

Conversion & Document Formats

Backfile conversion and format issues for
information stored in digital archives

AIIM Industry White Paper on Records,
Document and Enterprise Content Management
for the Public Sector



AIIM
International



© AIIM International Europe 2002

© DLM-Forum 2002

© Hewlett-Packard 2002

© PROJECT CONSULT 2002

All rights reserved. No part of this work covered by the copyright hereon may be reproduced or used in any form or by any means – graphic, electronic or mechanical, including photocopying, recording, taping, or information storage and retrieval systems – without the written permission from the publisher.

Trademark Acknowledgements

All trademarks which are mentioned in this book that are known to be trademarks or service marks may or may not have been appropriately capitalised. The publisher cannot attest to the accuracy of this information. Use of a term of this book should not be regarded as affecting the validity of any trademark or service mark.

Warning and disclaimer

Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied. The author and the publisher shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book.

First Edition 2002

ISBN 3-936534-00-4 (Industry White Paper Series)

ISBN 3-936534-02-0 (Industry White Paper 2)

Price (excl. VAT): 10 €

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Printed in United Kingdom by Stephens & George Print Group

Conversion & Document Formats

Backfile conversion and format issues for
information stored in digital archives

AIIM Industry White Paper on Records,
Document and Enterprise Content Management
for the Public Sector

AIIM International Europe
Chappell House
The Green, Datchet
Berkshire SL3 9EH - UK
Tel: +44 (0)1753 592 769
Fax: +44 (0)1753 592 770
europeinfo@aiim.org

DLM-Forum
Electronic Records
Scientific Committee Secretariat
European Commission SG.B.3
Office JECL 3/36, Rue de la Loi 200, B-1049 Brussels - Belgium
Tel. +32 (0)2 299 59 00 / +32 (0)2 295 67 21 / +32 (0)2 295 50 57
Fax +32 (0)2 296 10 95
A/e: dlm-forum@cec.eu.int

Author
Hewlett-Packard GmbH
Nancy Romany
Schickhardtstr. 32
71034 Boeblingen – Germany
Tel. +49 (0)7031 14 6837
nancy_romany@hp.com

Executive editors and coordinators
Dr. Ulrich Kampffmeyer
Silvia Kunze-Kirschner
PCI PROJECT CONSULT International Ltd.
Knyvett House, The Causeway
Staines, Middlesex TW18 3BA - UK
Tel.: +44 (0)1784 895 032
info@project-consult.com

Published by PROJECT CONSULT, Hamburg, 2002

Industry White Papers on Records, Document and Enterprise Content Management	Series	ISBN 3-936534-00-4
(1) Capture, Indexing & Auto-Categorization		ISBN 3-936534-01-2
(2) Conversion & Document Formats	HP	ISBN 3-936534-02-0
(3) Content Management	FileNET	ISBN 3-936534-03-9
(4) Access & Protection	IBM	ISBN 3-936534-04-7
(5) Availability & Preservation	Kodak	ISBN 3-936534-05-5
(6) Education, Training & Operation	TRW/ UCL/ comunicando	ISBN 3-936534-07-1

Preface

The Information Society impacts in many different ways on the European citizen, the most visible being the provision of access to information services and applications using new digital technologies. Economic competitiveness of Europe's technology companies and the creation of new knowledge-rich job opportunities are key to the emergence of a true European digital economy. Equally, the Information Society must reinforce the core values of Europe's social and cultural heritage – supporting equality of access, social inclusion and cultural diversity. One important element in ensuring a sound balance between these economic and social imperatives is co-operation between the information and communication industries and public institutions and administrations. Over the past 5 years, the European Commission in co-operation with EU Member States, has worked to create a multi-disciplinary platform for co-operation between technology providers and public institutions and administrations. The Forum aims at to make public administration more transparent, to better inform the citizen and to retain the collective memory of the Information Society. These objectives are at the heart of the eEurope Action Plan adopted by the European Summit in Feira on June 2000. I welcome the way the DLM-Forum has evolved over this period as a platform for identifying and promotion concrete solutions to many of the problems facing our public administrations.



In 1996 the initial focus of the DLM-Forum was on the guidelines for best practices for using electronic information and on dealing with machine-readable data and electronic documentation. More recently, at the last DLM-Forum in Brussels in 1999 a challenge was made to the ICT industries to assist public administrations in the EU Member States by providing proven and practical solutions in the field of electronic document and content management.

The importance of providing public access and long term preservation of electronic information is seen as a crucial requirement to preserve the “Memory of the Information Society” as well as improving business processes for more effective government. Solutions need to be developed that are, on the one hand, capable of adapting to rapid technological advances, while on the other hand guaranteeing both short and long term accessibility and the intelligent retrieval of the knowledge stored in document management and archival systems. Furthermore, training and educational programmes on understanding the technologies and standards used, as well as the identification of best practice examples, need to be addressed. I welcome the positive response from the ICT industries to these challenges and their active involvement in the future of the DLM-Forum, for example in the event proposed in Barcelona in May 2002, to coincide with the EU Spanish Presidency.

The information contained in the following pages is one of a series of six ICT Industry White Papers produced by leading industry suppliers, covering the critical areas that need to be addressed to achieve more effective electronic document, records and content management. I am sure that the reader will find this information both relevant and valuable, both as a professional and as a European citizen.

A handwritten signature in black ink, appearing to read 'Erkki Liikanen'.

Erkki Liikanen
Member of the Commission for Enterprise and Information Society

Preface Sponsor

Hewlett Packard is pleased to be sponsoring this informative white paper. The management and accessibility of document information resources is a key issue that face European business in the 21st Century.

Businesses and society at large stand to gain immeasurably through viable and effective systems for electronic document, records and content management. They will save time, money and increase efficiency.

This white paper provides companies and institutions in Europe with the information to create their own, appropriate, best-practice path to successful document management.

As we address the issues surrounding document and content management, the expectation is the solutions outlined here will play a crucial role as we move to universal archive management systems.

A handwritten signature in black ink, appearing to read 'R. Gerstner', with a stylized flourish at the end.

Rolf Gerstner
Manager Business and Marketing Strategy, Commercial Imaging
and Printing Solutions, EMEA, Hewlett-Packard

Table of Content

1.	Introduction	7
2.	The Bottleneck of Converting Existing Archives	8
2.1	Conquering the document mountain	8
2.2	Overcoming perceptions	8
2.3	Keeping pace with technology	9
2.4	Lost Information.....	10
2.5	Conversion Costs	10
2.6	Legality	11
2.7	Records management vs. document management.....	11
3.	Content in the Digital Age	13
3.1	Paper original	13
3.2	Raster or vector imaging	13
3.3	Scanning ‘norms’	14
3.4	Compression.....	15
3.5	Eliminating image ‘noise’	15
3.6	Raster image characteristics	16
3.7	Renditioning.....	17
3.8	Adaptive thresholding	17
3.9	Annotations	17
3.10	Recognition techniques	18
3.11	Colour and greyscale.....	20
3.12	Microfilm, Microfiche.....	21
3.13	Electronic objects	21
3.14	Forms	22
4.	Information Capture Technologies and Methods	23
4.1	In-house or bureau?.....	23
4.2	Traditional archives.....	24
4.3	COLD/ERM Enterprise Report Management	26
4.4	Forms capture.....	27
4.5	Electronic documents	29
4.6	Planning a Document Conversion.....	30

5.	Standards for Formats.....	34
5.1	Tagged Image File Format (TIFF)	35
5.2	Joint Photographic Experts Group (JPEG).....	37
5.3	Graphics Interchange Format (GIF).....	37
5.4	Moving Pictures Expert Group (MPEG).....	38
5.5	Audio Video Interleave (AVI)	38
5.6	Adobe Acrobat Portable Document Format (PDF)	38
5.7	Rich Text Format (RTF)	39
5.8	HyperText Markup Language (HTML)	39
5.9	Extensible Markup Language (XML)	39
6.	Best Practice Applications	40
6.1	Forestry Geographic Information System (FOGIS) Project of the Department of Forestry, Baden-Württemberg.....	40
6.2	Document Imaging System Project of Sanctuary Housing Association	43
6.3	COLD/ERM – Solution Project of Staffordshire County Council	45
6.4	Migration of paper documents into electronic files Project of Levy Gee	48
7.	Outlook.....	50
7.1	Formats.....	50
7.2	Conversion Strategy	50
7.3	Archive Value	51
7.4	Technology Developments.....	52
	Glossary.....	54
	Abbreviations.....	57
	References	58
	Authoring Company	59

1. Introduction

This white paper addresses the issues which arise when considering the conversion of existing physical archives, that contain documents of different formats and types, into electronic format. These issues are broad in nature including the logistics of capture involving high volumes; the determination of appropriate strategies and tactics, for both delivering the conversion and maintaining normal business operations in the process; and the adoption of appropriate, reliable and sustainable document formats.

- Chapter 2 – The Bottleneck of Converting Existing Archives
The route to conversion involves the negotiation of a number of hurdles, which are addressed in this chapter; sizing, legality, cost justification, lifecycle considerations and not least, a clear understanding of the business purpose of the archive itself.
- Chapter 3 – Content in the Digital Age
A single piece of paper can be used to represent many things; the same is not true of a digital document. To adopt a ‘one size fits all’ strategy is to marginalise the value of the document, this chapter looks at how to maximise the document’s value as a corporate asset by understanding and designing solutions around its true nature.
- Chapter 4 – Information Capture Technologies & Methods
Capture methods and technologies are equally varied. While some organisations have adopted approaches which involve printing an electronically generated document in order to scan it digitally (or re-key it manually), there are better methods available. Nor are such approaches limited to technologies; the human interface is vital also.
- Chapter 5 – Standards for Formats
This chapter addresses the various formats that exist for capturing different types of documents. While no format has a guarantee of immortality, it focuses on established formats whose pedigree and sustainability are reasonably well assured with emerging standards gaining widescale acceptance.
- Chapter 6 – Best Practise Applications
This chapter assesses European organisations in the public sector which have faced and overcome the challenges of creating a digital archive to support their front line processes.
- Chapter 7 – Outlook
A summary of proven conversion strategies, factors that need to be considered for different user requirements and comments on the developments and technologies that are shaping the conversion market.



John Mancini
AIIM International

2. The Bottleneck of Converting Existing Archives

2.1 Conquering the document mountain

Despite the advances that have been made in processing data by electronic means in the last thirty years, relatively little has been achieved in converting documents for access by electronic means.

In fact, it is estimated that 90% of all business information is still held in non-digital form: mainly on paper, but also on microfilm and microfiche. Corporate computer systems have been responsible for generating much of this paper mountain—management reports printed on the ubiquitous ‘pyjama paper’, bills, invoices, and pay advices, where typically the original content is held electronically for only a few months before being wiped, leaving a paper or microfilm copy as the only record or proof of transaction.

Nor have PC-based systems helped reduce the problem. Electronic objects generated in office systems (letters, spreadsheets and similar) are often stored on an individual’s PC hard drive, or on floppy disk or local CD copy, where their value to the organisation is marginalised. Not surprisingly, such documents are often printed for archive purposes. Perhaps of more concern is the growth in the use of email. Many key decisions are now based on the content of email, yet most organisations still have to develop an effective strategy for archiving that content – even on paper.

There may also be other reasons – for example documents bearing signatures – that necessitate the paper original taking precedence over the electronic copy.

2.2 Overcoming perceptions

Document overload is not always recognised as a problem. As the organisation grows, it needs larger premises, a larger car park, an expanded HR function; surely a growing filing problem is simply another symptom of success, a necessary price to be paid?

Organisations with carefully nurtured reputations for prudence and probity have been known to take a very simplistic view of their archival needs – ‘we went out and bought a scanner’, or ‘we just want to get our files from that filing cabinet onto this PC’ – with predictable results.

Quite simply, archiving often fails to ‘hit the radar’. Will the Finance Director invest in addressing the problem, unless convinced that it will help ensure the survival and competitiveness of the organisation? The established view pervades “ship off site in boxes for a third party to store”.

The value that this information asset can be realised by conversion is very rarely appreciated by senior management.

2.3 Keeping pace with technology

In common with many other people who built up libraries of vinyl records and Betamax videos, the author is only too aware of the issues that obsolescence creates for the archivist.

Document management has seen a proliferation of storage media in recent years: 14" WORM optical disk, 12" WORM, 5.25" WORM, 5.25" Rewritable, CD-R and CD-RW, DVD, DAT and RAID to name but some. Not only do the media differ, often vendors have developed different standards for the way in which they read/write that media, so that the demise of the vendor, let alone the medium, renders the archive obsolete.

Lifetime issues also need to be considered. Few optical archives are more than 15 years old, although manufacturers such as Hewlett Packard have undertaken rigorous testing and assessment that show that their discs have 50 plus years expectancy. But media of all types can suffer from short life expectancy: the 1911 UK census was reported on paper that now crumbles when you touch it, and microfilm at a MOD site is no longer readable after 10 years due to poor quality control when the archive was created. A long-term archive strategy needs to address the longevity of both the document format and the storage medium, and requires detailed examination.

Digital storage capacity is growing exponentially, and will need to continue to do so in order to meet the demands of the document industry. Some years ago a multinational charge card company adopted document scanning for its 'country club billing' system, which depended on presenting the customer with a facsimile of the original handwritten and signed transaction. However, it discovered that the largest available optical jukebox storage system only gave it the capability to store bills for around three months – just long enough to bill the customer and resolve queries arising. The company was obliged to maintain a microfilm archive for long-term retention, and envisaged that the problem would remain until optical storage was superseded – for example crystalline storage, utilising the individual molecules of a crystal for data storage.

Changes in technology also have their impact. That same charge card company found its requirements rendered largely obsolete by the introduction of the point of sale 'cardswipe' device...

Document formats do not stand still either, to a large extent driven by the demands of the world wide web and the ubiquitous use of the PC. There are many well- established document format standards, denoted here by their acronyms but described in more detail later in this paper - TIFF, JPEG, GIF, PDF, RTF, HTML, XML - while new formats such as JPEG2000, JPM and TIFF/FX continue to emerge. All Points Addressable (APA) print-out put formats such as IBM's Advanced Function Presentation (AFP) Technologies also continue to be developed, and are seen as key competitive weapons in areas such as electronic bill presentation.

