary methods to create smaller, space-saving indexes.

Another way of dividing this software category is by the format of the files to be searched. Most text retrieval software can search files in common word processor formats, while the less powerful programs can search only through ASCII text. Some text retrieval programs are now branching out, searching through database records, spreadsheets, and computer programs.

8.2. Text Analysis Software

Text analysis software is a loose collection of software that facilitates analysing text by performing one or more of the following operations: concordancing, coding, or statistical analysis.

Concordancing is the generation of lists of the words used in a text, accompanied by the location of the word and often some surrounding text. A concordance program offers more flexibility than a printed concordance. Users can specify what should be "concorded" (e.g., all words, all nouns, or all prefixes) and also context for the words (e.g., only a location or the surrounding sentence). More sophisticated programs allow accompanying translations or annotations. Some examples of this type of "interlinear text" are phonetic transcriptions, grammatical categories, intonation, and rhythm.

Coding is the assignment of codes to specific sections of the text to allow retrieval of those sections of text. Coding is similar to assigning keywords, except that each coded segment has a specific beginning and ending point, and codes can be overlapped and even nested. A search on "marriage" might retrieve a two paragraph coded segment in an oral history transcript, while a search for "children" would retrieve only the two sentences within those two paragraphs which were coded for children.

Statistical analysis is counting various text components, such as the number of unique words, the number of times words appear, or the distribution of words in parts of the text.

Two major uses for text analysis software are for literary or linguistic analysis of text. Text analysis software can be used to examine themes in an author's works, to determine authorship of texts of unknown origin, or to analyse the grammatical structure of a language. Fields such as history, anthropology, sociology, psychology, nursing, education, and journalism use text analysis to discover themes in interview transcripts, a process called qualitative or content analysis.

9. Relevant Standards

CD DA (Digital audio), Red Book	Physical specification of the CD-ROM	1980
CD ROM, Yellow Book	Continuation of Red Book	1983
CD-I (Interactive), Green Book	Complete Multimedia system, ISO 9660	1986
CD-R (Recordable), Orange Book		1995
CD-RW (Rewritable), Orange Book		1995
UDF	Universal Disc Format (DVD compatible, but not ISO 9660 compatible). Developed by the Optical Storage Technology Association; based on ISO 13346. UDF will replace ISO 9660.	1996
ISO 9660	Standard for the CD file formats. Predecessor: High Sierra standard (HSF). Disc size: 120 mm	1987
HSF	High Sierra Standard	1986
ISO 9171-1	Specification of the disc format (5,25“ = 130 mm)	1990
ISO 9171-2	Specification of the writing format	1990
ISO 10089	Specification of the disc format for MOs 130 mm	1991
ISO 10090	Specification of the disc format for ROMs and MOs 80 mm	1992
ISO 10091	Specification of the disc format for WORM 130 mm	1995
ISO 10149	Specification of the disc format for CD-ROM 120 mm	1995
ISO 10885	Specification of the disc format for 14“ WORM (Kodak)	1993
ISO 11560	Specification of the disc format for MO-WORM 130 mm	1992
ISO 11694-4	Specification of the file structure for optical storage cards	1996
ISO 12654	Hardware independent storage format (draft)	1996
ISO 13403	Specification of the disc format for 12“ WORM (CCS)	1995
ISO 13481	Specification of 1 GB discs 130 mm	1993
ISO 13549	Specification of 1,3 GB discs 130 mm	1993
ISO 13490-1/-2	Specification of the file structure for ROMs and WORMs	1995
ISO 13614	Specification of the disc format for 12“ WORM (SSF)	1995
ISO 13482	Specification of 2 GB discs 130 mm	1995

