

Availability & Preservation

Long-term Availability &
Preservation of Digital Information

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



AIIM
International



© AIIM International Europe 2002

© DLM-Forum 2002

© Kodak 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 Papers Series)

ISBN 3-936534-05-5 (Industry White Paper 5)

Price (excl. VAT): 10 €

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Printed in United Kingdom by Stephens & George Print Group

Availability & Preservation

Long-term Availability &
Preservation of Digital Information

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
Kodak Limited
Robert M. Young
PO Box 66
Station Road,
Hemel Hempstead, Herts HP1 2TL - UK
Tel : + 44 (0)1442 844791
robert.m.young@kodak.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

Creative thinking is necessary to meet the challenge of managing the records of the electronic age as new applications generate digital documents in ever-increasing volumes. These documents contain factual information that has business, regulatory, and cultural and historical significance. They merit preservation.

Kodak believes it is best to strive for balance. Information must be made accessible while meeting long term retention and preservation goals. Organisations must embrace new technology to achieve higher goals for service, efficiency and revenue. Meanwhile, funding levels must be set according to society's ability to pay.

Initiatives such as this one assure that archivists, document and content managers best serve citizens by preserving the unquestioned trustworthiness of the documents in their care. Kodak is pleased to be a partner in this noble effort, the benefits of which will be enjoyed by future generations.

A handwritten signature in black ink that reads "Michael Barrett". The signature is written in a cursive, flowing style.

Michael Barrett
Vice President of Imagelink Products, Commercial Imaging Group,
Kodak

Table of Content

1.	Introduction	7
2.	Electronic Archives Are a Virtual Memory of the Information Society ...	8
2.1	Digital Document Overview	8
2.2	Accumulating digital records of an online age.....	8
2.3	Preservation for availability across generations.....	9
2.4	The importance of content and form	9
2.5	Balancing processibility and inalterability	10
2.6	Digital Preservation with a Reference Archive.....	11
2.7	Current long-term preservation strategies	11
2.8	Conclusion.....	11
3.	Long-Term Availability	12
3.1	Reference Archive overview	12
3.2	Process overview.....	12
3.3	Application Requirements.....	12
3.4	Benefits of Analogue Rendering	12
3.5	Challenges of Analogue Rendering.....	13
3.6	Conclusion.....	13
4.	Constant Migration of Information	14
4.1	Migration overview	14
4.2	Process overview.....	14
4.3	Application Requirements.....	14
4.4	Benefits of Migration	14
4.5	Challenges of Migration.....	14
4.6	Analysis of digital preservation based solely on migration	16
4.7	Conclusion.....	16
5.	Standards for Reference Archive Media and Reference Archive Management Software	17
5.1	Reference Archive Media.....	17
5.2	Reference Archive Management Software.....	17
5.3	Conclusion.....	17
6.	Best Practice Applications	18
6.1	UK Census 2001 Dual-Path Record Keeping Project of UK Office of National Statistics.....	18
6.2	Records Management of State of Virginia Land Project of Library of Virginia & County Clerk’s Offices	19

7.	Outlook.....	21
7.1	Digital Preservation strategy and process overview.....	21
7.2	Benefits of Digital Preservation with Reference Archive protection.....	21
7.3	Future developments	21
7.4	Solutions availability.....	22
7.5	Conclusion.....	22
	Glossary.....	23
	Abbreviations.....	26
	Authoring Company	27

1. Introduction

This Industry White Paper offers Kodak's perspective on the long-term retention and availability of digital information. Digital documents require management just as their paper-based forerunners do. The electronic technologies used to create, distribute, and store them present special problems for archiving this information as time advances. Successive iterations of technology, inevitable media decay, and their inherent editability ill-suits them for long-term keeping in their native formats.

A Reference Archive of permanent document images offers a cost effective long-term solution. By rendering digital information to microfilm as uncoded, analogue images, Organisations may create technology-proof repositories.