2.4 Lost Information

Existing archives represent a very limited subset of the intellectual capital assets of any organisation. Most organisations are guilty of duplicating information they already possess simply because the information they require lies buried in a filing cabinet somewhere in the organisation but is poorly registered or indexed. This was once summarised as ‘if we only knew what we knew, we’d be four times as productive’.

The most valuable attributes of any document are the content and context (metadata) in which it was created – and these attributes are largely lost once active paper or microfilm documents are committed to archive, since the information ceases to have value unless someone knows that it exists and where it exists. Records Management activity to classify/register the document - which conventionally is seen as a basement activity that takes place after the documents have moved beyond regular access - is gaining a higher profile and is seen as a front office task to address these requirements.

The EC commissioned the report “Model Requirements for the Management of Electronic Records – MoReq” which was published in May 2001. MoReq sets out the functional requirements for Electronic Records Management and many of the leading Electronic Document Management suppliers now provide ERMS functionality that meet the MoReq specification.

There are also physical reasons for documents to be unavailable. Many documents will have suffered the ravages of time – mildew, damp, faded ink, or damaged paper. Or they may be lost completely – borrowed and never returned, or simply misfiled. Microfilm archives are also not immune to such risks – not least from the obsolescence of microfilm viewers or poor quality of transcription.

2.5 Conversion Costs

Paper handling is a major factor in any operation – as much as 30-40% of overall operating cost can be accounted for simply in paper handling activities. A conversion exercise is equally labour-intensive and costly, and compounded by the fact that it is often scheduled over a compressed timescale.

Costs to consider include both the overall costs for the conversion exercise and its management overheads, and the cost of ‘business as usual’ during conversion (e.g. disruption to business processes if files are offsite during capture, or of hiring extra staff for an onsite conversion).

A carefully planned conversion strategy is vital. Backfiling strategies include full conversion; on-demand conversion; incremental conversion; or a ‘from this day forward’ strategy.

Post capture costs must also be considered. The paper may have to be retained for legal reasons (for example deeds and policies), with a cross referencing system to ensure that it can be traced back through the digital record. Even if it is to be destroyed, there may be special controls and safeguards to be applied.

Significant attention should also be paid to weeding out documents before conversion – duplicate documents, superfluous notes, documents that are irrelevant or have been superseded. It is not unknown for organisations to discover that more than 50% of their archive is more suited to shredding than conversion.

2.6 Legality

To an extent, concerns over issues of legality have in the past constrained the widespread adoption of electronic archival. In the UK for example there were many concerns regarding the legality of electronic records – what was the value of an electronic archive if the paper archive had to be maintained also? The original UK Data Protection Act then came into play, which stated that while a paper record might not fall under the remit of the Act, it would do so once captured onto a DIP system.

Gradually this situation is changing. The latest version of the Data Protection Act has removed the DIP anomaly by including paper documents within its remit; the introduction of PD0008 – 'Legal Admissibility of Information Stored on Electronic Document Management Systems', and equivalent protocols in other European countries, has helped define clear guidelines; and initiatives such as the UK Modernising Government White Paper, the Freedom of Information Act, and the legalisation of electronic signatures being implemented across the EC are now acting as a stimulus for the move to electronic archival. Quite simply, organisations in general and public bodies in particular, are realising that first class service cannot be delivered without first class support systems.

2.7 Records management vs. document management

Electronic Document Archiving to be of value has to be structured so that the user can efficiently retrieve the required document for the query or task being undertaken. Records and document management both provide this support. An Archive may require management by either or both disciplines dependent on the archive content.

Electronic Document Management Systems – EDMS - provide general controls over storing and accessing electronic documents but do not have the more structured disciplines demanded for managing and conserving electronic records.

A record, whether in paper or electronic form, is a statement of a business activity that needs to be conserved without change and holds a description of the contents and the status of that record.

Electronic Records Management Systems - ERMS - provide the control and management functions to standardise file plans, record metadata and disposition schedules for an organisation, and provide comprehensive auditing and control over the access and actions on records. These systems often provide for the management of both electronic and paper records.

An EDMS...	An ERMS...
allows documents to be modified and/or to exist in several versions;	prevents records from being modified;
may allow documents to be deleted by their owners;	prevents records from being deleted except in certain strictly controlled circumstances;
Provides few “end-of-life” management tools;	Can REQUIRE that certain documents be deleted, destroyed “shredded” at a defined point in their lifecycle
may include some retention controls;	must include rigorous retention controls;
may include a document storage structure, which may be under the control of users;	must include a rigorous record arrangement structure (the classification scheme) which is maintained by the Administrator;
is intended primarily to support day-to-day use of documents for ongoing business.	may support day-to-day working, but is also intended to provide a secure repository for meaningful business records for short and long term archiving.

Figure 1: EDMS v. ERMS

3. Content in the Digital Age

3.1 Paper original

Paper comes in many shapes and sizes; from A0 and above to A6 and below, simplex and duplex, in loose leaf, bound and book form, in many different colours and weights. Most critical of all however is content. An A4 piece of paper can be endorsed with a letter; an invoice; a map; a Last Will and Testament; an engineering drawing; a cartoon drawing; a da Vinci sketch or Einstein's Theory of Relativity. One simple piece of paper, but the methods required for creating an electronic representation of each of those items are vastly different.

The quality of the original page also matters far more to a computer than to the human eye. While the human eye can make allowances for poor handwriting, smudgy scrawls, overwriting on a patterned background, spelling mistakes and similar, a document recognition system is less adaptable. It can interpret typed text with a reasonable level of accuracy. The reading of constrained handprint (capital letters or numbers hand-written into boxes) has become a viable technology, while the reading of unconstrained handprint (capital letters or numbers NOT in boxes) is close to widespread acceptance. But it is still very poor at lower-case and cursive (joined-up) hand writing recognition, and at interpretation – for example in deducing sub totals and grand totals in a column of figures.

3.2 Raster or vector imaging

A raster image is a bit-mapped representation of an original; each pixel in the original is represented digitally, with the simplest representation being a one-for-one match. The most accurate rendition is a reproduction at exact scale, since the requirement for enlargement or reduction (for example to size in order to fit a PC window) involves the use of scaling algorithms and therefore some compromise or loss of detail.

A vector imaging interprets and represents the shapes in an original; for example a curve or a straight line can be reproduced precisely at any scale, resolution or rotation.

Vector imaging is used in applications such as digital mapping. A large-scale map is typically reproduced at 1:25000 scale: the finest detail on the map is approximately 0.1mm in width, which translates to 2.5 metres 'on the ground'.

Consider therefore the impact on those organisations which rely on such maps. Traditionally, gas, electric and water utilities have been unable to define the position of pipes and cables with precision; hence the reason for repair gangs digging trenches in the road that are up to three metres wide or more, simply in order to find the offending supply. With vector-based digital mapping, utilities can pinpoint the precise location of their pipes and cables, and minimise the disruption that repair work causes.

Vector conversion can however be an expensive and time consuming exercise - the Ordnance Survey project to digitise their 230,000 maps started in 1973 and was not completed until 1995!

Most general office applications are more suited to simpler and faster raster imaging techniques.

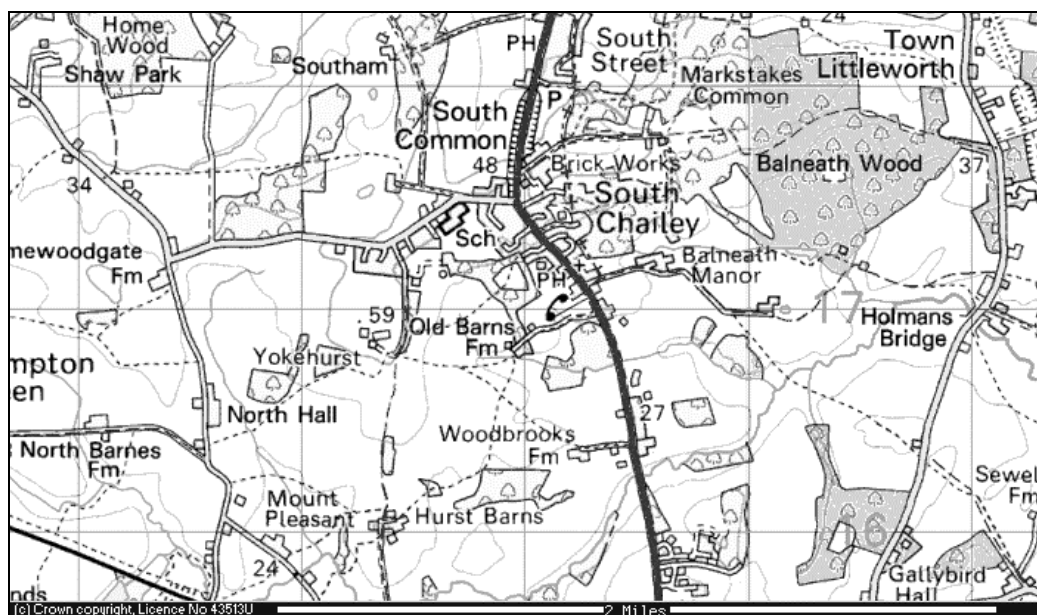


Figure 2: Ordnance Survey map extract

(Image reproduced with kind permission of Ordnance Survey and Ordnance Survey of Northern Ireland.)

3.3 Scanning ‘norms’

Consumer expectations are often now conditioned by the home scanner market; a basic scanner, costing in the order of €100 from a high street store, can scan and digitally separate a number of different items simultaneously, at resolutions of up to 1200 dots per inch (dpi) in full 24-bit colour.

Prospective users are often therefore surprised to find that a commercial scanner, costing €30,000 or more, may only be capable of monochrome scanning at resolutions of up to 300 dpi, and require page size to be pre-set manually.

But that is the nature of the commercial scanning market. There has been little requirement for high-resolution colour scanning simply because the sheer scale of the storage and network bandwidth capacity required, is not commonly a justifiable expense. For many organisations, the yardsticks of ‘acceptable quality’ remain monochrome fax, at a resolution of 100-200 dpi, and monochrome laser printing at 300 dpi.

In most instances also, the throughput capacity and robustness of the scanning engine (which translate directly into the cost of capture) are far more critical than individual image quality. While it may be cheaper to place a low cost scanner on every desk rather than a large scanner centrally, it is the ongoing labour costs that determine whether such a strategy is viable.

3.4 Compression

The most popular compression techniques for monochrome bi-tonal images are the ITU Group standards (formerly CCITT Group 3 and Group 4), which utilise a technique known as ‘run length encoding’.

Put simply, a compression engine scans a line of pixels, counting the number of pixels of the same colour (white) before a change of colour (black). The process is repeated until end of line.

For a typed page such as the one you are currently reading, consisting of 95% white and 5% black pixels, the technique works well. But for photographs and graphics, the technique is much less effective.

Group 3 compression scans each line at a time. Group 4 compression achieves greater efficiency by comparing each scanned line with the previous, and encodes only differences encountered. Hence it is described as a two-dimensional (2d) technique.

Group 4 compression is most commonly used in digital scanning, but Group 3 remains the accepted standard for fax transmission since it is most tolerant of faults arising due to noise on analogue phone lines (there is a checkpoint at the end of every line to enable the image to be ‘re-synchronised’).

3.5 Eliminating image ‘noise’

Document content may prove highly resistant to effective compression – especially poor quality photocopies and ‘cluttered backgrounds’ (for example on banknotes). This can easily be verified by faxing a ‘clean’ page and then repeating the exercise with a grainy photocopy of the same, and comparing the time taken to transmit.

Various techniques can be applied to mitigate this. These include colour dropout (choosing coloured backgrounds that are transparent to the scanning device), zoned imaging (scanning only pre-designated areas of a document), and forms dropout (electronically eliminating the background to a form).

One particularly challenging conversion exercise involved a local council, whose housing benefit forms were designed with a dithered background (a pattern of black dots simulating a shade of grey). The assumed sizing of 30 kilobytes per page for a Group 4 compression was rapidly revised when it was discovered that compression was delivering images that were over 2500 kilobytes per page!

3.6 Raster image characteristics

The ‘jagged edge’ nature of a fax document is characteristic of raster images, where an original document is rendered as a series of square pixels. This pixellation is common to all formats, although some document image viewers may be differentiated by the extent to which they minimise the effect through suitable software algorithms.

One technique in common use is scale to grey, or anti-aliasing. Consider an image scanned at 300 dpi, which is to be displayed on a PC monitor with a maximum resolution of 100 dpi. In effect, the display software must reduce each block of nine pixels to a single pixel.

The simplest technique to achieve this is a cut-off technique; if more than four of the pixels in the block are black, then the output is rendered in black; otherwise it is rendered in white. This results in the jagged effect described above.

Anti-aliasing, in contrast, counts the number of black pixels in the block and renders this as a shade of grey, with a result which is much easier on the eye.