ISO 14517	Specification of 2,6 GB discs 130 mm	1996
ISO 8879	SGML	1986
ISO 12083	XML DTDs [9]	1998
ISO 646	ISO 7-bit coded character set for information interchange	1983
ISO/IEC 10179	DSSSL - Document Style Semantics and Specification Language. [10]	1996
ISO/IEC 10744	HyTime - Hypermedia/Time-based Structuring Language. [10]	1997
ISO 10180	SPDL - Standard Page Description Language. [10]	1995
ANSI/NISO Z39.56-1996	Serial Item and Contribution Identifier	1996
ANSI Z39.9-1992	International Standard Serial Number (ISSN)	1992
ISO 8601	Date and Time Formats	1997
RFC 1766	Tags for the Identification of Languages	1995
ISO 639	Code for the representation of names of languages - The International Organization for Standardization	1988
RFC 1327	Kille, S., "Mapping between X.400(1988) / ISO 10021 and RFC 822", University College London	1992
RFC 1521	Borenstein, N., and N. Freed, "MIME Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", Bellcore, Innosoft	1993
ISO 3166	Codes for the representation of names of countries - The International Organization for Standardization	1988
ISO 3297	International Standard Serial Numbering (ISSN)	1986
ISO 639	Code for the representation of names of languages [12]	
RFC 2482	Language Tagging in Unicode Plain Text [13]	
RFC 2231	MIME Parameter Value and Encoded Word Extensions: Character Sets, Languages, and Continuations [14]	
RFC 2277	IETF Policy on Character Sets and Languages [15]	

Other relevant standards

ANSI/NISO Z39.47	Extended Latin Alphabet Coded Character Set for Bibliographic Use.	1993
ANSI/NISO Z39.64 (R1995)	East Asian Character Code for Bibliographic Use.	1989
ISO 9	Information and documentation -- Transliteration of Cyrillic characters into Latin characters -- Slavic and Non-Slavic languages.	1995
ISO 233	Documentation – Transliteration of Arabic characters into Latin characters.	1984
ISO 259	Documentation – Transliteration of Hebrew characters into Latin characters.	1984

ISO/R 843	International system for the transliteration of Greek characters into Latin characters.	1968
ISO 8859-1	Latin alphabet no. 1. For use with at least: Danish, Dutch, English, Faeroes, Finnish, French, German, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish, and Swedish.	1987
ISO 8859-2	Latin alphabet no. 2. For use with at least: Albanian, Croatian, Czech, English, German, Hungarian, Polish, Romanian, Serbian, Slovak, and Slovene.	1987
ISO 8859-3	Latin alphabet no. 3. For use with at least Afrikaans, Catalan, Dutch, English, Esperanto, German, Italian, Maltese, Spanish, and Turkish.	1988
ISO 8859-4	Latin alphabet no. 4. For use with at least Danish, English, Estonian, Finnish, German, Greenlandic, Lappish, Latvian, Lithuanian, Swedish, and Norwegian.	1988
ISO/IEC 8859-5	Latin/Cyrillic. For use with at least Bulgarian, Belorussian, Croatian, English, Macedonian, Russian, Serbian, and Ukrainian.	1988
ISO 8859-6	Latin/Arabic alphabet.	1987
ISO 8859-7	Latin/Greek alphabet. Is suitable for multiple language applications involving Latin and Greek Scripts.	1987
ISO 8859-8	Latin/Hebrew alphabet.	1988
ISO 8859-9	Latin alphabet no. 5. For use with at least: Danish, Dutch, English, Finnish, French, German, Irish, Italian, Norwegian, Portuguese, Spanish, Swedish, and Turkish.	1989
ISO 8859-10	Latin alphabet no. 6. For use with at least: Danish, English, Estonian, Finnish, German, Greenlandic, Icelandic, Sami (Lappish), Latvian, Lithuanian, Norwegian, Faroese, and Swedish.	1992

10. Technical Glossary and Acronyms

24-bit image	A digital image that can include approximately 16 million possible colours. In this kind of image, 24 bits are allocated for the storage of each Pixel, allowing 2 to the power of 24 (or more than 16 million) colours to be represented.
8-bit image	A digital image that can include as many as 256 possible colours. In this kind of image, 8 bits are allocated for the storage of each Pixel, allowing 2 to the power of 8 (or 256) colours to be represented.
Access points	Index terms used to search a database.
Adaptive palette	Image-specific set of colours chosen to most closely represent those in the original source. Part of a custom colour look-up table.
Algorithm	A rule (often mathematical) governing computer processes.
Alpha	A value representing a pixel's degree of transparency. The more transparent a pixel, the less it hides the background against which the image is presented. In PNG, alpha is really the degree of opacity: zero alpha represents a completely transparent pixel, maximum alpha represents a completely opaque pixel. But most people refer to alpha as providing transparency information, not opacity information, and we continue that custom here.
Annotation	Information (such as arrows, pointers, words) added to an image. Annotations to a digital image might be stored in layers separate from the image.
API	Application Programmable Interface
Archival image	An image meant to have lasting utility. An "archival" digital image is generally an image kept off-line in a safe place; it is often of higher quality than the digital image delivered to the user.
Article	see Contribution.
Attribute	A characteristic of an object. This term refers to a specific XML syntactic construct; the name="value" portions of an XML tag.
Automatic Indexing	Indexing of a text done by computer without human intervention (usually by finding the words occurring most frequently within the document).
Bandwidth	The transmission capacity of a communications channel, usually expressed in bits or bytes per second (the former is also called baud rate).
Bit depth	The number of bits per palette index (in indexed-colour PNGs) or per sample (in other colour types). This is the same value that appears in <tt>IHDR</tt>.
Bit Mapped Image	An image created from a series of bits and bytes that form Pixel. Each pixel can vary in colour or gray-scale value. Also known as a raster image.