For the time that it remains economically feasible to keep digital archives available online, a Reference Archive will provide proof of their trustworthiness. Later, a Reference Archive may be used to access the stored document images over many human and technological generations. This approach avoids the pitfalls of constant migration and risk caused by eventual technological obsolescence without placing a significant financial burden on society.

The technology necessary to implement a Reference Archive strategy is available today. A variety of Organisations are using it to render scanned documents and an increasing variety of "born-digital" documents. Existing standards are available to guarantee the quality and continuity of the archival and retrieval processes.

The Reference Archive process is expected to provide a useful foundation for future content management schemes. By containing trustworthy records of communications and transactions, the Reference Archive will safeguard agencies and the rights of citizens. Thus long-term preservation and availability of digital information will be assured.



John Mancini
AIIM International

2. Electronic Archives Are a Virtual Memory of the Information Society

Permanent Trustworthiness Requires the Validation of a Reference Archive.

2.1 Digital Document Overview

2.2 Accumulating digital records of an online age

Accountability and trust form the foundation that allows political institutions to govern. The very dynamism that makes electronic technology so useful runs counter to the preservation of unchangeable “facts” in digital form. As technology changes, there must be an unquestionable record – a “Reference Archive” – to which those governed can use to check the validity of the information available online.

We are witnessing the birth of electronic government. Already the business of government is facilitated by the ever-growing application of electronic tools for communicating, processing transactions, and managing information. These activities are recorded as digital documents – electronic documentation held on digital media, whether stored in servers, on CDs, or on magnetic tape.

Digital documents may be found in virtually every area of government. Legislative, executive, administrative, regulatory, and judicial agencies now traffic within and among themselves and with citizens using networks and the Internet. In the course of doing its business, government produces official records that document issues, decision-making processes, procedures, laws, and rulings. These must be preserved and made available to citizens as part of the social contract. Only with oversight and accountability will citizens continue to trust their governments and political institutions.

Another class of digital documents has to do with the services provided by government. Vital records, environmental and safety compliance documents, tax records, benefit disbursement, and census information are all being managed online. Citizens depend on government record keepers for accuracy and expect a clear audit trail when questions arise about funds paid, collected, or held for them.

Many of these digital documents are born digitally as email, word processing or spreadsheet files, sometimes routed as email attachments and as on-line forms. Other digital documents are converted from paper documents to digital formats by scanners and facsimile devices. The existence of this store of documents has brought about sweeping improvements in the accessibility of information to citizens via the Internet and interactive kiosks networked to government agencies. Process productivity and expectations for service levels have also risen.

With each passing day, the reservoir of digital documents grows. Often, there is no associated hard copy output to archive via conventional means. In many cases, paper documents are essentially a disposable vehicle for data capture. The challenge is to preserve, unchanged, those documents that must survive for very long time periods, regardless of how they were created.

These digital documents reside in a variety of places under greater and lesser degrees of security. Servers and disc jukeboxes may be connected to public and private networks. Hard disk drives store files in the personal desktops used by legislators and administrators. Some documents may be transitory, such as a daily Web page. Others may be more durable in form, such as a data base report of current pensioner benefits written to CD. They may be distributed across a multi-national enterprise. They may exist as “meta” documents, with content distributed in various locations. Over time, the problem is that media decays and hardware and software platforms evolve, placing the electronically stored information – and the public trust – at risk.

Government must now find a way to preserve the trustworthiness of its digital documents in a fiscally and legally responsible manner. Electronic archives alone do not meet this test for the reasons cited above. This paper lays out the argument for a long-term preservation strategy based on a permanent, low-maintenance Reference Archive using future-proof practices and technology.

2.3 Preservation for availability across generations

“Long term” in the context of this paper refers to storage for a period of ten years or more, as outlined by industry consultants Gartner Group.

Legally, records may have to be retained for the life of a product, customer relationship or the life of the company. To support historical research, governmental agencies may keep documents in perpetuity. Individual records may be accessed extremely rarely, although the archive as a whole may be accessed daily (Source: “Management Update: Important Issues About Digital Data Preservation,” IGG-08082001-04).