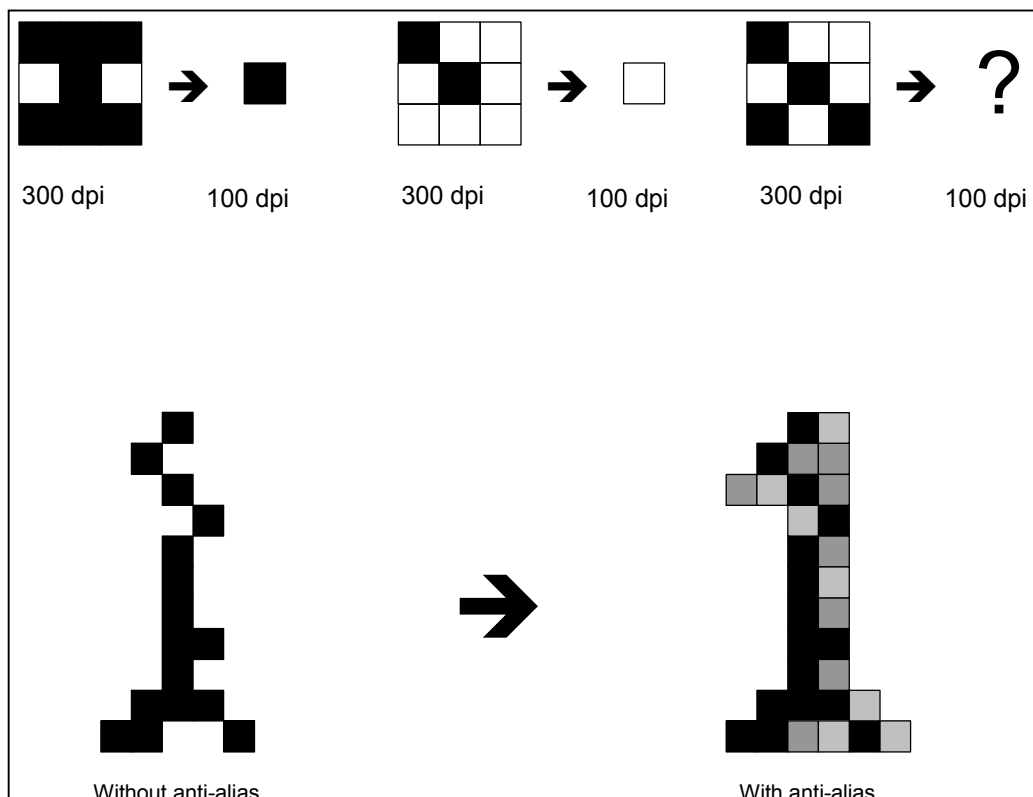


Figure 3: Anti-Aliasing

3.7 Renditioning

One option for better managing network traffic is to produce both a high quality and a thumbnail rendition of the original document. The small-size thumbnail is accessed during search and browse operations, with the full-scale image being retrieved only once the appropriate thumbnail has been identified.

Similar techniques may be used in OCR/ICR operations. For example, a high resolution image may be captured for input to the character recognition engine, after which it will be discarded. A lower quality rendition is simultaneously captured for long-term archive.

3.8 Adaptive thresholding

Adaptive thresholding is a technique, applied either in scanner hardware or software, for compensating for changes in contrast even when monochrome scanning a multi-coloured document.

Many vendors incorporate automated facilities for optimising the contrast, brightness, edge definition and other aspects of an image for maximum quality. Without such techniques, difficult originals (dark green on light green, for example) may require several manual attempts to achieve the best result, which will significantly raise the cost of the conversion exercise.

3.9 Annotations

The application of annotations within an image – for example to highlight a segment of text, to add a ‘sticky note’, or to endorse a digital date stamp – raises a number of challenges in terms of both legal admissibility and long term viability.

Some vendors offer the ability to ‘burn in’ an annotation on a TIFF document – in other words to recreate the image with the annotations fully encapsulated. This is a vendor-independent technique, but dependent upon effective version and audit controls for regulatory compliance purposes, which in themselves may be vendor dependent.

Vendors may also or alternatively, offer the ability to store annotations on separate ‘layers’ – these may be held as separate electronic objects and overlaid onto the image when displayed or printed. There are no independent portable standards in general use for the application of annotations, therefore the potential loss of such annotations needs to be considered if images are migrated from one application to another.

For many organisations, this issue could creep up on them unexpectedly. For example, the Microsoft Windows operating system has bundled a default TIFF viewer from Windows 95 onwards, with the capability to annotate images.

In the latest version, Windows XP, a new default viewer has been bundled which does not support annotations and does not display annotations made with the previous viewer unless 'burnt in'.

3.10 Recognition techniques

A number of automated recognition techniques are commonly applied during document capture.

Recognition technologies

- ICR 346 175 5073
- OCR 3461755073
- BCR 
- OMR Yes No

Figure 4: Recognition Techniques

3.10.1 Intelligent character recognition (ICR)

ICR uses neural recognition techniques to detect shapes - for example a capital 'A' contains a forward slash, a backward slash and a horizontal dash. ICR is effective on both typed and hand printed text (but not joined up handwriting), and can still be effective even when encountering a previously unseen font.

ICR is also capable of learning. A user is able to 'train' the system to recognise new patterns, which it then incorporates in its neural network.

3.10.2 Optical character recognition (OCR)

OCR is used to recognise text characters, normally by comparing the scanned bitmap of a character against stored character sets, repeating the process until a match is found (template or matrix matching).

OCR typically achieves accuracy of over 98% on typed text. This represents a mis-read of one character in fifty so that for e.g. a credit card processing application (where the average credit card number is 16-18 digits long), it means that one transaction in three will be incorrectly processed.

A number of techniques are used to improve accuracy – check digit verification, cross matching of address against zip code, totals balancing and spell checking, for example.

3.10.3 Bar coding (BCR)

BCR provides a simple and highly accurate technique (incorporating checksums) for capturing information – pre-printed barcodes are often applied to forms which will then be hand completed and returned, avoiding the need to index such documents when scanned. Barcode labels are often used in applications where documents are indexed prior to scanning, for example if indexing is carried out in-house but scanning is outsourced.

3.10.4 Optical mark recognition (OMR)

OMR is a technique for recognising pre-defined shapes in pre-determined positions – for example a tick or a cross in a box. It is used in applications such as the National Lottery, where a 100% accuracy rate is (hopefully) achieved through simple box marking.

Software verification can be used to detect errors (such as ticking too many boxes), or even to distinguish between a ticked box and one that has been scrubbed out.

It can also be used for signature detection – not to recognise the signature, but to check that ‘something’ has been entered in a signature box.

3.10.5 From OCR to ICR

Essentially, OCR is a template-matching technology. It knows what an “R” in 10 point Bookman Old Style looks like, for example, and applies that template to the character to be recognised.

If it matches, it knows it has an “R” in 10-point Bookman Old Style. If not, it tries another template until it finds a match. Such technology is relatively simple in today’s terms and gives high recognition rates and good performance against documents containing known typefaces. Note also, however, that an “R” in 12 point Bookman Old Style, or an emboldened, italicised “R” will defeat it, unless it is in its library of templates.

ICR, on the other hand, is rules based, and knows about the topology of letters and numerals. It understands that an “R” is made up of a vertical bar, a semi-circle on the top half of the vertical bar, and an angled leg at the bottom. It knows it can be different fonts and point sizes, it can be serif or sans serif, italicised, bold, or a poor quality scan. It can apply the rules for recognising an “R” to typewritten and handwritten text alike.

As such, ICR techniques can be used to recognise a much wider range of documents, and degrades less on poor quality or hand-written documents. When used with dictionary and context checking facilities, as it is in most forms processing applications, it gets closer to human-level recognition capabilities than OCR could ever do.

3.10.6 Output format

The output from any recognition process needs to be clearly defined. Does the application involved require a simple data stream, a text stream, or a formatted database record? Can the image on which recognition was performed be discarded once its content has been extracted and verified, or must it be retained for archive purposes? Must the original paper also be retained, for example if it contains a signature?

3.11 Colour and greyscale

So far, we have looked primarily at the application of bi-tonal monochrome digital capture techniques. They are highly efficient – a pixel is either black or white, and it can be represented in a single binary bit. Greyscale is more complex – each pixel can be rendered in up to 256 shades of grey (8 binary bits). Colour is more complex still – 24 binary bit or higher. The image is much larger, and much more difficult to compress efficiently.

Developments such as JPEG compression have helped overcome this limitation, and can now produce colour images occupying only marginally more storage space than their monochrome equivalent. (One reason for this is that such images can be scanned at lower resolutions – the eye ‘sees’ a photographic image as sharp even at resolutions as low as 75 dpi, whereas text at 200 dpi is clearly jagged).

Archivists should now be considering a strategy for both monochrome and colour content. Improvements in compression techniques, and reductions in the cost of storage and bandwidth, are rapidly closing the gap.

Colour can add content to a document – the clear representation of a chart, or a colour photograph, might be critical.

Colour can add context to a document – for example key phrases might be highlighted in yellow to aid understanding and to reinforce a message. Colour coding is also used in manual systems (e.g. “pink copy for Accounts”), and colour recognition can replicate that functionality

Colour can aid recognition. Character recognition software is being developed to exploit colour more fully, and appears to deliver higher accuracy than bi-tonal monochrome recognition.

The need for trade-off also needs to be appreciated. Consider a colour brochure containing a photograph of a new product, and its technical specification: a JPEG image will deliver the photograph clearly, but the text may be illegible. A TIFF monochrome image will render the text clearly, but the photograph may be indecipherable. Hence the

development of multi-layered compression techniques such as JMP, used by DjVU, which analyse the topology of the document, and apply different compression techniques according to whether a part of the page is colour or monochrome, text or picture, foreground or background, etc.

3.12 Microfilm, Microfiche

Microfilm and microfiche continue to play a significant role in document archives, especially for large archives requiring large storage capacities (terabytes and beyond). Some document scanners support dual capture – a digital image is taken for short-term use, for example data capture or workflow, and a microfilm image for long-term archive.

Some organisations moving from microfilm to digital archive have carried out a backfile conversion on their film archive, and many capture agencies specialise in such services. Others have developed strategies which involve on-demand conversion – as a microfilm archive is requested, it is retrieved and digitally transformed.

Many microfilm archives are referenced through an online database, perhaps even an automated retrieval system. The microfilm is limited to one physical place, but the index is widely available and a digital image of the film can be routed electronically to its requestor. It may be appropriate to preserve the investment in such a system, by implementing a records management strategy that accommodates hybrid records (for example electronic images, electronic objects, microfilm and paper – books and bound reports).

For records that need to be kept for a lifespan i.e. 75+ years bureaus are now offering a facility of writing digital images to microfilm on the basis that the medium will always be able to be read in the future.

3.13 Electronic objects

Paradoxically, documents created electronically may not be in a suitable format for electronic archive. Obsolete word processor formats are an obvious hurdle, but consider also the use of an autodate function; how valid is an archived document if it always prints and displays with today's date?

The archivist needs to consider storing such documents in a portable form. Conversion to a TIFF image is one option, though that will impact the ability to search on the full textual content of the document. (It is still however preferable to the approach of printing a document in order to then scan it). Rich text format or even ASCII text are other options with some guarantee of longevity, although such options may lose graphical and other non-textual content.

Adobe's Acrobat portable document format (PDF) offers another alternative. Some software packages allow separate renditions of a document to be held, for example a

WORD document as a master for authoring and a PDF rendition for circulation and comment.

The PDF format is particularly flexible because it can convert source documents generated in multiple formats. These include accurate renditions of common electronic formats (word processing, spreadsheet and presentation graphics); still image, moving image and sound, and hypertext links; and capture from paper, in either image, text or 'image + text' format.

The object might also be converted to HTML format for use over the web. Better still, XML/XSF offers the ability to completely separate presentation and content.

3.14 Forms

Forms processing is a rapidly growing market for many aspects of document management. OCR/ICR techniques, combined with sophisticated error verification procedures, have evolved to the stage where the level of accuracy achieved has enabled many previously labour-intensive key-to-disk operations to become fully automated.

The Internet has also helped make the concept of the paperless form a reality. Increasingly, consumers are filling in forms on their PC – and some on their palmtop PDA, WAP phone or digital TV – using forms delivered in spreadsheet, PDF, HTML or XML format.

XML in particular offers attractive possibilities for form completion. The entire capture process – data entry fields, dropdown completion lists and validation rules can be encapsulated in the XML form, allowing it to be downloaded and completed offline.

4. Information Capture Technologies and Methods

4.1 In-house or bureau?

One of the first considerations for any conversion exercise is whether to tackle the project in-house or to outsource it to a document capture agency. This applies to both an initial backfile conversion requirement, and to the ongoing capture operation thereafter. The need to match appropriate skills to the conversion exercise should not be underestimated; many organisations have simply assumed that their existing workforce can be assigned to the conversion exercise and it will take care of itself, therefore there is no need to consider an external agency. Unfortunately this is seldom true.

In-house staff are usually well-skilled, with a detailed knowledge of the business process and the ability to accurately categorize the work. But are they suited to the work? Document scanning and indexing can be a mundane task, and experienced staff can feel devalued if assigned to it. It is not unknown for turnover rates and sickness levels to rise markedly during such an exercise.

The organisation may decide to employ temporary or junior staff as part of their in-house strategy. One risk of this is that target levels of accuracy and quality may not be achieved, due to inexperience; another is that the high fliers among the group, who prove themselves to quickly develop the necessary expertise, may be just as quickly snapped up by other departments.

The external agency may also experience some of these issues, but a well-established agency is one that has learnt how to deal with this effectively, and the organisation can further safeguard itself against risk through a tightly defined service level agreement.

Pros	Cons
In-house	
Confidentiality Full control over information at all times Utilise existing staff skills and expertise Minimise disruption to 'business as usual'	Capital investment in scanning and IT equipment and software Obtaining, motivating and retaining suitable staff Learning curve Managing peaks and troughs – e.g. staff holidays
Outsource	
No capital investment No learning curve No recruitment or staffing issues It's their core business, not yours Economies of scale	Loss of control Less understanding of your business Disruption to 'business as usual' – for example taking documents off site for capture May be constrained to specific vendor solutions

Figure 5: In-house v. Outsourced Conversion

Regardless of the approach to be adopted, the operation itself will follow broadly similar lines:

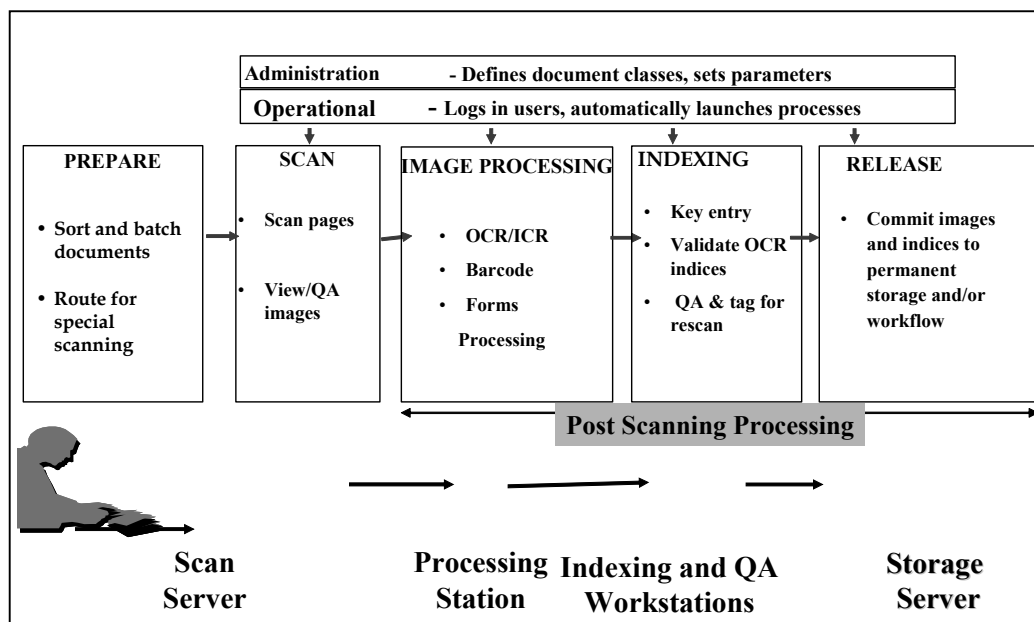


Figure 6: Conversion Process

One possibility that an organisation can consider is to reverse the above procedure; index then scan.