Bits per Second	Measure for the speed of data transfer through communication media..
Bordering	Automatically locating the correct edge of an image on a scan so that marking from the edge, frame, etc. is not captured.
Browse Image	A small image (usually derived from a larger one). Browse images (often called "thumbnails") permit a user to view a dozen or more images on a single screen.
Browser	Software for the interpretation and presentation of HTML documents.
Buffer	Temporary storage space within a computer system.
Byte	Eight bits; also called an octet.
Capture Device	see Scanner
CCD Array	Charge-coupled device array. Light-sensitive Diodes used in scanners and electronic cameras. These usually sweep across an image and, when exposed to light, generate a series of digital signals that are converted into Pixel values.
CCITT	Comité Consultatif International de Télégraphie et Téléphonie (ITU-T)
CCITT Group III or Group IV	The standards adopted by the International Telecommunications Union (ITU), formerly the Comiteacute; consultatif internationale de teacute;leacute;graphique et teacute;leacute;phonique (CCITT), to compress page images. All fax machines in common use employ one or both of these standards.
CD-ROM	Compact Disc Read-Only Memory. A form of write-once, disc-based, random-access data storage, usually mass-produced and distributed as a publication. At present, capable of holding approximately 550 megabytes of data.
Centering	Positioning an image properly within the digital field of vision so that it is framed appropriately.
CGI	Common Gateway Interface. A technique that allows a Web server to interface to external application such as databases.
Channel	The set of all samples of the same kind within an image; for example, all the blue samples in a true colour image. (The term "component" is also used, but not in this specification.) A sample is the intersection of a channel and a pixel.
Chromaticity	A pair of values x,y that precisely specify the hue, though not the absolute brightness, of a perceived colour.
Chronology	The date(s) on the work used by the publisher to identify the individual bibliographic unit of a serial, i.e., the cover date.
CI data	Coded information; e.g. text files.
Client/Server Architecture	A systems architecture design that divides functions (which might be part of a single application) between two or more computers. The client is the machine that requests information; the server is the machine that supplies it. A typical client/server architecture for imaging might allow a server to store and transmit a compressed file, and the client to decompress, process, and display the image.

CMYK	Cyan Magenta Yellow Black. A system for reproducing colour in print, which creates the colour spectrum using cyan, magenta, yellow, and black. Used in four-colour printing.
Colour correction	The process of altering colours as they appear in a digital image or in print to ensure they accurately represent the work depicted.
Colour depth	see Dynamic Range
Colour look-up table (CLUT)	see Palette
Colour Management System	Systems that attempt to produce consistency in the representation of colour in image files, across image capture, display, and output devices.
Colour Space	A means of representing the spectrum.
Composite	As a verb, to form an image by merging a foreground image and a background image, using transparency information to determine where the background should be visible. The foreground image is said to be "composited against" the background.
Compression/Decompression	Compression is the process of squeezing more data into a smaller storage space. Decompression is the retrieval of compressed data and its reassembly to resemble its original form (before compression). see also Lossless compression, Lossy compression
Contribution	A published item, normally identified by a title and one or more authors, that constitutes an intellectually separable part of a publication. A contribution could be an article, paper, story (as in a newspaper), column, editorial, letter to the editor, meeting calendar, news item, obituary, erratum, book review, etc.
Cover date	This is the date as it appears on the cover of the item. It is the date most prominently stated by typography or location, often appearing on the cover, spine, and/or title page. It is the date most often appearing in conjunction with the title or with the enumeration. It is the date by which most users would expect to request the publication.
CRC	Cyclic Redundancy Check. A CRC is a type of check value designed to catch most transmission errors. A decoder calculates the CRC for the received data and compares it to the CRC that the encoder calculated, which is appended to the data. A mismatch indicates that the data was corrupted in transit.
Data Model	A description of the organisation of a database. It is often created as an entity relationship diagram. Today's modelling tools allow the attributes and tables (fields and records) to be graphically created. The SQL code that defines the data structure (schema) in the database is automatically created from the visual representation.
Datastream	A sequence of bytes. This term is used rather than "file" to describe a byte sequence that is only a portion of a file. We also use it to emphasise that a PNG image might be generated and consumed "on the fly", never appearing in a stored file at all.
DBMS	Database Management System