Government has its own compelling set of reasons to retain documents for long periods of time. The very fabric of the social contracts that weave together nations and communities is based on documents. The categories extend beyond constitutions, charters, and compacts. Citizens expect a continuous trail of more mundane transaction documentation, legal determinations, and vital records to be maintained forever. When necessary, these must be available so that citizens can prove their legal rights and status, as well as the contractual relationships they have with various government agencies.

2.4 The importance of content and form

The realm of digital documents is much more fluid than the world of traditional paper-based documents. Electronic records (e-records) are composed of binary data. To be complete, the e-record must include content, context, and structure, the data for which are

frequently distributed across multiple sources, tracked as metadata. For human comprehension, this data must be interpreted by technology for presentation in analogue form on printed output and CRT screens. For ongoing availability, all of the distributed data and the metadata must be maintained intact. Questions about legality and economical manageability become more apparent as time passes.

Contrast this with paper documents and conventional microfilm records. Content, context, structure, and signature all appear as a unitary, integrated record. Taken together, archived paper and microform document images have centuries of legal acceptance behind them as faithful analogue repositories of public and private records. They are human-readable, essentially independent of technology.

2.5 Balancing processibility and inalterability

Those pursuing strategies for archiving digital documents find ourselves in a somewhat contradictory situation. We want the dynamism of digital processing and manipulation, but we also want the trustworthiness of analogue records. The answer is to do both in an economical manner that is virtually transparent to the people using the digital archives.

On the one hand, government is being urged to maintain electronically-accessible information repositories. In this scenario, the documents must be in forms that may be readily processed by information systems and made available on line.

However, the processibility that makes digital files so desirable also makes them vulnerable to technological changes and media degradation.

On the other hand, while databases, email routing, and document versions change from second to second, we need to archive slices of time for later reference. To be trusted, these records must capture – and preserve – the status of the account, transaction, or contractual relationship exactly as it was at that moment. Unalterable permanence is the only assurance of the preservation of the facts involved as truths.

The best approach to creating trustworthy digital archives follows a dual track. For short- and medium-term needs, maintain structured electronic archives in their native formats for a reasonable period of time before allowing them to expire.

For long-term trustworthiness, create a Reference Archive based on analogue renditions of the documents as images of their application formats, which can be converted back to digital at any time. It will be seen that these two archives complement each other.

The Reference Archive provides the proof that validates the content of any electronic archive. The Reference Archive also provides economical record management that satisfies the desire for digital document preservation and availability spanning generations.

2.6 Digital Preservation with a Reference Archive

As we examine strategies for long-term availability and preservation further, we will consider the following criteria:

- permanence – is the trustworthiness of the digital document preserved?
- accessible – can the digital document be retrieved?
- available – can the digital document be viewed, routed, and output?
- economical – is the cost of ongoing storage and management sustainable?

2.7 Current long-term preservation strategies

The Gartner Group identifies four avenues to electronic archiving in their paper, “Long-Term Strategies for Digital Data Preservation.” These may be summarised as follows:

- Technology Preservation – keeps the hardware and software used to create and access the data originally. Considered impractical.
- Data Migration – moves data from old technology formats and media to new. Provides continual processibility, but incurs ongoing expense and undetected data corruption.
- Technology Emulation – preserves the environment and functionality of old technology using a detailed metadata description. Standards are not in place and metadata is subject to obsolescence.
- Converting Digital to Analogue – renders digital files as human-readable document images. Because the information must be re-digitised when required, it is not suitable for frequently or widely accessed records.

Source: “Management Update: Long-Term Strategies for Digital Data Preservation,” IGG-07042001-04.

2.8 Conclusion

It will be seen that a blend of short-term data migration coupled with digital conversion (or analogue rendering) to create a long-term Reference Archive provides the dual track strategy referred to above.

3. Long-Term Availability

3.1 Reference Archive overview

A Reference Archive is created by transferring digital files to a humanly readable medium – microfilm. Digital document content is thus committed to a static format. In the short-term, it is available to answer any challenges to the authenticity of a digital archive. In the long-term, it replaces expired or obsoleted digital files with reliable, technology-independent records.