This can be particularly suitable where the indexing element of the operation is particularly complex and requires skilled staff. In this instance, documents can be indexed in-house and a barcode label attached to the indexed document; the documents are then sent to an agency for scanning. As the document is scanned, it is automatically cross-matched to its index entry via the barcode.

4.2 Traditional archives

The relationship between transaction, document and page can be highly complex. A transaction might consist of several sub-transactions, each represented by one or more documents; and each document might consist of many pages. Page size, colour, weight and shape may alter from document to document, or even within a document

In some organisations, for example, a person's change of job may be represented by a set of documents recording the termination of one role, and a set of documents recording the commencement of a new role. Failure to treat the entire document set as a single transaction could be critical; it might affect that person's pension rights. Often such subtlety is managed easily within a paper handling system – perhaps by stapling the

documents together rather than paperclipping them - and it is important that the electronic archive can represent this metadata accurately and consistently.

The nature of the paper to be processed will play a large part in the determination of the optimum capture technology and method of capture.

If original documents are of a consistent size and quality – for example A4 single page, single sided, good quality originals – then the capture process is relatively simple. Originals can be loaded into the automatic document feed (ADF) of the scanner for rapid capture, and then routed for quality control and indexing.

If the originals are of consistent size and good quality, but the number of pages can vary, then the operation is slightly more complex. This can be addressed either by inserting separator pages between each document, or by routing the batch to the next stage (quality control or indexing) for electronic separation into distinct documents. In addition to a scanner with ADF capability, there may be a need for separator recognition technology.

For complex documents, which may be stapled or clipped and comprised of pages of different sizes and colour, a more demanding regime is required. There may need to be a separate stage for document preparation (removal of staples etc), and each page of the document may require hand feeding, making an ADF superfluous. Adaptive thresholding could also become a critical requirement in order to accommodate multiple page types.

Alternatively, documents may be separated into their component page types at preparation stage, and batches of similar pages are then fed through the scanner ADF, applying pre-determined scanner settings that have been optimised for each page type. After scanning, the separate pages must be electronically recombined prior to or during indexing.

A third approach is to index the document before scanning, generating barcode labels which are then affixed to each page of the document. It can then be separated into its component pages for scanning, since the barcode will ensure that the pages are then recombined electronically.

There is no ‘best way’ of approaching this; it depends upon the nature of the operation, and may even depend on whether the paper original documents can be discarded or must be reconstituted after capture.

Another issue to consider, is whether a document should be indexed from the paper original or from the captured image.

Indexing from the image provides a quality check, since if the document is unreadable then the indexer will not be able to index from it. On the other hand, indexing from the

original paper provides a batch control check, for example where two pages have been pulled through the scanner simultaneously.

4.3 COLD/ERM Enterprise Report Management

Computer Output to Laser Disk (COLD) was the acronym originally assigned to the process of capturing computer-generated print spool files for long term archive. As the dependence on optical disk diminished, the acronym was superseded by COOL (Computer Output On Line) and finally by the more generic term of Enterprise Report Management - ERM

This renaming exercise reflects a growing recognition of the true nature of this form of archive. Initially it provided a cost effective alternative to the storage of report archive on mainframe disk. As storage costs fell, and organisations migrated increasingly to client server systems, this particular need reduced: organisations could just as easily add more storage to their line of business systems as purchase a separate COLD system. But at the same time, however, system needs were growing more complex. As gas utilities supplied electricity to their customers while electric utilities supplied gas to their customers, customer bills became increasingly complex, often representing the output from several different systems, which only ever came together at bill production time. While in theory the information can be reconstituted at a later date (for example if the customer contacts the call centre with a bill query), in practise it is very difficult to synchronise this historical data and the only way of ensuring accuracy is to retain a record of the actual bill – hence ERM.

Indeed, the entire topic of presentation has become a key competitive differentiator in its own right.

ERM systems were initially developed to handle simple print streams. The development of APA – All Points Addressable print structures has emerged as an area that ERM products must support if they are to address the large corporate market place.

APA provides the technology to personalise documents. These could be bills or statements, incorporating, for example, targeted special offers. Clearly, if helpdesk staff are able to view the documents as the customer sees them, they can give a personalised service and capitalise on cross-selling opportunities when the customer calls in. IBM's AFP format and Xerox Metacode are the two most common APA formats.

A key fundamental requirement, particularly for a multinational organisation, is to present a bill to its customer in the right language and on the correct stationery, to be inserted into the correct envelope and correctly sorted for posting. This can be achieved through sophisticated processes which match the stationery to the bill, by reading barcodes on the stationery itself (these are printed on the page edge, which is guillotined during the print process).

A second requirement is presentation quality. Few customers now find plain text delivered through a dot matrix print as an acceptable level of service, so high quality formatted laser printing is essential – requiring COLD vendors to upgrade their products beyond support for ASCII text formats alone. The archive system must be able to present an exact facsimile of the document that the customer received. Timing is also critical for this, if the customer bill were presented on 31st December in francs then it may not be much use to reproduce a facsimile in euros when the customer queries it on 11th January!

A third requirement is one-to-one marketing; the bill is perhaps the only regular communication between supplier and customer, and provides a unique opportunity for developing the customer relationship. This might include for example, printing discount vouchers for a customer who has satisfied certain criteria; it may not include re-printing those vouchers if the customer later requests another copy of the bill...

A fourth requirement is presentation medium. Customers do not necessarily want to receive a paper copy of a bill, and some organisations offer discounts to encourage customers to receive bills in electronic form (via email or the world wide web). Egg, for example, sends an email with a hypertext link to the bill, which is available through its secure website in HTML format; First Direct delivers its customers bank statements in HTML format, but with the option to download in Excel or Quicken format; One-Tel presents its bills as PDF documents.

The ERM system may simply store an exact facsimile of the report, representing both content and presentation in a single source. Alternatively, the report ‘backdrop’ may be held separately, and combined with the report content as an overlay during display and printing. In this case, it may be necessary to apply controls to ensure that the correct overlay can be assigned whenever required.

4.4 Forms capture

Forms capture is a particular subset of the typical archive process. The prime objective of the forms capture process is to extract data from the document that can be used in subsequent processes, and therefore the prime consideration is the accuracy of the recognition techniques.

The QA process would tend therefore to focus on the accuracy of the data extracted rather than the quality of the image itself, but the technologies involved are broadly the same.

Forms recognition is essential to the process; many vendors are able to offer solutions which will scan a mix of different form types, recognising the form upon scanning and automatically routing it to the appropriate character recognition process, eliminating the need to pre-sort forms. Solutions are also often capable of recognising forms fed ‘footer first’ rather than ‘header first’, and will electronically rotate the form before routing. Optimisation techniques such as contrast adjustment, despeckle (remove noise), deskew

(straighten the page if fed at an angle) and form dropout (electronically remove the form background, leaving only the content) can also be carried out to ensure that the best possible quality is delivered to the recognition engine.

The recognition process itself will be optimised to the form. If the process requires a customer reference to be captured, then OCR/ICR techniques will be applied at the precise location on the form where the reference number is expected (zonal recognition). The recognition technique itself can be focused to that field – for example, if it cannot distinguish between ‘B’ and ‘8’ but the field should contain numeric digits only, it can ‘vote’ accordingly. And context-based verification can also be applied, for example check digit verification.

Human intervention may still be required, but kept to a minimum. If the recognition process cannot resolve a query, then the relevant portion of the original image and the interpreted characters for that component can be presented to an operator for correction – not the entire image.

Further techniques can be applied for QA purposes. For example, the system can present to an operator a list of every character it has interpreted, but grouped by character – ‘a’, ‘A’, ‘b’, ‘B’, etc. If the operator detects a ‘4’ among the ‘A’s, selecting the character will display the form from which it was obtained and highlight the offending digit.

One drawback in forms processing is that the application must be ‘trained’ to recognise each form. In the case of a large organisation dealing with hundreds of suppliers, each presenting invoices in different formats, this is a significant administrative overhead.

The latest developments in recognition software are now addressing this; for example software can interpret an invoice however it might be laid out,.

Others are exploiting the training aspect further; an application that can be taught the layout of a form, can also be used to generate a form. Some applications can also reproduce the form in different formats, for example on paper, in PDF or in HTML format for use over the web.

These new developments in forms processing called Intelligent Document Recognition – IDR provide automation of two key areas. First the ability to recognise the form type by its layout will allow the document to be automatically indexed. It will recognise the layout as an invoice and will look for the purchaser’s address and other metadata and automatically store it, adding the metadata to the record in the customer’s folder. The other use is having recognised the form it will be automatically routed to relevant business unit through the workflow system.

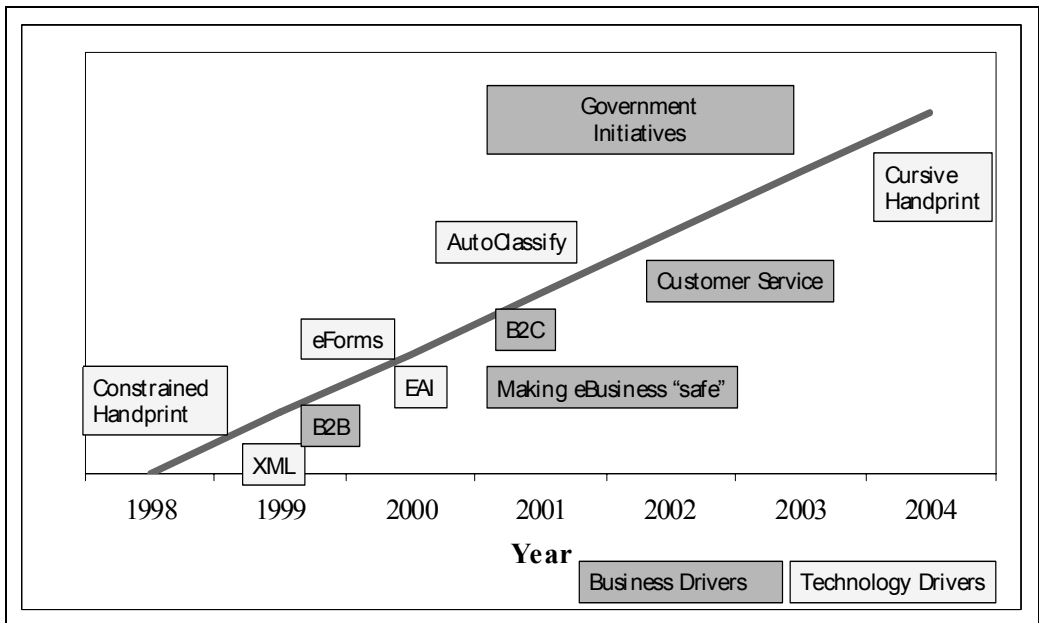


Figure 7: Drivers for Intelligent Document Recognition

Source: Document Capture for eBusiness 2001-3. Pub. Strategy Partners

4.5 Electronic documents

In many circumstances, the capture of electronic documents, and the capture of an email message and/or its attachments, are relatively straightforward; but should not always be assumed to be so.

Most document management systems offer simple integration with office systems, for capture of word processing documents and similar. ODMA-compliant applications provide the necessary 'hooks' to enable their existing save and retrieve interfaces to be adapted so that for example, the user can choose to work with the local PC hard drive or with the document archive. The interface may simply present the operator with an option to 'fill in the blanks' in order to categorize a document, an email, or an email attachment, using the same rules that are applied for scanned originals.

Other options allow for documents to be auto-indexed based on content, alternatively individual users can route documents requiring archive to a central location, where a designated archivist will complete the task. This may include transformation of the document into a more portable format.

The same techniques can be applied to automate the process of capturing documents generated in line of business (LOB) systems. As the LOB system generates a document – a letter to a customer for example – the document and its context can be automatically routed to the archive. (This will usually require some programming and is dependent upon the API capabilities of both the LOB system and the archive system).

One particular challenge for such interfaces is the ‘mail merge’ document. Here, the LOB application generates a data stream which will be used to create a single report in a word processor – common for example in HR applications generating annual ‘salary review’ letters for all staff.

What needs to be determined is whether such a document in fact a report – more suited to an ERM application, which will treat it as a single entity – or is it in fact a collection of individual documents, each of which must be correctly captured and categorized against individual employee? Payslips for example, might be deemed to be documents which can be held in a COLD report, while salary review letters are records which must be stored individually.

If the latter, then does the archive system have the tools to do so, with the necessary controls and safeguards in place? One application with which the author was familiar, worked well at separating each page of the report into separate documents and archiving them by employee – until the year in which the salary review letter became two pages long.

4.6 Planning a Document Conversion

4.6.1 Step 1: Choose a Location

Should the conversion be carried out on site or off site? On site conversion ensures the documents remain close at hand should they be required during conversion, but requires that appropriate resource is available; space for equipment, preparation areas, perhaps upgrades to network bandwidth. If a bureau is providing the service, then there will be additional costs for installing their equipment on site, and accommodation and travel costs for their personnel.

Off site conversion can be easier to manage, but does involve the unavailability of documents during the conversion. For some organisations, this is an advantage, especially for ongoing conversion. Documents can be delivered direct to a capture bureau (using a PO Box No or similar), scanned in the morning of delivery, and the scanned images routed back to the organisation electronically for processing the same afternoon. Paper handling costs are eliminated.

4.6.2 Step 2: Determine Tactics

There are three basic options.