Decompression	see Compression/Decompression
Deflate	The name of the compression algorithm used in standard PNG files, as well as in zip, gzip, pkzip, and other compression programs. Deflate is a member of the LZ77 family of compression methods
Derived Image	An image that is created from another image, usually by eliminating part of it. Common techniques used to create a derived image include taking a detail, Subsampling to a lower resolution, using Lossy compression, or using Image Processing techniques to alter an image. Also called derivative image.
Digital Camera	A camera that directly captures a Digital Image without the use of film.
Digital Envelope	A digital "container" that surrounds an image with information (or metadata). Such information might be used to find the image, guarantee its authenticity, or limit access to authorised users.
Digital fingerprint	see Watermark
Digital Image	An image composed of bits and bytes. see Bit Mapped Image or Vector graphic.
Digitising	To convert an image into binary code. Visual images are digitised by scanning them and assigning a binary code to the resulting Vector graphic or Bit Mapped Image data.
Diodes	Light-sensitive electronic components used in image capture. They function as one-way valves that sense the presence or absence of light and create a digital signal that the computer converts into Pixel values.
Document Provider	Organisation which provides on line access to primary electronic material
Document Server	Server from which is processed the secure electronic transmission of documents to the end user.
Documentation	Textual information that describes a work of art or image, recording its physical characteristics and placing it in context.
DOI	Digital Object Identifier
Download	The transfer of information from one computer to another. Frequently used to describe file transfer from a network file server to a personal computer.
DPI Dots Per Inch	Dots per inch. A measurement of the scanning resolution of an image or the quality of an output device. Expresses the number of dots a printer can print per inch, or monitor can display, both horizontally and vertically. A 600-dpi printer can print 360,000 (600 3 600) dots on one square inch of paper.
Drum Scanner	A high-quality image-capture device. The image to be captured is wrapped around a drum that spins very fast while a light source scans across it to capture a digital version of the image.
DTD	Document Type Definition. A DTD describes an SGML document. For example, Mozilla is known to be Netscape's DTD
DVD	Digital Versatile Disc

DVD RAM	Erasable DVD
DVD ROM	Read Only DVD
DVD-R	Recordable DVD (WORM technology)
Dynamic Range	The colour depth (or possible Pixel values) for a digital image. The number of possible colours or shades of gray that can be included in a particular image. 8-bit can represent as many as 256 colours; 24-bit can represent approximately 16 million colours.
ECMS	Electronic Copyright Management System
EDD	Electronic Document Delivery. Generic term which involves the identification of the user, the searching of bibliographical reference and the requesting of document (SOD or Online delivery).
Element	As used here, this term refers to a specific XML syntactic construct; i.e., the material between matching XML start and end tags.
Entity	In a database, anything about which information can be stored; for example, a person, concept, physical object or event. Typically refers to a record structure.
Entity Relationship Model	A database model that describes the attributes of entities and the relationships among them. An entity is a file (table). Today, ER models are often created graphically, and software converts the graphical representations of the tables into the SQL code required to create the data structures in the database. See data model.
Enumeration	The non-chronological scheme used by the publisher on the bibliographic unit to identify the individual bibliographic units of a serial and to show the relationship of a bibliographic unit to the serial as a whole.
EPS	Encapsulated PostScript. An image-storage format that extends the PostScript page-description language to include images.
Filter	A transformation applied to image data in hopes of improving its compressibility.
Firewall	see Security Firewall
Flatbed Scanner	An image-capture device resembling a photocopy machine. The object to be scanned is placed face-down on a glass plate. The CCD Array passes beneath the glass.
FTP	a)File Transfer Protocol. Defines how files will be transferred from one computer to another. b)A software to transfer files using the File Transfer Protocol. File Transfer protocol. Reliable file transfer protocol used on the top of the TCP/IP stack.
Gamma	The brightness of mid-level tones in an image. More precisely, a parameter that describes the shape of the transfer function for one or more stages in an imaging pipeline. The transfer function is given by the expression *output = input ^ gamma*