Paper and microfilm have proven to be trustworthy archive media for periods measured in centuries and approaching a century respectively. Microfilm has the additional familiar benefits of offering compact, secure, and easily managed storage. Microfilm is both cost-effective and easy to use. It has a life expectancy of 500 years, allowing for the availability of information years down the road.

3.2 Process overview

While the details vary, the process of analogue rendering to microfilm can be described as four primary steps. Digital documents or their storage addresses are assembled and organised electronically according to the chosen indexing scheme or schemes. Document queues are created and released to the rendering system. Software converts the files to digital images for output by the rendering hardware. Post processing transfers the microfilm roll and frame information to the host document management system for eventual search, retrieval, and rescanning to digital form for electronic routing, viewing, and printing.

3.3 Application Requirements

Beyond business needs, regulatory compliance, and preservationist concerns, image to film is best suited for agencies that:

- Place a premium on the absolute trustworthiness of their records
- Are concerned over the cost of ownership involved in preserving digital records
- Keep files that may need to be printed in the future
- Can afford a slight delay in the electronic retrieval of rarely accessed documents

3.4 Benefits of Analogue Rendering

It is a well-understood technology that is acceptable in virtually any court of law. Legal standards have been written around microfilm.

There are no migration issues because microfilm is immune to the evolution of technology:

- It is a robust storage medium, with a life expectancy of 500 years
- It offers a high level of reproduction quality
- It is a low-cost solution for preserving digital documents and their content
- Proven disaster recovery schemes exist already
- Duplicate records are easy to make for distribution and off-site backup
- Current retrieval technology is interoperable with digital systems
- Retrieval times are similar to those required for loading offline backup or archival digital media

3.5 Challenges of Analogue Rendering

Documents are non-executable without rescanning. Information is contained in a static format that cannot be changed, which is indeed a highly desirable characteristic in a Reference Archive. It does not support multimedia (animation), hypertext, GIS, or interactive Web pages. Not all digital documents and files lend themselves to this solution. Retrieval is not as fast as on-line digital solutions.

Analysis of a Reference Archive based on analogue rendering.

As a strategy for long-term availability and preservation, a Reference Archive makes good operational and economic sense:

- permanence – the trustworthiness of digital document content is preserved for generations
- accessible – document images can be scanned and managed digitally
- available – analogue document images will remain compatible with all future systems
- economical – analogue rendering to microfilm is an automated, write-once process using low-cost media and technology available and proven today

3.6 Conclusion

For availability and preservation across generations, it makes sense to build a Reference Archive based on analogue rendering to microfilm. The resulting images of digital documents are essentially permanent and unchangeable. No migration or refreshing is required. The trustworthiness and readability of the documents endures.

4. Constant Migration of Information

4.1 Migration overview

Migration transfers data from old technology formats and media to new. While this path provides continual processibility, it incurs massive ongoing expense and the risk of undetected data corruption. Experience shows that even the most disciplined migration and translation processes include a rate of error of five percent or more.

4.2 Process overview

This solution requires the periodic copying and reformatting of digitally stored information to the newest storage media and application or viewing software.

4.3 Application Requirements

Migration appears to be a sufficient solution for government agency applications that:

- Can accept a high degree of risk that records will be corrupted or rendered unreadable over the long term
- Need to retain the executability of the documents
- Are not bound by financial concerns; in other words, "money is no object"
- Are willing to go through migration every five to seven years

4.4 Benefits of Migration

- It can maintain digital records in immediately executable form
- It can employ technology neutral interchange format and updates as required
- It can utilise backward compatibility within vendor line as appropriate
- It can convert legacy applications to new technology platforms

4.5 Challenges of Migration

There are no industry standards in place for migration. Setting ultimate standards is not realistic in this ever-changing technology age. Evolving standards would require translation and will always be a step behind state-of-the-art use. Users are unlikely to conform to an ultimate set of standards that essentially ask them to relinquish their use of new capabilities.