- Backfile conversion – a full conversion of all existing documents, either prior to or in parallel with the conversion of ongoing work. Some organisations will outsource the backfile conversion to a bureau but retain the ongoing conversion in-house.
- Incremental conversion – a gradual conversion of existing documents in parallel with the conversion of ongoing work. This can be an on-demand approach - where existing documents are scanned as and when a document folder needs to be retrieved for other reasons - or a phased approach based on document ‘value’.
- From this day forward – conversion of ongoing work only. This works well where documents are retained for only a short lifetime.

4.6.3 Step 3: Determine What to Scan

Does every document in the paper archive need to be scanned? Or does it contain duplicates, out-of-date records, superfluous or irrelevant information? A purge exercise incorporated into the conversion project can be highly cost effective.

Do the individual documents in a file folder need to be separately indexed, or will it suffice to index at folder level? An indexing strategy that relates the effort of indexing to the value of the document can also be cost effective.

How long must documents be retained, and can those retention periods be set automatically during capture? Which documents can be shredded after scanning, which must be retained regardless?

Who requires access to the document? If the archive is to be opened for public access at a future date, are the necessary security controls incorporated?

What strategy should you consider to detect and remove duplicate documents?

4.6.4 Step 4: Select Capture Software

Many document management systems operate in conjunction with a specialist front-end capture system: the features such systems provide to reduce manual intervention soon justifies any additional investment. These include edge detection, double feed detection, image cleanup and optimisation, line removal, deskew and despeckle, are all designed to eliminate the need to rescan a document (or batch) and speed up subsequent indexing or recognition processes.

Quality control systems also save considerable time – verifying image quality, managing the repair process if images do need rescanning, batch control procedures, and generating management statistics (throughput, volumes, performance against target) that would otherwise be collated manually.

4.6.5 Step 5: Select a Scanner

The correct choice of document scanner is critical to a successful operation, and needs to be carefully sized. A scanner able to handle the high volumes of a backfile may be overkill for ongoing work, while a scanner sized for ongoing work is simply not up to the job of managing a backfile.

- What volumes need to be fed – daily average and peak? Are documents suited to an automatic feed (ADF), or would manual feed be more appropriate? Is there a need for flatbed scanning?
- What kind of paper needs to be processed – A3, A4, simplex, duplex, monochrome, colour?
- What is the paper quality – weight, condition etc? Is it prone to double feed? What resolution(s) need to be supported?
- Is one scanner sufficient? Might it be more appropriate for example to deploy a fast ADF scanner for high volumes ‘fast track’ work and a slower flatbed scanner for ‘slow track’ and ‘repair’ work?
- Is the scanner matched to the capture software – or does it provide features that could be duplicated in software, for example barcode recognition?
- What happens if the scanner breaks down – how quickly can it be repaired or replaced?

A recommended approach is to ‘try before you buy’. Scanners with identical specifications do not perform identically, and a trial feed of ‘live’ documents through a set of scanners will always prove worthwhile. Scanner distributors often provide a laboratory environment for this purpose, demonstrating a range of scanners from different vendors.

There are also specialist scanners. Examples are book scanners that allow scanning the page without breaking the spine, X-Ray and microfilm/microfiche scanners. A more limited selection and unless there is a high volume requirement most organisations contract out the capture to a service bureau.

4.6.6 Step 6: Define Preparation Procedures

A backfile exercise can involve huge volumes of paper, and a strategy is needed for managing it effectively.

- How will work be scanned – a filing cabinet at a time, by customer surname, by department?
- How will the paper be ‘checked out’ for scanning, and an audit trail maintained until its disposition?
- What procedures will be established to prevent a file from being scanned twice, or not at all?

4.6.7 Step 7 – Size the System

The conversion will generate large volumes of data – which may be uploaded directly onto the archive system, or supplied on CDs for a batch upload.

- How will the new system manage that data – transfer to optical jukebox, transfer the data to RAID, insert the CDs in a CD jukebox?
- Has sufficient capacity been allowed?
- What impact will that traffic have on the network, both during upload and then for subsequent retrieval?
- If a jukebox solution is being utilised, what are the implications for retrieval if the documents required to satisfy a search request are scattered across many disks?

4.6.8 Step 8: Select and Train Staff

The various roles in the process need to be evaluated – scanner, indexer, verifier, administrator – and appropriate skill sets for each role defined.

- What training is required?
- will roles be rotated?
- Will roles be rotated within the conversion team, or throughout the department?
- Will there be a specific career path?
- What contingencies will be made for holidays, sickness and leavers?
- What measures and feedback mechanisms will be deployed to maximise productivity?

5. Standards for Formats

Not so many years ago, the document management market was simple to describe. Documents came in only one flavour, i.e. scanned images; they were processed by only one type of application, known as document image processing (DIP); images were either black on white, or white on black; compression options were ITU Group 3 and 4 (formerly CCITT Group 3, CCITT Group 4), or none; if the original document was in poor condition, then the user was obliged to scan a photocopy of it; colour and greyscale were interesting concepts that might happen one day.

Times have moved on. Digital scanners and the charge couple devices (CCD) which drive them are now more powerful and sophisticated, (to the extent that digital cameras now rival high quality film cameras), and document scanners can also act as photocopiers; compression algorithms achieve ratios previously undreamed of; and storage and bandwidth capacities have grown exponentially. The pipedreams of 1992 are a reality in 2002.

One issue to be aware of with colour documents is ‘lossy’ compression, one example of which is the fractal compression technique. A famous instance of this involved the picture of a crop-duster flying over a cornfield; in the compressed image, the cornfield remained but the crop-duster disappeared! And all standards involve some compromise, simply because the output is digital rather than analogue.

Other techniques involve zonal compression, where different techniques are applied to different elements of an image (photographs, text, diagrams etc) for optimum results. These techniques have tended to remain in the research lab rather than become mass market, although the DjVu format is one example of seeking to break that mould.

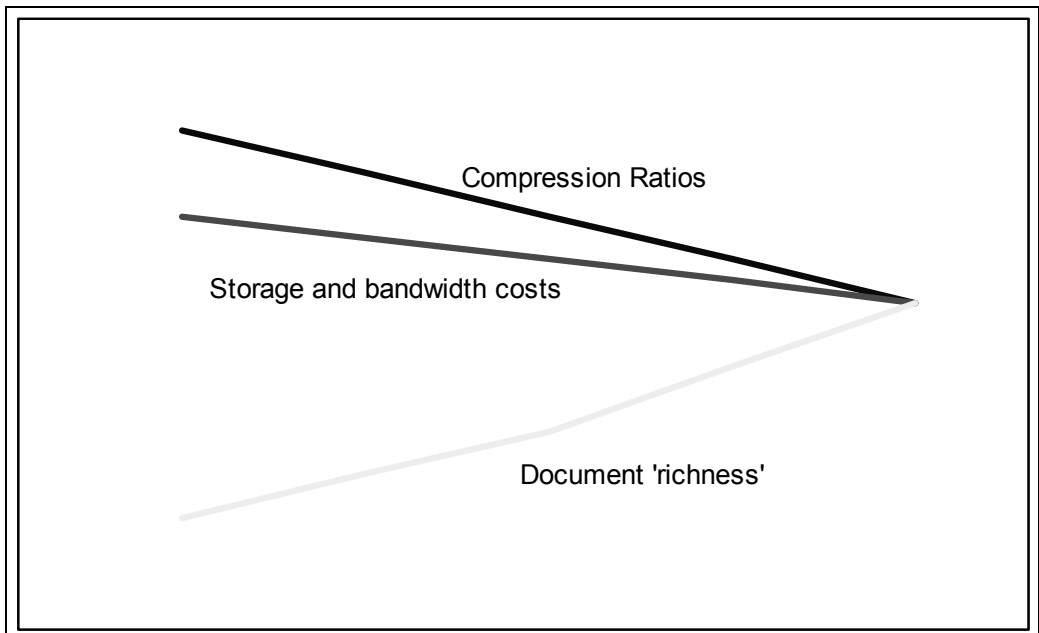


Figure 8: Technology Convergence

5.1 Tagged Image File Format (TIFF)

The TIFF standard was originally defined in 1986 by a group of seven vendors, including Hewlett Packard and Microsoft, and is now owned by Adobe. Most mainstream vendors support it, for example a default TIFF viewer is bundled with all versions of the Microsoft Windows operating system.

The key attributes of the TIFF standard are extensibility - new image types can be incorporated without invalidating existing supported formats - and portability – full hardware and software independence.

A TIFF image object consists of a header part and image data. The header contains ‘tags’ (hence ‘tagged image’) which define such attributes as resolution, height and width, compression type and similar. It also allows vendors to include proprietary information – keywords and similar metadata - or mechanisms for detecting whether an image has been altered subsequent to capture (for example by holding a count of the total number of pixels in the original scan). Generally speaking, proprietary tags introduced by one vendor can simply be ignored by other vendors and the image is still ‘valid’ – the only exception being where different vendors utilise the same portion of the header for different purposes.

In the early days of the standard, different interpretations of the tags were only too obvious; a TIFF image created on one system might display as a negative on another system, for example, since one developer chose to represent black by setting a bit ‘on’

and the other chose to set the bit ‘off’ instead. But such discrepancies are now largely eliminated.

TIFF supports a number of compression modes; some software applications may automatically try different compression techniques for each image and then retain the version which achieves greatest efficiency.

Compression Type	Application
Uncompressed	No compression – may be required if the image fails to compress well, for example a poor quality photocopy
Group 3 (1d)	Usually applied for documents designed for fax transmission, for example integration of a fax gateway with the image archive
Group 3 (1d) Modified Huffman	Also designed for fax transmission, essentially a hybrid of 1d and 2d techniques
Group 4 (2d)	The most common compression technique used for scanning of monochrome ‘business’ documents, achieving the best compression ratio for non-dithered monochrome images
Packed bits	Used for compressing black and white image files
LZW	For colour and greyscale images, LZW is a lossless technique achieving lower compression ratios than JPEG
JPEG	For true colour and 256-bit greyscale images, JPEG is a lossy technique which achieves good compression ratios

Figure 9: TIFF Compression Formats

Note that the TIFF standard does not directly support the use of annotations. Such annotations must either be ‘burnt in’ to an image - which creates a separate version of the image, with version control and audit trail implications – or held as a separate object, which can be overlaid onto the image at display or print time. The overlay approach is seldom portable and annotations may be lost if TIFF images are migrated from one system to another.

At the moment however, the TIFF standard represents the most reliable option for the long-term storage of monochrome ‘business’ documents. It is universally accepted as the de facto standard; the vast majority of PCs sold today are pre-bundled with software for viewing TIFF images; the vast majority of business-oriented scanning software applications support the creation of TIFF images; the standard is portable; and it is extensible.

But it is insufficient to simply stipulate ‘TIFF’ – some attention needs to be paid to the methods of tagging that will be applied, and in particular to the compression technique or techniques applied to the image data itself. If there is a need for documents to be annotated, then this must also be carefully considered.

5.2 Joint Photographic Experts Group (JPEG)

The JPEG standard was developed by a group of experts nominated by national standards bodies and major companies. The committee, known as ISO/IEC JTC1 SC29 Working Group 1, was charged with the development of standards for continuous tone image coding. The 'joint' refers to the involvement of both ISO and ITU-T (formerly CCITT).

One particular standard produced was IS 10918-1 (ITU-T T.81), aimed at still image compression. A version of this standard was introduced to the public domain and has become the widely adopted standard generally known as JPEG.

The JPEG standard is not only used extensively in the commercial sector; millions of home users make extensive use of the format when emailing photographs (captured on a digital camera or on a SOHO colour scanner) to friends and relatives, or when building personal websites. JPEG viewers are also widely available, being bundled with the most popular web browsers; adopted as default by most digital camera vendors; and incorporated in a plethora of image manipulation packages.

The respectability of the origin of the JPEG standard, its portability and the universality of its adoption, make JPEG one of the safest choices of standard for the long-term archive of still colour images such as photographs.

Note however that not all vendors implement the standard in a consistent way; a JPEG image created in one application may become unrecognisable to that application if it has been modified elsewhere (for example, if rotated from landscape to portrait mode by a separate application).

All the major scanner makers are now offering a broadening range of colour scanners capable of handling business volumes. There are costs and performance issues- with colour as it places a greater demands on storage capacity and network bandwidth than its monochrome equivalent

JPEG2000 is set to replace JPEG as the colour compression standard of choice and addresses some of the performance issues of JPEG.

5.3 Graphics Interchange Format (GIF)

GIF and JPEG represent the two most common file formats used for graphic images on the Internet.

GIF uses the Lempel-Ziv-Welch (LZW) compression algorithm, owned and licensed by Unisys. It is a lossless, two dimensional compression technique, which for larger documents can represent a significant storage size overhead. One version of the format

(GIF89a) supports short animation and also allows interlacing (useful for including an image as a backdrop in a web page with overlaid text).

A patent-free replacement for GIF has also been developed, the portable network graphics (PNG) format.

5.4 Moving Pictures Expert Group (MPEG)

The standard was also developed by the ISO/IEC JTC1 SC29 committee (Working Group 11 developed the MPEG standard while Working Group 1 developed the JPEG standard, but the groups have little commonality).

The standard is aimed at digital video and audio compression. It allows different vendors to apply proprietary compression algorithms within the same standard

MPEG is sometimes compared with H.261, a teleconferencing application standard. In both cases, there are elements of proprietariness which the intended user should explore carefully before establishing a strategy for the long term archival of moving image.

5.5 Audio Video Interleave (AVI)

AVI is defined by Microsoft, and an AVI viewer is pre-bundled with Microsoft Windows. This in itself has helped to make AVI a de facto standard for audio/video data. It is commonly used for short animations and video clips – for example the animated display of a file being copied from one location to another, or crunched into a recycle bin.

5.6 Adobe Acrobat Portable Document Format (PDF)

The PDF standard has also become ubiquitous, due in large part to the widespread availability of the Acrobat viewer and its widespread uptake on the web. Like TIFF, it is a portable and extensible standard, and is widely recognised as an open and de facto standard for electronic document distribution worldwide.

It preserves the fonts, formatting, graphics and colour of a source document, regardless of the application and platform on which the document was created, and is able to reproduce an accurate rendition of the original for display or print. Adobe has also opened up the PDF format to third party developers.