	where both input and output are scaled to the range 0 to 1.
GIF	Graphics Interchange Format. File format for graphics, developed by CompuServe, Inc. GIF offers the inclusion of Inline graphics to HTML.
Greyscale	An image representation in which each pixel is represented by a single sample value representing overall luminance (on a scale from black to white).
GUI	Graphical User Interface
Header	Technical information packaged with an image file, which may be of use in displaying the image (e.g., length and width in pixels), identifying the image (e.g., name or source), or identifying the owner.
HTML HyperText Markup Language	HyperText Markup Language. The language that describes Web pages contents. HTML is derived from ISO/SGML using a specific DTD. See also SGML, World Wide Web.
HTTP	HyperText Transfer Protocol.
HTTP / `httpd`	HyperText Transfer Protocol Daemon. This is the world wide Web server.
Hyperlink	see Link
Hypermedia	see Hypertext
Hypertext	A document which contains links to other documents.
IAB	see Internet Architecture Board
ICONCLASS	A system of letters and numbers used to classify the iconography of works of art, developed in the Netherlands.
ICR	Intelligent Character Recognition
IDF	International DOI Foundation
IETF	see Internet Engineering Task Force
Image capture	Employing a device (such as a scanner) to create a digital representation of an image. This digital representation can then be stored and manipulated on a computer.
Image Manipulation	Making digital changes to an image using Image Processing.
Image Processing	The alteration or manipulation of images that have been scanned or captured by a digital recording device. Can be used to modify or improve the image by changing its size, colour, contrast, and brightness, or to compare and analyse images for characteristics that the human eye could not perceive unaided. This ability to perceive minute variations in colour, shape, and relationship has opened up many applications for image processing.
Imagepac	File storage format used with Kodak's PhotoCD.
Index	A systematic guide to the contents contained in, or concepts derived from, any work or group of works. It comprises a series of entries arranged in alphabetical, chronological, numerical, or other chosen order, such as subject, and with references or indicators to show where

	each indexed item or concept is located.
Indexed colour	An image representation in which each pixel is represented by a single sample that is an index into a palette or lookup table. The selected palette entry defines the actual colour of the pixel.
Inline-Image	Graphic as part of a hypertext document. See also Linked Image.
International Standard Serial Number (ISSN)	A number that provides for the unique identification of a serial publication. It is assigned to a serial's title. The ISSN appears as two groups of four characters separated by a hyphen. A unique one-to-one correspondence exists between each assigned ISSN and the serial title to which it is registered; once registered, an ISSN is not reassigned (ANSI Z39.9-1992).
Internet	a) In general, a number of single networks which operate together like one big network. b) The worldwide network of networks.
Internet Architecture Board (IAB)	Committee for standardisation and other important decisions for the Internet.
Internet Engineering Task Force (IETF)	Committee for the analysis and clearing of technical problems with respect to Internet. The members of IETF report to the Internet Architecture Board (IAB).
Internet-Resources	Any kind of information accessible via Internet.
IPR	Intellectual Property Right
ISDN	Integrated Services Digital Network. The natural evolution of the PSTN towards a fully digital network. Allows data transfer speed up to 64 KBPS for a basic rate interface BRI
Issue	see Item
Item	A unit of publication containing one or more contributions, usually under a title, as a part of a serial.
JPEG	Joint Photographic Experts Group. Used to refer to the standard they developed for still-image compression, which is sanctioned by the International Standards Organisation (ISO).
Jukebox	A stand-alone device that can hold several optical disks or magnetic tapes at a time, making it possible to switch among them at will.
Link	Reference to an other document. If the link is used the corresponding document will be loaded.
Literal	The most primitive value type represented in RDF, typically a string of characters. The content of a literal is not interpreted by RDF itself and may contain additional XML markup. Literals are distinguished from Resources in that the RDF model does not permit literals to be the subject of a statement.
Location	The numbering of the pages, or equivalent units, of an item. In non-