Questions about the authenticity and trustworthiness of migrated records arise for two principle reasons. The process itself is error-prone, leading to transcription and transpositions that undermine the accuracy of the content. Dropped bits result in invalid files that cannot be opened. Migration is also subject to inevitable degradation of software functionality hampers the correct presentation of the records' content.

Migration is expensive and labour intensive and detracts from revenue-generating activities. The 100 percent QC/QA required to be absolutely sure content and formatting have been preserved is economically and operationally impossible to sustain.

Essentially, migration merely defers the problem. At some point, either the massiveness of the conversion or the rate of technological change is likely to overwhelm almost any migration scheme.

Consider a hypothetical high-volume imaging application, such as a centralised regulatory application capturing licensing applications and associated documents, may capture 50,000 documents a day, resulting in an annual volume of 12.5 million images. Over 6 years' time, technology changes, software becomes outdated, or service providers change. For whatever reason, an Organisation typically must migrate its existing images to a new format in this timeframe. The accompanying table and graphic demonstrate how the magnitude of the task increases.

Year	Migrate	New Images	Total
6	12 M	22 M	34 M
7	15 M	24 M	39 M
11	34 M	36 M	70 M
12	39 M	39 M	78 M
16	70 M	57 M	127M

Figure 1: Capture and Migration Image Volumes
 50,000 images/day = 12,500,000 images/year
 10% growth/year = 22M in yr 6 36M in yr 11

Another illustration of the magnitude of migration is provided by an analysis of the US Census reported in the July 20, 2000 National Archives Assembly Resolution.

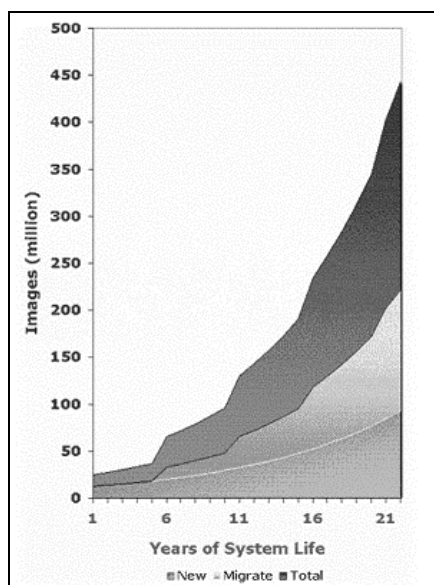


Figure 2: Years of System Life

This was based on an estimated volume of Census 2000 scanned images of approximately 60 terabytes. NARA's imaging experts estimated costs for maintaining that volume at \$5.3 million to \$10.5 million per year (\$53 million to \$105 million for the first 10 years).

The Bureau of Census has now revised its estimate of the amount of data involved upwards to 160 terabytes. NARA could expect the annual cost to easily reach \$14.31 million to \$28.4 million per year.

Here are some additional cautionary observations relating to the costs of a maintaining a digital archive based on a 1998 Gartner study:

- Long-term archive costs will exceed budgeted amounts by 300 percent to 500 percent through 2002 (0.9 probability)
- Actual media lifetime will under deliver claimed lifetimes by 50 percent to 100 percent through 2002 (0.9 probability)

Source: Long-term planning assumptions, N. Allen

4.6 Analysis of digital preservation based solely on migration

As a strategy for long-term availability and preservation, migration comes up short:

- permanence – the trustworthiness digital document degrades over generations
- accessible – records must be sent through a suspect transfer process or software must retain infinite backwards compatibility
- available – in order for the digital document to be viewed, routed, and output, its technology platform must migrate forward to remain compatible with future systems
- economical – the cost of the migration process, ongoing storage, and management appear impossible to sustain or justify in the long term

4.7 Conclusion

Migration is, at best, a short-term approach to digital archives. As technology changes accumulate and the rate of access falls off, migration's costs and risks outweigh the benefits of executability. The process of moving data to successive generations itself degrades the trustworthiness of the digital documents as accurate records – defeating their fundamental purpose as archives of historical activity.