Monochrome TIFF and colour LZW-compressed images can also be converted to PDF format, with options of preserving the image, converting the image to text (OCR), or an ‘image + text’ option. This latter option has proved popular in knowledge management applications because it retains the image – complete with graphics and signatures, etc – and also the textual content for use in full text retrieval searches.

5.7 Rich Text Format (RTF)

RTF is a method of encoding formatted text and graphics for portability between applications. It can be used with different output devices, operating environments and operating systems.

RTF uses the American National Standards Institute (ANSI), PC-8, Macintosh, or IBM PC character sets to control the representation and formatting of a document, both on screen and in print. It allows documents created under different operating systems and with different software applications to be transferred between those operating systems and applications.

5.8 HyperText Markup Language (HTML)

HTML was defined by the ISO/IEC JTC 1 committee, and is an application of ISO 8879 – Standard Generalised Markup Language (SGML).

HTML is used extensively in web-based applications and provides a portable framework for the interchange of documents independent of the application which created them.

5.9 Extensible Markup Language (XML)

XML was designed ‘as an extremely simple subset of SGML, designed for ease of implementation and for interoperability with SGML and HTML’. It is primarily intended to ‘meet the requirements of large-scale Web content providers for industry-specific markup, vendor-neutral data exchange, media-independent publishing, one-to-one marketing, workflow management in collaborative authoring environments, and the processing of Web documents by intelligent clients’ (W3C XML).

Whereas HTML describes the structure and display characteristics of a document, XML describes the content and structure of the data contained within the document.

XML is platform independent, and its simplicity provides a low cost of entry for participating in data exchange applications, compared with alternatives such as Electronic Data Interchange (EDI).

XML supports the creation of document type definitions (DTDs). Trading partners who adopt the same DTDs, whether as a small group or industry-wide, can trade data seamlessly regardless of the applications each of them uses to generate and process that data. Companies can also adopt XML as a mechanism for exchanging documents and data between disparate applications even within their own organisation.

6. Best Practice Applications

6.1 Forestry Geographic Information System (FOGIS) Project of the Department of Forestry, Baden-Württemberg

Managing natural resources - balancing sometimes divergent environmental and economic interests - is a major challenge for many governmental organisations around the world.

The Department of Forestry in Baden-Württemberg, a state in the southwest corner of Germany, is responsible for the cultivation and management of the state's forestland, including the famous Black Forest.

The department oversees four administrative regions, which include 190 forestry offices and more than 1,400 forestry districts, as well as a Forestry Research and Development Institute. It must comply with both state and federal regulations and guidelines.

The department's alphanumeric database contains massive amounts of administrative, ecological, inventory and statistical data. Most of this data is spatially oriented and can be effectively updated and used only through map visualisation.

But traditional paper maps and simple computer-based displays could no longer keep up with user demands for rapid, flexible and accurate access to up-to-date forestry information.

6.1.1 Objectives

The department also needed a means to share information among its four administrative districts and the regional environmental information system.

Faced with the need for timely and accurate forest maps using spatial data processing, and a means to communicate that data, the department decided to create a next-generation GIS (geographic information system). This system would be based on high-performance hardware and both custom-developed and off-the-shelf software.

The objectives were to:

- Cultivate and manage 3.34 million acres of public, private and corporate forest land
- Produce timely and accurate forestry maps using spatial data processing, replacing analogue cartographic systems
- Continually update spatially related data via rapid database access and fast, flexible processing of integrated graphic and alphanumeric information.

6.1.2 The technical solution

The solution was to create a geographic information system (GIS) for the forestry. The hardware on which the system is based is composed of HP workstations, HP personal computers, and HP ENVIZEX X terminals. The application itself is based on both internally developed software and ESRI's ARC/INFO GIS software, and the application is supported on an Oracle relational database.

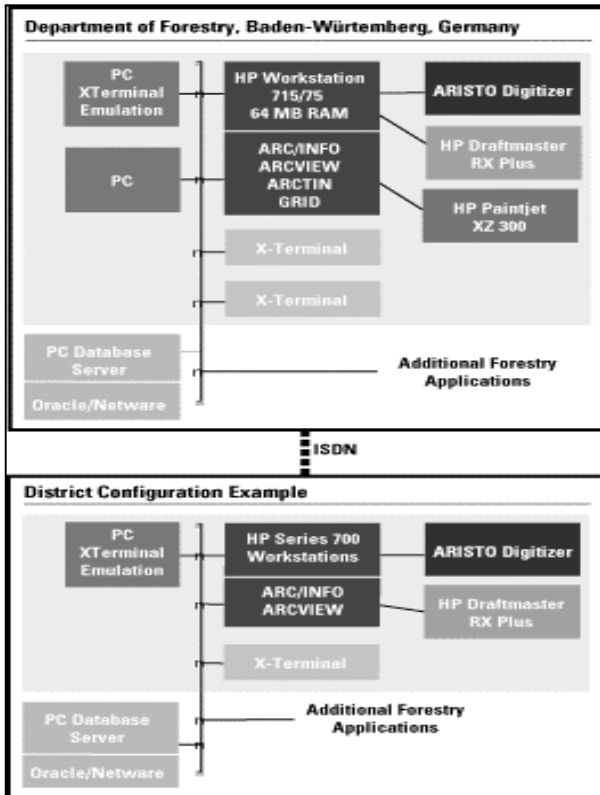


Figure 10: FOGIS System

The department created a three-level forestry GIS management system, known as FOGIS:

- The operations level consists of non-graphical data processing for inventory management and the procedures for standard map production.
- The information level supports spatial and alphanumeric queries and performs statistical calculations on the entire database.
- The planning level selects and aggregates spatial and non-spatial data for simulation and analysis of forestry operations and control procedures.

In selecting hardware and GIS software, the forestry department wanted a platform that could support the conceptual data model it was developing, capture data efficiently, maintain data consistency, and perform the required cartographic, thematic and analytical processing and output.

For GIS, the forestry department required powerful processing capability, and HP 9000 UNIX workstations were selected for this purpose. The department also decided to standardise on the HP platform for simplified and more economical system maintenance.

ARC/INFO, from ESRI was selected as the software platform for several reasons, including its ability to handle complex cartographic and analytical data processing, and support both the developing data model and data management consistency state-wide. It also provided a development platform for internally engineered applications.

The system links four regional administrative offices, Stuttgart, Karlsruhe, Freiburg and Tubingen, with the research and maintenance centre of the Baden-Württemberg Ministry of Lands and Forests.

Each regional office utilises HP 9000 Model 735 workstations with 64MB of RAM driving HP ENVIZEX X terminals, and running ESRI's ARC/INFO and ARC/VIEW software. The workstations are also connected to HP Vectra personal computers running an X terminal emulation front-end. Peripherals include HP DesignJet and HP Draftmaster RX Plus plotters and a digitizer.

At the research and maintenance centre, an HP 9000 Model 715/75 workstation with 64MB of RAM drives two X terminals and is connected to two personal computers. A PC-based database server using the Oracle RDBMS is also in each installation. Additional forestry applications can also be accessed. The five installations are connected via an ISDN network.

The system is being implemented in stages according to department priorities. The objective of the first stage is data capture and management for the production of forestry planning maps and the inventory of regional forest areas.

6.1.3 Benefits and ROI

The department has identified three main categories of benefits based on the early stages of system implementation, according to Andreas Hohne, the project manager:

"First of all, a productivity and quality increase is to be expected after switching to digital mapping. There are several reasons for this, including our ability to use digital survey data for base information, rapidly update and easily produce maps, make maps in different scales and formats, and integrate area computations and area inventory."

"Second, combining geographic data with forestry data leads to a product enhancement, because the ability to analyse expensive-to-obtain forestry data using spatial query, overlay, intersection and display tasks is now possible."

"Third, we have prepared the way for the future by adapting ARC/INFO to our three-level system concept to effectively share spatial data among users, applications and departments in an open systems environment."

"The benefits from the first two categories are already sufficient to justify investment in GIS technology. Benefits expected from the third category will be a decisive factor in determining the size of our future investment."

6.1.4 Adaptability

GIS systems are representative of archival needs that can be addressed through vector digitisation techniques. While such applications may not seem typical to many organisations, they do illustrate a particular genre of document conversion and powerfully illustrate the opportunities that can be created when documents are transformed onto digital media.

The UK Ordnance Survey, for example, is a €160M organisation producing digitised maps for business, leisure, administrative and educational purposes. OS data is used in almost every area of public life, ranging from collating the 2001 Census, through identifying suitable 'brown field' sites for new urban housing development, to helping the police detect crime patterns. The same digital information can be re-purposed to suit a myriad of different needs.

6.2 Document Imaging System Project of Sanctuary Housing Association

Sanctuary Housing Association is the fifth largest housing association in England, with a property base consisting of over 25,000 properties. Sanctuary have 35 offices spread all around the country, and in 1996/1997 their turnover was £80.8 million. The challenge they were facing was that over the last five years, government grants for building housing has decreased as a percentage of each property built. This necessitated that Sanctuary must seek to find more efficient ways of operating their business.

The Association is a 'not for profit' organisation and must therefore be extremely careful in how it spends its money, how it is seen to spend that money, and to make sure that any investment is fully justified.

Their housing stock is growing each year by 10% therefore the process of managing the properties is generating increasing volumes of paper. By the end of 1995 they realised that the volumes of paper were becoming unmanageable and that using a more modern method to handle the paper was required.

6.2.1 Issues and Requirements

An initial investigation of their processing requirements showed that the area with the highest profile was in the storage and retrieval of committee papers. The association is run slightly differently from a typical company in that most of the decisions and policy making is carried out through voluntary committees. These committees – in particular Central Council, Executive and Finance - meet several times a year and the agendas for each meeting are typically between 100-150 pages in length. The process of retrieving past agendas therefore involved a laborious task of sorting quite literally through hundreds of pages.

The second important area was the storage space taken up by the 2,000 paid invoices received each week. All invoices are sent to their head office in Hertford for processing. If there were any inquiries received from a local office, then these would have to be passed to the head office for processing. These invoices were stored in Lever Arch files initially but growth in volume meant they were running out of space. A migration of the files to microfilm storage media reduced the space required but did not help with the time it took to retrieve a document.

In addition they also had a picture library containing many thousands of colour photographs, stored in photograph albums. These pictures were used for internal publications and for public relations. There was a concern that if for any reason they were destroyed then their replacement would be difficult to achieve and enormously expensive.

Careful consideration was given to all the issues by fully analysing their requirements. After carefully considering the options they decided they would require a document imaging system. It would require the capability to handle colour as well as black and white images. It would also have to be able to index documents and be flexible enough to suit the radically different indexing requirements for each archive.

6.2.2 Technical Solution

In May 1996 they selected a supplier who could firstly, provide the necessary software to meet their requirements; and secondly, provide a managed service to handle the bulk capture of paid invoices and committee agenda papers. HMSL were chosen as supplier as they met all the requirements including financial viability, which was established through an excellent financial rating from Dun & Bradstreet.

The system was developed based upon HMSL's FileFlo software. Using this software, Sanctuary can now retrieve invoices by supplier reference, name or invoice amount, which has greatly improved the speed and ease of retrieval for any invoice.

In terms of backfile migration, the existing invoice archive was simply left on microfilm. Over the past two years, the microfilm system has fallen largely into disuse

since there are virtually no retrieval requests on invoices that old, bearing out the soundness of the initial strategy on backfiling. "When we do retrieve a microfilmed invoice we get a reminder of the dubious quality we used to put up with," commented Peter Stowe of Sanctuary.

Each month's invoices are scanned and supplied incrementally on CD. At the end of the year all the indexed information is put onto a single CD to provide a complete record.

The search for previous minutes has also been greatly improved as they can now retrieve them using keywords in addition to committee name, data and agenda item. When the minutes and agendas were stored as paper records it could take hours to retrieve the relevant document, now it takes seconds.

The internal publications department can now access the complete photo library electronically and they have for the first time a back up of the thousands of photographs on CD. Additional photographs can also be added by using a colour scanner.

6.2.3 The next stage

Once the installation of their national network has been completed, all enquiries to local offices will be dealt with using the local staff rather than by transferring such enquiries to the head office for action.

Another future consideration will be to install a CD jukebox to allow automated access to all the information. This will be essential when the local offices get connected as they will require remote access to the data.

It is also planned to use the system for the capture and storage of Human Resource files.

6.3 COLD/ERM – Solution Project of Staffordshire County Council

In 1995, Staffordshire County Council began to search for a more appropriate and efficient method for the production of financial reports - in an attempt to reduce the time taken to process the large quantities of data produced by the Education Department each month in respect of Local Financial Management of Schools.

After careful evaluation of the requirement, Staffordshire focused on an initial objective of storing financial data, which was at that time generated via mainframe computer resulting in large volumes of computer printout, requiring a significant amount of effort to burst, sort and distribute.

6.3.1 Technical solution

The Council decided upon a solution based on COLD/ERM (Computer Output to Laser Disc/Enterprise Report Management).



Figure 11: Paper-based bureau

Staffordshire County Council selected Hitec's DataStore™ product for this purpose. Prior to the implementation of DataStore, production of the financial reports was extremely labour intensive, amounting to one week of manual effort per month on a 14 months per year basis.

This manual process involved printing, sorting, collating and sending the relevant reports to each of the 540 pre-LGR (430 post-LGR) Primary, Special and High Schools within the County. The Education Department had to perform this manual process twice - once for the copy sent to the school and then a second time for the centrally filed office copy. Each School could receive up to 30 different types of reports – resulting in over 400,000 sheets of paper in a year!

The implementation of DataStore into the Education Department at Staffordshire County Council alleviated many of the problems associated with their previous paper system.

All relevant financial reports are now automatically indexed by report type, school, period and year and then archived electronically on-line. Where the department previously held three years worth of financial data in paper copy, one optical disc holds ten years worth of data. DataStore has enabled the Education Department to reduce process timescales to hours rather than days. "We are playing a supportive role," comments Dave Cheeseman, Principal Education Officer, Management and Financial Services.

"DataStore enables the schools to concentrate on teaching and learning and takes away many of the more mundane financial administration aspects."