	print media, the location could be, for example, frame number, screen number, reel number, etc.
Lossless compression	Process that reduces the storage space needed for an image file without loss of data. If a digital image that has undergone lossless compression is decompressed, it will be identical to the digital image before it was compressed. Document images (i.e., in black and white, with a great deal of white space) undergoing lossless compression can often be reduced to one-tenth their original size; continuous-tone images under lossless compression can seldom be reduced to one-half or one-third their original size.
Lossy compression	A process that reduces the storage space needed for an image file. If a digital image that has undergone lossy compression is decompressed, it will differ from the image before it was compressed (though this difference may be difficult for the human eye to detect). The most effective lossy-compression Algorithm work by discarding information that is not easily perceptible to the human eye.
Luminance	Perceived brightness, or greyscale level, of a colour. Luminance and chromaticity together fully define a perceived colour.
LZW	Lempel-Ziv-Welch. A proprietary lossless data-compression Algorithm.
MARC	Machine Readable Catalogue. MARC is an exchange format used to import/export bibliographic records. e.g. UNIMARC, USMARC
Medium/Format Identifier (MFI)	A two-letter alphabetic code used to indicate the form that an item takes (e.g., TB = braille, CO = online (remote), TX = printed text).
Metadata	Data to describe the documents (as e.g. libraries catalogue data). Metadata should guarantee an unique identification key for each document. Activities for standardisation: – Dublin Core Set – Warwick Framework – PURL-Concept of OCLC
MIME	Multipurpose Internet Mail Extensions. Enhancement to SMTP / 7-bit limitation. Handles all types of content through E-mail (e.g. images)
NCI data	Non-coded information; e.g. image files
Network topology	The arrangement of computers and storage devices on a network. Different topologies will create higher use on various segments of the network.
NIST	National Institute of Standards and Technology
NLC	National Library of Canada. Implementers of CanSearch, Z39.50 origin
Node	A representation of a resource or a literal in a graph form; specifically, a vertex in a directed labelled graph.
Noise	Data or unidentifiable marks picked up in the course of scanning or data transfer that do not correspond to the original.
Numbering	see Enumeration.

OCLC	On-line Computer Library Centre, Dublin, Ohio.
OCR	Optical Character Recognition
ONE	OPAC Network in Europe. European Project that aims at investigating and evaluating Z39.50 implementations and search and retrieval APIs.
OPAC	On-line public access catalogue. A common term for automated, computerised library catalogues, made available to a wide range of users.
OSTA	Optical Storage Technology Association. World-wide association of optical storage systems producers (represents > 70% of the market). Specification of the UDF based on ISO 13346 (1996)
Pagination	see Location.
Palette	The set of colours that appear in a particular digital image. Becomes part of a colour look-up table. see Adaptive palette and System palette
Pattern recognition	Computer-based recognition of forms or shapes within an image.
PDF	Portable Document Format. PDF is Acrobat's favourite format
Perl	Practical Extraction and Report Language. An interpreted language. Mostly used to implement CGI scripts for Web servers.
PhotoCD	A popular storage method for digital images. In the basic Kodak PhotoCD configuration, five different levels of image quality are stored for each image in an Imagepac.
Photoshop	A sophisticated software program, produced by Adobe Systems, for editing and processing of images.
PICT	Macintosh Picture. A storage format for digital images designed primarily for the Macintosh.
PIN-code	Personal Identification Number code usually used in order to authenticate an end-user
Pixel	The picture elements that make up an image, similar to grains in a photograph or dots in a half-tone. Each pixel can represent a number of different shades or colours, depending upon how much storage space is allocated for it. see 8-bit or 24-bit
Property	A specific attribute with defined meaning that may be used to describe other resources. A property plus the value of that property for a specific resource is a statement about that resource. A property may define its permitted values as well as the types of resources that may be described with this property.
PURL	Persistent URL.
Quality control	Techniques ensuring that high quality is maintained through various stages of a process. For example, quality control during image capture might include comparing the scanned image to the original and then adjusting colours.
RAID	Redundant Array of Inexpensive/Independent Disks. A storage device that uses several disks working together to provide large storage