5. Standards for Reference Archive Media and Reference Archive Management Software

The Reference Archive solution is based on readily available micrographic technology with a history of thoughtful, stepwise evolution. Because the content is preserved as analogue document images, a Reference Archive is inherently stable.

Currently there are no specific standards for Reference Archives. However, widely accepted ISO/ANSI standards for micrographics transfer directly to the media and management software used.

There is one nuance of Reference Archive creation that should be taken as a de facto standard until formal standards are in place. This is the desirability of capturing the digital document as close in time to its creation or the posting of an associated transaction as possible. “Instantaneous” capture provides the most trustworthy record of the action or decision documented by the digital document.

5.1 Reference Archive Media

The media used for analogue rendering conforms to ISO/ANSI standards for 16 mm microfilm. This means it is backwards compatible with current readers, scanners, and management software and should remain useful far into the future. Available Reference Archive media has a certified 500-year lifetime when processed and stored under specified conditions.

5.2 Reference Archive Management Software

A variety of document image management software packages in use today may be used to manage a Reference Archive of digital documents. Current book, batch, roll, frame indices interface with electronic document management systems to provide on-line retrieval via local and Internet network access. It is anticipated that proper indexing will also provide connections to emerging content management schemes.

5.3 Conclusion

Archivists and record managers may easily adopt the discipline of a Reference Archive using familiar standards and practices in use today. Once again, the permanence and analogue format of the captured document images assure forward integration without any software or hardware interpretation. For these and other reasons stated elsewhere, a Reference Archive is a low-risk, immediate solution for the long-term preservation of digital information.

6. Best Practice Applications

6.1 UK Census 2001 Dual-Path Record Keeping Project of UK Office of National Statistics

6.1.1 User Organisation

The Office for National Statistics (ONS) is the government department that provides statistical and registration services. ONS is responsible for producing a wide range of key economic and social statistics which are used by policy makers across government to create evidence-based policies and monitor performance against them. The Office also builds and maintains data sources both for itself and for its business and research customers. It makes statistics available so that everyone can easily assess the state of the nation, the performance of government and their own position.

6.1.2 Problem

The purpose of the UK 2001 Census Project was to collect data about the population in a timely, accurate, and economical manner. United Nations members are charged with compiling their national census data by 2004. These goals prompted the agency and its service providers to automate the process with an electronic imaging system that scanned forms for data collection. An estimated volume of over 30 million forms, resulting in over 600 million pages were involved. Process improvements notwithstanding, the project also needed to meet the long-term archive and access requirements agreed with the Public Records Office.

6.1.3 Technical Solution

A score of high-speed scanners were used to digitise the bulk of the Census forms. Exception forms that were damaged or in fragile condition were diverted to dual-purpose scanners with flatbed capabilities. Incoming digital documents were transferred to an image database as TIFF images and the host workflow management system updated accordingly. Recognition technology was used to automatically record responses for data tabulation. The images were also sent to a bank of Document Archive Writers, where they were rendered to 16 mm microfilm.

6.1.4 Organisational Solution

After receipt and checking of completed forms gathered from the Census field Organisation, pages were scanned in batches. By use of control information on batch headers, the system ensured that every form and page was scanned and processed. This helped assure that 100 percent of the data was captured from 100 percent of the population, despite a high throughput rate.

Two other key efficiency gains came at the end of the process when the scanned images were output to reference archive media. Firstly, the images will meet the 100-year archive and access requirements set by the Records Office. Images will remain available to

historical researchers and genealogists and can be used to recreate an electronic image for public viewing.

6.1.5 Project Experiences

The integration of technologies and service providers proved to be successful and met the goals of the UK Census 2001 project.

6.1.6 R.O.I.

For the UK Office of National Statistics, this solution limited capital expenditure, minimised labour costs and training, and accelerated the execution of the project.

Gains were also realised from the perspective of the Public Records Office. Long-term, the reference archive component of this project eliminates the risk and expense of regularly migrating the entire image base to new platforms as existing hardware and software inevitably becomes obsolete.