In 1999 a further initiative was identified by Staffordshire County Council in line with Government targets as set out in Connecting the Learning Society, National Grid for Learning - The Government Consultation Paper. "From 2002, general administrative communications to schools and further and higher education bodies, and the collection of data from schools, should largely cease to be paper-based (the benefits would consist of environmental gains through diminished reliance on paper copies, more rapid transfer of information, and savings in postage costs)."

Using the same technical infrastructure installed as part of the NGfL, the monthly financial statements will be sent out electronically. The initial Pilot system, installed in ten High Schools, will roll out to the other 41 High Schools and all Schools within 18 months.

Staffordshire County Council and Hitec jointly began to investigate the best method for providing the reports electronically for each of the schools. The first option was for each of the schools to be connected to the DataStore system and be able to perform searches against the current financial data.

However, due to both the cost involved in a 400 + concurrent user license and the fact that all schools are not yet connected to the Staffordshire County Council network, this was not a viable option. Other options Staffordshire County Council considered included Word Documents and viewing the electronic reports through a Web Browser. However, both options proved limiting. After much analysis, Staffordshire County Council finally opted for a hybrid DataStore for Windows and Microsoft Excel solution - it was the only solution to offer the full functionality required.

Through the use of Hitec's API (Application Program Interface) the DataStore solution will produce the school reports electronically - using an Excel Spreadsheet as the viewing mechanism. Each school will be processed sequentially, and only the data/report types that are relevant to that individual school will be extracted into the Excel spreadsheet. Each spreadsheet will consist of a number of pages containing only the report types that are specific to the individual school - not all schools receive the same report types. The different report types will be written away to a separate worksheet, assisting users when searching for data. Once the DataStore extracts have been performed, the application will send the reports as an attachment to each of the schools by email. This will be achieved using the school code extracted from the saved file name, accessing the school look-up table and extracting the email address.

Each school will be provided with a User Interface and DataStore-style search screen. This will perform searches against the extracted spreadsheets - enabling each of the schools to make searches against specific spreadsheets and report types.

This user interface will present the user with a means of searching all of their spreadsheets for data. "This project demonstrates that even though the data has been archived, it can still be used by external applications for reporting purposes," comments

Alastair Allars, Automation Specialist at Hitec. “Once again, it highlights the benefits of both DataStore and the advantages of using the API.”

Staffordshire County Council is responsible for ensuring each of the schools receive timely and accurate financial information. The automation of spreadsheet creation through DataStore will help to improve the quality of service they provide the schools.

“DataStore will provide the reassurance that every school will receive their tabulations on time,” comments Dave Cheeseman. “It also ensures certainty of production of the reports and hence, the response we are able to provide is immediate - DataStore is a more effective solution.”

6.4 Migration of paper documents into electronic files Project of Levy Gee

Levy Gee, one of the UK’s top twenty accountancy firms, is a financial and business consultancy firm which has developed a deep understanding of the added value that IT can bring to both its customers and consultants.

As a technology driven national practice, Levy Gee is able to service businesses around the UK and internationally. By its large investment in Information Technology, it can reduce time spent in data handling and improve the efficiency of the audit fieldwork. Because of its dealings on a day-to-day basis with large numbers of paper documents, Levy Gee took the decision to migrate its paper documents into electronic files.

These files can be permanently stored in a central location thus guaranteeing that the documents will always be available, from any place, at any time.

By using electronic files, Levy Gee’s employees avoid having to physically transport original paper files through the office. It also promotes the ability for a team to work on one document at the same time, thus giving clients better service and staff more working flexibility.

6.4.1 Technical Solution

Levy Gee chose HP Digital Sender devices and NSi AutoStore software for the integration of paper documents into its existing Lotus Notes database which is currently being migrated to Microsoft SharePoint Portal server. It deployed 32 Digital Senders throughout its UK and Spanish offices.

6.4.2 Benefits and ROI

“This combination of software and hardware truly integrates the worker and their knowledge. The impact on our business will be far reaching as we open our minds to the opportunity that this technology represents in terms of improved service delivery to

our clients, flexible working for staff, and lower operating costs”, said Julian Synett, managing partner, Levy Gee.

For Levy Gee, electronic files have helped to solve a serious space shortage issue in the London office. Recently Levy Gee started working with a large client, which without electronic files would not have been possible due to the large volume of paper storage space required by the case.

Tracking all its client correspondence such as letters, phone calls and emails, into a central location guarantees Levy Gee consultants shared and accurate understanding of its customers.

The electronic files have also helped the company’s Human Resources department to be more organised, and reduce administrative tasks and costs. Having scanned all past and current employee documents with the HP Digital Sender, they have created a paperless office, which is a great benefit for the whole company.

Kim Sands, Human Resources Director says, “The introduction of scanned files has enabled us to increase efficiency and improve the service we provide to internal clients.

We have estimated that departmental administration costs have been reduced by 60% and that our improved processes with external suppliers have lowered outsourcing fees.

Based on this successful collaboration, HP and Levy Gee IT Consulting Department are now working together to deliver Business and Technical services around its existing Microsoft Share Point Portal Server. Levy Gee’s IT Consulting Department will provide services for Taxonomy and Business Categorization.

“HP Digital Sending Solutions give customers a fast, inexpensive way to turn paper into electronic documents and distribute them at the push of a button. This is a major leap forward for Levy Gee which can only have the most positive impact on our business”, says Julian Synett, managing partner, Levy Gee.

7. Outlook

7.1 Formats

Certain document formats have become timeless and can be seen as ‘safe bets’ for long-term storage. ASCII, TIFF, RTF are well established standards in universal use, while PDF, HTML, JPEG and GIF have rapidly established themselves as part of that group through their widespread adoption on the web.

XML is emerging as ‘the’ standard for web-based trading applications. Consider for example an application in which a form needs to be circulated to a group of applicants, for return and completion. The form can be distributed in Word format, or PDF, or HTML; most of us have some experience of handling forms electronically in this way. What such processes cannot do however, is validate the form. Once the form has been returned, it is checked by the initiator for completeness and accuracy, and if necessary re-issued for correction. While such rework is much faster than in a paper-based process, it is still rework.

What XML offers, is the ability to embed the validation rules in the document itself; in other words, it checks itself for accuracy and completeness, eliminating the need for subsequent checking and rework.

The technologies themselves are also well established. Most vendors offer solutions that support multiple media and formats, enabling your organisation to determine both the best solution to address your archival needs and the best format or formats to use for storing documents long-term.

7.2 Conversion Strategy

Empirical results from the many case studies that have been published in this area show clear guidelines for creating a successful conversion strategy:

- The strategy must be planned and co-ordinated with the same attention to detail as any other major project; an archive is a critical long-term corporate asset.
- The nature and lifecycle of the documents concerned must be thoroughly understood: how are they created; at what point is capture best undertaken; what is the retrieval profile at different stages in the lifecycle; how should the archive be classified; what are the retention and disposition requirements for documents.
- The long-term sustainability of the archive must be considered. It may not be sufficient to assume that formats and storage media will endure for the lifetime of the archive, a pro-active policy may be required which presumes that re-migration will be a periodic event; to more advanced formats, to higher density storage media, or to meet changes in access requirements (for example as a result of Freedom of Information).

- The portability of the archive must also be considered. In a human resource application for example, it is entirely feasible that documents created for a new joiner at age 16 will remain dormant until retirement at 65, and even beyond (pension benefits payable on death for example), then become critical. The HR system in use today may be long obsolete by that time, but the archive must remain durable.

7.3 Archive Value

The archive must be considered as an asset in its own right, independent of the system used to access it and the methods used to store it. Developing an archive strategy therefore needs to take account of a number of factors in developing both the conversion approach, access and storage media.

Lifetime of the media must be considered in relation to the documents lifecycle. – for example storing documents that must be retained for 80 years on media that may last only 30 years. Also, the systems used to access that archive will inevitably change over time – Word 2097 may not be able to read documents created on Word 1997.

An archive strategy is not just concerned with achieving the conversion, but on ensuring that the archive can be migrated to new storage devices, and transformed to new formats, that have not yet been developed.

The major cost of conversion is the initial clerical effort in physically converting and categorising pieces of paper or other media. These costs will vary with the intended future use of the archive.

A pharmaceutical drug research archive may require conversion of all research notes to text whereas a client policy file will only require indexing by client name and policy number. Time therefore must be invested in understanding the future requirements for accessing the documents and their use so that the archive is fit for purpose. The need to undertake further conversion at a later stage can prove difficult and certainly more costly.

The archive strategy must also consider the ongoing maintenance of that archive; a film archive will at some stage in the future need to accommodate new forms of digital media, and an archive for paper-based forms will need to allow for future electronic form submission.

7.4 Technology Developments

Backfile Conversion has benefited from a number of recent advances in Technology and the progress in International/Nationally accepted standards and codes of practice:

- Recent developments in Intelligent Document Recognition IDR techniques provide more cost effective approach to capturing the metadata and automating the conversion process.
- Storage costs have plummeted for both RAID and Optical discs and their performance improved. There are also many new storage concepts becoming available to users. Two such configurations are Network Attached Storage (NAS) and Storage Area Networks (SAN). These are in addition to Direct Attached Storage which is in common use . They provide improved access and performance to the Archive with built in security and backup. Manufacturers have also addressed the need for backward compatibility of optical discs. Hewlett Packard 5¼inch optical disc drive have a capacity of 9.1 gigabytes but still allow the original 650 megabyte device, launched in 1988, to be read on that drive. The manufacturers, such as Hewlett Packard have committed to providing similar backward compatibility with their new range of 40 gigabyte opticals to be launched in the next two years.
- XML and JPEG 2000 are examples of standards that are future proofing against changes in operating systems and line of business operations. All mainstream suppliers of technology and software are incorporating these standards.
- National and International codes of practice to address the legal admissibility of electronic documents are now in place nationally in most European countries and an ISO standard is in draft format. EDMS and ERMS suppliers are ensuring their systems can meet these requirements.
- Standards for the management of electronic records have also been addressed at all levels. The ISO 15489 Information and documentation: Records management was published in 2001 so was the European Commission MoReq specification. The UK PRO have also set up a minimum requirement for ERMS which ERMS suppliers can submit their product for assessment and accreditation.

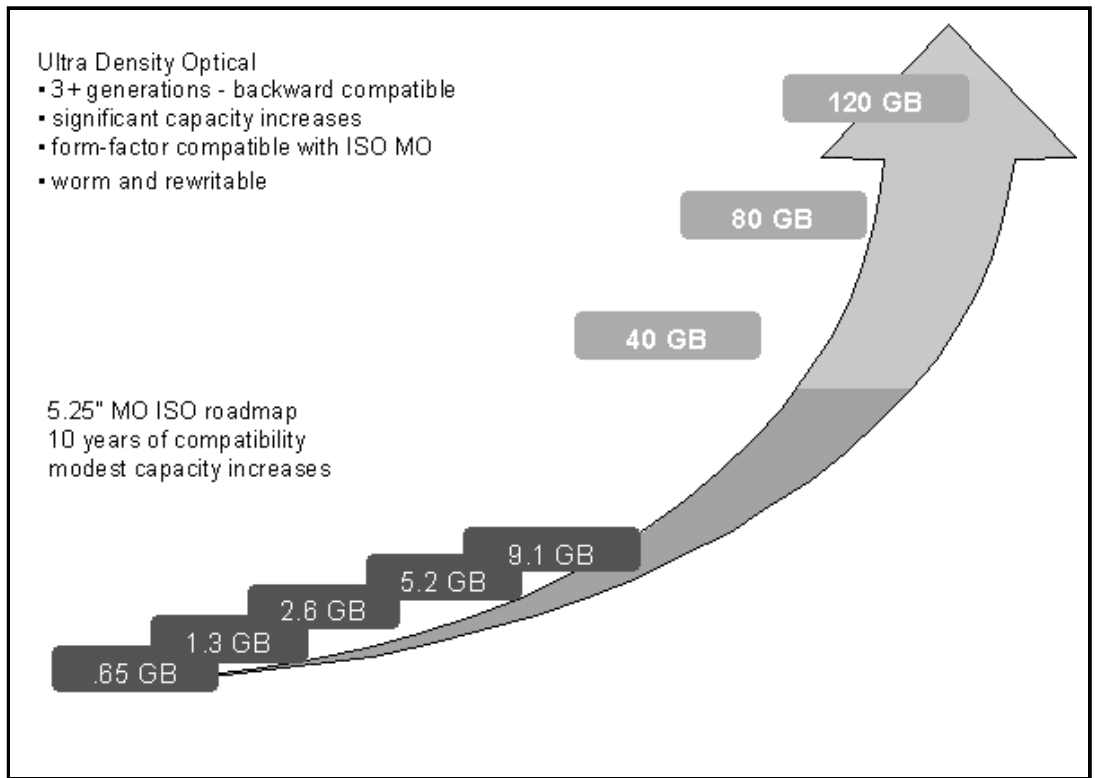


Figure 12: Trends in Optical Storage

Glossary

ADL (Advanced Distributed Learning)	ADL is an initiative by the U.S. Department of Defence to achieve interoperability across computer and Internet-based learning courseware through the development of a common technical framework, which contains content in the form of reusable learning objects.
Associative Access	Knowledge retrieval based on pattern matching between an unstructured query (text paragraph) and a document content store.
Authoring tools	Tools/SW to create and adapt content to the web for use in an online course. They assist in creating e-learning solutions and provide a “do-it-yourself” option for placing content and materials online.
Categorization / Category	Assigning documents to different groups by performing content-related analysis - so called categories. Categorization schemes are typically built upon business processes and business rules or rely on knowledge domains within an organisation.
CD-ROM assessment	An assessment or survey that can be accessed and completed by using a CD-ROM launched through a company’s intranet. CD-ROM based assessments also can be used on a desktop stand-alone computer if the assessment is a self-assessment for the benefit of the trainee only. Alternatively, a CD-ROM-based survey can be printed (if the CD-ROM has a print capability) and used as a paper-based survey.
Computer-based training	A term used to describe any computer-delivered training, including CD-ROM, the Internet and Intranets. Sometimes referred to as Computer-assisted instruction (CAI), CBT is asynchronous learning.
Classification / Class	Collection of methods applied to categorize documents by analysing their content. In many cases, categories and classes are identical. Categories incorporate the semantics of the application, whereas classes may also be of formal nature.
Classify	Classification is a method of assigning retention/disposition rules to records. Similar to the Declare function, this can be a completely manual process or process-driven, depending on the particular implementation. As a minimum, the user can be presented with a list of allowable file codes from a drop-down list (manual classification). Ideally, the desktop process/application can automate classification by triggering a file code selection from a property or characteristic of the process/application.
Content Search	Information retrieval based on pattern matching between a query (text paragraph) and a document repository.
Distance learning/ Interactive Distance Learning (IDL)	Traditionally refers to a broadcast of a lecture to distant locations, usually through video presentations. IDL is a real-time learning session where people in different locations can communicate with each other. Videoconferencing, audio conferencing or any live computer conferencing (e.g., chat rooms) are all examples of IDL.
Document	A document (any form or format), an email message or attachment, a document created within a desktop application such as MS Word, regardless of format. There are two forms of document: Electronic Document: Body (text) of the document is stored in electronic format and can be read. If declared as a record, an electronic document becomes a managed record (i.e. a document may or may not be a (declared) record) Non-Electronic Document (Ndoc): A physical document of any form (maps, paper, VHS video tapes, etc.). Body is not recorded in electronic form, but descriptive metadata is stored and tracked within CM (profile). If declared as a record, an Ndoc becomes a managed record (i.e. an Ndoc may or may not be a (declared) record).
Document Life Cycle Management	The records life cycle is the life span of a record from its creation or receipt to its final disposition. It is usually described in three stages: creation, maintenance and use, and final disposition. e-Records applies management to all three stages. With e-Records, the records manager can create and maintain the official rules that will dictate when to destroy (or permanently keep) electronic records, as well as record and enforce any conditions that apply to destruction (e.g. destroy 2 years following