	capacity and redundant backup. The use of a set of cheap small disks instead of a single large disk. Five level system, was defined at the University of Berkeley in 1987. RAID 7 Architecture (7 levels) gives access from several hosts to one array system.
Raster Image	see Bit Mapped Image
Resolution, image	Number of pixels (in both height and width) making up an image. The higher the resolution of an image, the greater its clarity and definition.
Resolution, output	The number of dots per inch, DPI, used to display an image on a display device (monitor) or in print.
Resource	An abstract object that represents either a physical object such as a person or a book or a conceptual object such as a color or the class of things that have colors. Web pages are usually considered to be physical objects, but the distinction between physical and conceptual or abstract objects is not important to RDF. A resource can also be a component of a larger object; for example, a resource can represent a specific person's left hand or a specific paragraph out of a document. As used in this specification, the term resource refers to the whole of an object if the URI does not contain a fragment (anchor) id or to the specific subunit named by the fragment or anchor id.
RGB	Red Green Blue. An additive system for representing the colour spectrum using combinations of red, green, and blue. Used in video display devices.
RPN	Reverse Polish Notation. A query basic language used in Z39.50 in order to issue search requests
Scanline	One horizontal row of pixels within an image.
Scanner	A device for capturing a digital image. see Copystand Scanner, Drum Scanner, Flatbed Scanner, and Slide Scanner
Scanning	see Image capture
Secure Location Identifier	Unique identifier which protects the document from unauthorised access.
SGML	Standard Generalised Mark-up Language. SGML is an ISO standard widely adopted by publishing professionals
Slide Scanner	A scanner with a slot to insert 35-mm slides; usually capable of scanning only 35-mm transparent material.
SSL	Secure Socket Layer. Low level packet encryption mechanism. Current version is SSL 3.0
Statement	An expression following a specified grammar that names a specific resource, a specific property (attribute), and gives the value of that property for that resource. More specifically here, an RDF statement is a statement using the RDF/XML grammar specified in this document.
Subsampling	Using an Algorithm to derive a lower-resolution digital image from a higher-resolution image (for example, eliminating every other pixel in each direction). see Derived Image
System palette	A colour palette chosen by a computer system and applied to all digital

	images.
Tags	a) In HTML: the structure and presentation of documents will be defined via tags. b) In TIFF: categories for the description of the TIFF file.
TCP/IP	Transmission Control Protocol/Internet Protocol Transmission Control Protocol / Internet Protocol. Denotes the required stack of software that allows a machine connect to the Internet. This covers layers 3 and 4 of the 7-layer OSI model.
TEI	Text Encoding Initiative. Guidelines for Electronic Text Encoding and Interchange (final version 1994). Based on SGML.
TGA	TrueVision Targa file. A storage format for Bit Mapped Image video images.
Thumbnail Image	see Browse Image
TIFF	Tagged Image/Interchange File Format. A file-storage format implemented on a wide array of computer systems. Considered an industry standard, but so open that header information is used in many different ways.
TIFF Header	Contains the file description of an TIFF file.
Title	The identifying name given to a contribution within a specific item. Note: The title of a serial does not appear in the Serial Item and Contribution Identifier.
Triple	A representation of a statement used by RDF, consisting of just the property, the resource identifier, and the property value in that order.
True-colour	An image representation in which pixel colours are defined by storing three samples for each pixel, representing red, green, and blue intensities respectively. PNG also permits an alpha sample to be stored for each pixel of a truecolour image
True-colour Image	Generally refers to 24-bit (or better) images.
UNIMARC	UNIversal MARC. Widely supported bibliographic MARC records
URL	Uniform Resource Locator. A standard addressing scheme used to locate or reference files on the Internet. Used in World Wide Web documents to locate other files. A URL gives the type of resource (scheme) being accessed (e.g., gopher, ftp) and the path to the file. The syntax used is: scheme://host.domain[:port]/path filename
URN	Universal Resource Name/Number. A storage-independent scheme under development to name all resources on the Internet, which is likely to be adopted by the Internet Engineering Task Force by late 1996. URNs are likely to supersede URL (Universal Resource Locators) for identification and referencing of networked resources.
VCR	Video Cassette Recorder. A videotape recording and playback machine that is available in several formats. One inch tape is used for mastering video recordings. Sony Umatic 3/4" tape was widely used in commercial training. VHS 1/2" tape, first used only in the home, has