A second benefit is that the storage space taken up by the microfilm will be approximately 300 square feet (27,7 square meters) of floor space compared to the almost 40 miles (64 km) of shelving for completed forms, which would need special housing and ongoing maintenance.

6.1.7 Adaptability of this Solution to Similar Problems

This example demonstrates how the creation of a reference archive complements digital systems. Digital technology is used to its full advantage to capture and analyse the data and to facilitate easy access in the short-term. Microfilm provides a proven, low-cost, permanent storage medium that guarantees that the reference archive will be accessible for hundreds of years to come.

6.2 Records Management of State of Virginia Land Project of Library of Virginia & County Clerk's Offices

6.2.1 User Organisations

County Clerk's Offices, Library of Virginia Records Management and Imaging Services Division.

6.2.2 Problem

The Public Records Act gives the Library responsibility for the preservation of public records in Virginia, including those pertaining to land. Historical documents under the Library's management include patents dating back to the early 1600's when the Commonwealth was a colony of Great Britain. In the late twentieth century, County Clerk's Offices began transitioning to electronic filing of land documents by use of digital imaging systems. The Library needed to find a way to fit these digital documents within its mandate to preserve records.

6.2.3 Technical Solution

Digital Image Transfer System software manages the network transfer of electronic deed books to the Library of Virginia's State Records Center from County Clerk's Offices around the state. Images are queued on a level-five RAID device and rendered to analogue records on a document archive writer. Deed books can also be uploaded via 4 mm DAT tape or compact disc.

6.2.4 Organisational Solution

Land records are scanned at County Clerk's Office as part of the recording process. They become part of a digital image base of official recorded documents integrated with a relational database and available for on-line research. The landowners keep their original records. Matching records are sent to the State Records Center in batches of less than 2,200 encrypted images. Here, they are routed into a directory, verified, and decrypted before being rendered on 16 mm microfilm for permanent storage.

6.2.5 Project Experiences

The success of this project depended to a great degree on a consortium made up of public authorities, technology specialists, and a private philanthropist.

Revisions to the Code of Virginia had made digital images legal substitutes or replacements for original documents. This facilitated the transition to electronic records management at the County level. But the Library, which was responsible for maintaining a permanent record, could find no recognised permanent media in the digital arena. By making it easy and inexpensive to transfer document images electronically for rendering to microfilm, the consortium drove archiving to the Library without placing a burden on the County Clerk's Offices. Indeed, this project allowed the Library to phased out the practice of sending field operators to the State's 95 counties to microfilm land documents on site.

6.2.6 R.O.I.

According to a published report, the Library's managers believe that this solution has reduced the cost and streamlined the process of providing trustworthy preservation copies of the valuable and historical records under their care.

6.2.7 Adaptability of this Solution to Similar Problems

This solution can be applied to a wide range of provincial and Federal applications that involved digital documents, included those captured by scanners. Distributed input can be managed and preserved at a central location for economies of scale. Agencies can maximise the effectiveness of their digital solutions confident that the records are backed up by a trustworthy, analogue reference archive that can be made accessible via a variety of present and future technologies.

7. Outlook

7.1 Digital Preservation strategy and process overview

The marriage of a sensibly-managed digital archive to a Reference Archive should provide an economical short-, mid-, and low-term solution to the availability and preservation of digital information. As records are added to the Reference Archive, their digital versions may be allowed to expire according to schedules based on a best fit to cost and accessibility, thus avoiding or mitigating the burden of perpetual migration.

Eventually, this may become a background process, part of an overall information storage management strategy. The system could identify which classes of digital documents are rendered to a Reference Archive and at what stage of their life cycles. Automating the process would assure that archiving is an integral part of information management rather than an afterthought, adding to the trustworthiness of the overall digital archive.

The physical aspects of analogue rendering would be a transparent, back-office operation. Depending on circumstances, this could be handled by a third-party service provider for economies of scale. Electronic documents no longer needed could be purged from online storage as part of a routine management schedule.