	contract completion). Finally, the records manager can carry out the physical destruction of electronic records, maintaining a legal audit file.
Document Security Control	Access control to documents (non-declared records) Note: Document security control is different from Records Security Control.
Electronic Recordkeeping	The practice of applying formal corporate recordkeeping practices and methods to electronic documents (records).
Electronic Signature	A signature is a bit string that indicates whether or not certain terms occur in a document.
Enterprise Content Management	Manage all content (i.e. unstructured information) relevant to the organisation. It embraces three historically separate technologies: web content management, document management, and digital media asset management. While outwardly dissimilar, all of these forms of enterprise content share similar needs for mass storage, search and access, personalisation, integration with legacy applications, access and version control, and rapid delivery over the internet.
EPSS (electronic program support system)	An electronic system that provides integrated, on-demand access to information, advice, learning experiences and tools. In essence, the computer is providing coaching support (i.e. the principal of technology based knowledge management).
File	A disk "file", something stored on electronic media, of any file. Does not necessarily denote a record. For example, "image files are stored on a server" simply refers to the electronic images, and implies nothing about the records status. Will be used in the context of describing the storage of documents and related information to electronic media.
File Plan Administration	Design and administration of the corporate file plan. The records manager can design file plan components. With Tarian's file plan designer, the records manager can design classes of file plan objects (files, records, folders, etc), then define the attributes of these classes. Relationships between classes are then defined (i.e. files can contain files, records and folders). Various views of the file plan may be defined. For instance, a warehouse view might present a view of the physical folders in the organisation, whereas a numeric view might present the sorted numeric structure for maintenance purposes. The records manager can create pick-lists enforcing consistency within the file plan, component profiles that define the characteristics of the file plan, and default values to simplify daily file creation tasks. Policies, Permissions, and Suspensions may be assigned to file plan objects.
Information mining	Linguistic services to find hidden information in text documents on content servers
Information Retrieval	An information retrieval (IR) system informs on the existence (or non-existence) and origins of documents relating to the user's query. It does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. This specifically excludes Question.
Keyword Search	Information retrieval method based on literal match of words.
Learning Resource interchange (LRN)	LRN is the Microsoft implementation of the IMS Content Packaging Specification. It consists of an XML-based schema and an LRN toolkit. It enables a standard method of description of content, making it easier to create, reuse and customise content objects with an XML editor, whether initially developed from scratch or bought under license from vendors.
Neural Networks	In information technology, a neural network is a system of programs and data structures that approximates the operation of the human brain. Typically, a neural network is initially "trained" or fed large amounts of data. A program can then tell the network how to behave in response to an external stimulus (for example, to classify a document based on its content).
Pattern Matching/Recognition	Matching/Recognition of objects based on features. Pattern Matching with regard to text documents means to identify and match words and phrases from different documents under the assumption that the more features match, the more similar the contents are.
Personalisation	The ability to provide the user with the right content both from the user's and Web site owner's perspective. A personalisation algorithm determines whether content is

	presented to the user, and if so, in what order of priority.
Portal	A single integrated point of comprehensive, ubiquitous, and useful access to information (data), applications, and people.
Record	Any form of recorded information that is under records management control. Records are either Physical or Electronic. Records may take any of the following four forms: Document: A document (see above) that has been declared as a record. Once declared as a record, the document is under records management control Folder: A folder of (paper) documents. Individual documents within the folder may or may not be treated as records (declared Ndocs). The physical handling of the folder is managed by Tarian's Physical Records Module Box: A box of (typically) paper documents. Usually contains folders (see above), which are individually managed as records, but may alternatively contain records other than folders such as loose documents of a given subject. The physical handling of the box is managed by Tarian's Physical Records Module Ndoc: A declared Ndoc (See above for definition of Ndoc) Important: A document (electronic or Ndoc) will not be considered to be a record until has been declared.
Record, Electronic	Electronic Records (e-Records). Any information (document) recorded in electronic form, on any digital media, that has been Declared to be a record. Characteristics of an e-Record: Document is in electronic form Metadata is associated with the document Document has been classified against a file plan Only the authorised Records manager has the means by which to apply retention/disposition to the document.
Record, Physical	Folders, Boxes, Ndocs to which records management control has been applied. A document (electronic or Ndoc) becomes an e-Record only once it has been declared.
Records Administration	The administrative infrastructure represents the tasks that the records manager carries out on the entire organisation's collection of declared records. Conducted within Tarian's Records Administration Client, a browser-based web application. End users never see this process. Consists of the following four broad activities; File Plan Administration, Records Security Control, LifeCycle Management, and Reporting.
Records Manager	Conducts one or more records administrative functions.
Records Security Control	Access control to declared records. Users and Groups of users may be created, and assigned roles and policies that will interact to determine the records users are able to access. Note: Records security control is different from Document Security Control.
Reporting	The process of generating reports from data managed by eRecords solution. It is a tow-step process. Reports are first designed, and the design is saved for later reuse. Second, reports are generated by running the report design against the data.
Repository	Physical storage are for documents and/or electronic records.
Retention Rules	(Retention Schedule). The set of rules which specify how long to keep (retention) records, and what to do with them at the end of their lifecycle (disposition).
Syntactical Analysis	Syntactical analysis derives the syntactic category of words or phrases based on (language dependent) dictionaries and grammars. Example: house – noun.
Thesaurus	A book that lists words in groups of synonyms and related concepts.
Volume	Folder. A Volume will be referred to as a folder (common US terminology).
Virtual Reality (VR)	Virtual Reality simulations (usually involving wearing headgear and electronic gloves) that immerse users in a simulated reality that gives the sensation of being in a three-dimensional world.

Abbreviations

ASP	Application Service Provider
AVI	Audio Video Interleaving
BCR	Bar Coding
BPM	Business Process Management
CBT	Computer Based Training
CCD	Charge Couple Devices
CM	Content Management
COLD	Computer Output to Laser Disk
COM	Component Object Model
COOL	Computer Output On Line
DBMS	Database Management System
DMS	Document Management System
DRT	Document Related Technologies
ECM	Enterprise Content Management
E-Learning	Education, training and structured information delivered electronically
ERM	Enterprise Report Management
ERP	Enterprise Resource Planning
E-Term	European programme for Training in Electronic Records Management
FDDI	Fibre Distributed Data Interface
GIF	Graphic Interchange Format
HTML	Hypertext Mark-up Language
ICR	Intelligent Character Recognition
ICT	Information and Communication Technology
IDM	Integrated Document Management
ISDN	Integrated Services Digital Network
ISO	International Standards Organisation
JPEG	Joint Photographic Experts Group
KM	Knowledge Management
LAN	Local Area Network
LDAP	Lightweight Directory Access Protocol
MoReq	Model Requirements for the management of electronic records
MPEG	Moving Pictures Expert Group
NAS	Network Attached Storage
OCR	Optical Character Recognition
ODCB	Open Database Connectivity
OLE	Object Linking & Embedding
OMR	Optical Mark Recognition
PDF	Portable Document Format
PPP	Point-to-Point Protocol
RMS	Records Management System
RTF	Rich Text Format
SAN	Storage Area Networks
SQL	Structured Query Language
TCP/IP	Transmission Control Protocol/Internet Protocol
TIFF	Tag Image File Format
WAN	Wide Area Network
WAV	Audio Format File
WCM	Web Content Management
WebDAV	Web-based Distributed Authoring & Versioning
WORM	Right once read many times
XML	eXtensible Mark-up Language

References

The references are largely derived from organisations rather than individuals, and often a synthesis of comments from diverse sources about a topic.

This is a list of some of the key topics and an authoritative website on the topic.

UK Public Records Office publications on records management	http://www.pro.gov.uk/recordsmanagement
Code of Practise for legal admissibility and evidential weight of information stored electronically (PD0008)	http://www.bsi-global.com/
International Standard on Records Management Practises	http://www.records.nsw.gov.au/publicsector/standards/introintl.htm
Ordnance Survey	http://www.ordsvy.gov.uk/
JPEG format	http://www.jpeg.org
MPEG format	http://www.mpeg.org/MPEG/index.html
GIF/LZW format	http://www.unisys.com/unisys/lzw/
XML format	http://www.w3.org/XML/
PDF format	http://www.adobe.com/products/acrobat/adobepdf.html
Sliger, Susan	Planning a document conversion http://www.convertdoc.com/

Authoring Company

Hewlett-Packard



i n v e n t

Hewlett-Packard Company - a leading global provider of computing and imaging solutions and services - is focused on making technology and its benefits accessible to all.

HP had total revenue of \$45.2 billion in its 2001 fiscal year.

Information about HP and its products can be found on the World Wide Web at <http://www.hp.com>.

Contact Authoring Company

European headquarter

Hewlett-Packard GmbH

Nancy Romany

Schickardstrasse 32

71034 Boeblingen - Germany

Tel. +49 (0)7031 14 6837

E-Mail: nancy_romany@hp.com

Industry White Paper Series on Records, Document and Enterprise Content Management for the Public Sector

Capture, Indexing & Auto-Categorization

Intelligent methods for the acquisition and retrieval of information stored in digital archives

ISBN 3-936534-01-2

Hewlett-Packard GmbH

Conversion & Document Formats

Backfile conversion and format issues for information stored in digital archives

ISBN 3-936534-02-0

FileNET Corporation

Content Management

Managing the Lifecycle of Information

ISBN 3-936534-03-9

IBM

Access & Protection

Managing Open Access & Information Protection

ISBN 3-936534-04-7

Kodak

Availability & Preservation

Long-term Availability & Preservation of digital information

ISBN 3-936534-05-5

TRW Systems Europe / UCL - University College London / comunicando spa

Education, Training & Operation

From the Traditional Archivist to the Information Manager

ISBN 3-936534-07-1

Publishing Information

The series of six Industry White Papers is published to address the needs of public administration and archives at the European, national, federal and local level and to educate the public sector throughout Europe about available solutions for archival problems on relevant topics about acquisition, management, long term storage, multilingual access, indexing and training issues.

DLM-Forum

The current DLM acronym stands for *Données Lisibles par Machine* (Machine Readable Data). It is proposed that after the DLM-Forum 2002 in Barcelona this definition be broadened to embrace the complete "**Document Lifecycle Management**". The DLM-Forum is based on the conclusions of the Council of the European Union, concerning greater co-operation in the field of archives (17 June 1994). The DLM-Forum 2002 in Barcelona will be the third multidisciplinary European DLM-Forum on electronic records to be organised. It will build on the challenge that the second DLM-Forum in 1999 issued to the ICT (Information, Communications & Technology) industry to identify and provide practical solutions for electronic document and records management. The task of safeguarding and ensuring the continued accessibility of the European archival heritage in the context of the Information Society is the primary concern of the DLM-Forum on Electronic Records. The DLM-Forum asks industry to actively participate in the multidisciplinary effort aimed at safeguarding and rendering accessible archives as the memory of the Information Society and to improve and develop products to this end in collaboration with the users.

European Commission SG.B.3

Office JECL 3/36, Rue de la Loi 200, B-1049 Brussels, Belgium

A/e: d1m-forum@cec.eu.int

AIIM International - The Enterprise Content Management Association

AIIM International is the leading global industry association that connects the communities of users and suppliers of Enterprise Content Management. A neutral and unbiased source of information, AIIM International produces educational, solution-oriented events and conferences, provides up-to-the-minute industry information through publications and its industry web portal, and is an ANSI/ISO-accredited standards developer.

AIIM Europe is member of the DLM-Monitoring Committee and co-ordinates the activities of the DLM/ICT-Working Group.

AIIM International, Europe

Chappell House, The Green, Datchet, Berkshire SL3 9EH, UK

<http://www.aiim.org>

Industry White Paper Series on Records, Document and Enterprise Content Management for the Public Sector

The Industry White Papers are published by the DLM-Forum of the European Commission and AIIM International Europe to address the needs of public administration and archives at the European, national, federal and local level and to educate the public sector throughout Europe about available solutions for archival problems on relevant topics about acquisition, management, long term storage, multilingual access, indexing and training issues. The leading suppliers of Enterprise Content Management technologies participate in this series and focus on electronic archival, document management and records management for the public sector in the European Community.

Conversion & Document Formats

This White Paper addresses the issues which arise when considering the conversion of existing physical archives, that contain documents of different formats and types, into electronic format. These issues are broad in nature including the logistics of capture involving high volumes; the determination of appropriate strategies and tactics, for both delivering the conversion and maintaining normal business operations in the process; and the adoption of appropriate, reliable and sustainable document formats.

ISBN 3-936534-00-4 (Series)

ISBN 3-936534-02-0