	mostly replaced the 3/4" tape. Sony's 1/2" Beta tape, the first home VCR format, is defunct. Although VCRs are analog recording machines, adapters allow them to store digital data for computer backup.
Vector graphic	A digital image encoded as formulas that represent lines and curves.
Watermark	Bits altered within an image to create a pattern which indicates proof of ownership. Unauthorised use of a watermarked image can then be traced.
White point	The chromaticity of a computer display's nominal white value.
World Wide Web	WWW. An interconnected network of electronic hypermedia documents available on the Internet. WWW documents are marked up in Hypertext Markup Language. Cross references between documents are recorded in the form of URL.
WORM	Write Once Read Multiple
WORM disc	Disc, optical disc, applying WORM technology
WWW	see World Wide Web
X.400	CCITT Messaging System
X-Windows System	A network based window system which has been developed originally by the Massachusetts Institute of Technology (MIT). X-Windows (also called „X„) is mainly used for UNIX computers.
Z39.50	ANSI and ISO standard. Z39.50 is a search and retrieval protocol widely accepted within the libraries community in the USA and Europe
Zooming	Enlarging a portion of an image in order to see it more clearly or make it easier to alter. Opposite of zoom-out, which is useful for viewing the entire image when the full image is larger than the display space.

11. References

[1] Fractal and Wavlet Compression, Steven Puglia, National Archives and Records Administration
http://www.thames.rlg.org/preserv/diginews/diginews23.html

[2] The Scan FAQ, Tips and Techniques of Image Scanning, Copyright © 1993-1997 Jeff Bone, All Rights Reserved.
http://www.infomedia.net/scan/The-Scan-FAQ.html / 09.29.97 / jbone@jbone.com

[3] Stigleman, Sue. "Text Management Software." Public-Access Computer Systems Review 1, no. 1 (1990): 5-22.
http://info.lib.uh.edu/pr/v1/n1/stiglema.1n1
Copyright (C) 1990 by the University Libraries, University of Houston. All Rights Reserved. Copying is permitted for non-commercial use by computerised bulletin board/conference systems, individual scholars, and libraries. This message must appear on copied material. All commercial use requires permission.

[4] The Promise of the FlashPix Image File Format
Kevin Donovan, New Media Applications, Intermuse, a Division of Willoughby Associates, Limited
kdonovan@willo.com
http://www.thames.rlg.org/preserv/diginews/diginews22.html

[5] Library and Research Documentation, Metadata-Tags zur Erschließung von Internetquellen, Metadata-Elemente des Dublin Core
eingedeutscht von Diann Rusch-Feja
http://www.mpib-berlin.mpg.de/DOK/metatagd.htm

[6] Beyond the Dublin Core: Rich Meta-Data and Convenience-of-Use Are Compatible After All, Roger Clarke
http://www.anu.edu.au/people/Roger.Clarke/II/DublinCore.html

[7] INTERNATIONAL STANDARD SERIAL NUMBER
TAME THE SERIALS JUNGLE WITH ISSN ONLINE
http://www.issn.org/

[8] Resource Description Framework (RDF) Model and Syntax Specification, W3C Proposed Recommendation 05 January 1999
Ora Lassila (ora.lassila@research.nokia.com), Nokia Research Center
Ralph R. Swick (swick@w3.org), World Wide Web Consortium
http://www.w3.org/TR/1999/PR-rdf-syntax-19990105

[9] ISO 12083 XML DTDs http://www.xmlxperts.com/12083xml.htm

[10] Standards Related to SGML
http://www.isgmlug.org/sgmlhelp/related.htm

[11] Dublin Core Metadata Element Set: Reference Description
http://purl.org/DC/about/element_set.htm#

[12] ISO 639, ”Code for the representation of names of languages”,
http://www.indigo.ie/egt/standards/iso639/iso639-1-en.html

[13] RFC 2482, “Language Tagging in Unicode Plain Text”,
ftp://ftp.isi.edu/in-notes/rfc2482.txt

[14] RFC 2231, “MIME Parameter Value and Encoded Word Extensions: Character Sets, Languages, and Continuations”
ftp://ftp.isi.edu/in-notes/rfc2231.txt

[15] RFC 2277, “IETF Policy on Character Sets and Languages”, ftp://ftp.isi.edu/in-notes/rfc2277.txt

[16] JPEG image compression FAQ, part 1/2
http://www.dcs.ed.ac.uk/%7Emxr/gfx/faqs/JPEG.faq

[17] Extensible Markup Language (XML)
World Wide Web Consortium 8-December-1997
http://www.w3.org/TR/PR-xml.html