7.2 Benefits of Digital Preservation with Reference Archive protection

First and foremost, the desired outcome of a reliable, accurate record is obtained. A Reference Archive will shield government from unwarranted challenges by providing accurate, authenticated documentation. The same applies for the governed – the records of activities related to taxation, regulation, benefits, and legislation remain permanently accessible to citizens.

By avoiding the expense of migration, agencies can allocate resources to other tasks and goals. This also removes the economic barrier migration can present to moving to successive generations of information technology. Citizens get the continuing benefits of evolving information services with clear access to historical information.

The mission of the institutional archivist is thus fulfilled.

7.3 Future developments

Kodak expects content management technology to drive the next phase of records management. For reasons of cost and convenience, more interactions are moving to on-line services via kiosks and the Internet. Today records of these activities are being lost. However, they need to be managed to protect against fraud and achieve regulatory compliance. Therefore, records management and archiving will emerge as an enterprise application. The Reference Archive will provide an unchallenged repository for this information in the most authentic and trustworthy way. Analogue rendering to microfilm is a viable technology within the context of this need.

The majority of current on-line forms are static and two dimensional. These can be managed by the Reference Archive products on the market today. As the world shifts to

more complex documents, so does the need for more sophisticated products with a wider range of attributes. Future solutions must be easy to use, economical, and scalable to growing volumes. They must also address changing document demographics, including the use of colour.

7.4 Solutions availability

End-to-end Reference Archive solutions are available from a variety of commercial sources, including system integrators, independent software developers, and value-added resellers. Source document images captured by scanners and converted from various digital formats are passed to Document Archive Writers for analogue rendering on 16 mm microfilm. In some implementations, electronic images are sent from multiple distributed offices to a central location for efficient rendering, storage, and management. Retrieval requisitions are typically handled by a host document management system. Intelligent Microimage Scanners automatically search for and digitalise images for electronic display and routing.

7.5 Conclusion

Agencies interested in preserving electronic information with assured long-term availability and trustworthiness should begin creating Reference Archives.

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 two-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

Authoring Company

Kodak



Kodak is the leader in helping people take, share, enhance, preserve, print and enjoy pictures - for memories, for information, for entertainment. The company is a major participant in “infoimaging” - a \$225 billion industry composed of devices (digital cameras, production scanners and micrographics equipment), infrastructure (online networks and delivery systems for images) and services & media (software, film and paper enabling people to access, analyze and print images). Kodak harnesses its technology, market reach and a host of industry partnerships to provide innovative products and services for customers who need the information-rich content that images contain. The company, with sales last year of \$13.2 billion, is organised into four major businesses: Photography, providing consumers, professionals and cinematographers with digital and traditional products and services; Commercial Imaging, offering image capture, output and storage products and services to businesses and government; Components, delivering flat-panel displays, optics and sensors to original equipment manufacturers; and Health, supplying the healthcare industry with traditional and digital image capture and output products and services.

See <http://www.kodak.com/go/docimaging> for full details of Kodak's global & European offices and services.

Contact

Eastman Kodak Company
343 State Street
Rochester, NY 14650 - USA
Tel.: + 1 (0)716 724 4000

Contact Authoring Company

Kodak Limited
Robert M. Young
PO Box 66
Station Road,
Hemel Hempstead, Herts HP1 2TL - UK
Tel : + 44 (0)1442 844791
E-Mail: robert.m.young@kodak.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: dlm-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.

Availability & Preservation

This Industry White Paper offers Kodak's perspective on the long-term retention and availability of digital information. Digital documents require management just as their paper-based forerunners do. The electronic technologies used to create, distribute, and store them present special problems for archiving this information as time advances. Successive iterations of technology, inevitable media decay, and their inherent editability ill-suits them for long-term keeping in their native formats. A Reference Archive of permanent document images offers a cost effective long-term solution. By rendering digital information to microfilm as uncoded, analog images, organisations may create technology-proof repositories. The information stored has to be made available for decades even centuries including issues of migration and secure storage media.

ISBN 3-936534-00-4 (Series)

ISBN 3-936534-05-5