Managing Information Assets in the Public Sector Conference Edmonton, Alberta – October 28-29, 2004

# Preserving Electronic Information For Future Generations

Peter Van Garderen
President / Consultant

artefactual
peter@artefactual.com

## **Preserving Electronic Information Assets for Future Generations**

- 1. The Digital Preservation Problem
- 2. Products and Technologies
- 3. Digital Preservation Initiatives
- 4. Digital Archives Framework
- 5. Resources and Links

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **Enter the Digital Age**

- Digital information has revolutionized modern administrations and business processes
  - email, web-content, office productivity documents, database information systems, etc
- Digital information is usually more efficient, convenient,re-usable, distributable, and effective than paper-based information.





Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, (2) rtefactual

# Digital Information Will Never Survive By Accident

### A serious gap exists between:

- 1. the volume and variety of digital information that is being created
- 2. the mistaken assumption that these electronic records and digital assets will automatically be managed and preserved over time

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **Beware the Digital Dark Ages!**

- Ongoing, active management and strategic planning is required
- Organizations will lose critical and valuable digital information if they fail to plan ahead



Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactua

### Implications for the Public Sector

Digital information or e-records will be the basis for:

- confirming pension and other entitlements
- · registering births and deaths
- verifying citizenship and certifying voting rights
- enabling the collection of taxes
- supporting financial management and enabling audits and evaluations
- tracking and managing public assets (e.g. buildings, equipment, etc.)
- registering land titles
- helping resolve land claims
- documenting inter-governmental agreements
- supporting litigation
- monitoring the nation's governance, economic development, culture
- enabling countless other information-intensive activities

Long-term or permanent access to the digital information produced by these government functions is required

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### How 'long' is 'long-term'?

- "A period of time which is <u>long enough to be concerned</u> <u>about the impacts of changing technologies</u>, including support for new media and data formats, and with a changing user community, on the information being held in a repository. This period extends into the indefinite future."
  - ❖ ISO-STD 14721: Open Archival Information System

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

### **The Digital Preservation Problem**

- Physical instability and deterioration of digital storage media
  - Magnetic hard disk
  - Magnetic tape
  - Optical (CD/DVD)
  - ❖ Flash/Solid State
  - Holographic
  - no digital media is permanent (compare to durability of paper or clay tablets)

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **The Digital Preservation Problem**

- 2. Technology obsolescence and incompatibility
  - at the interacting layers of:
    - computer hardware



\* system software



\* application software



data formats



storage hardware devices



storage media

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, (a)rtefactual

### **The Digital Preservation Problem**

- 2. Technology obsolescence and incompatibility, e.g.
  - ❖ Wordstar file on 5¼ inch floppy
  - \* website that requires specific browser and plug-in versions
  - ❖ Access97 file that can't be opened with Access2003
  - 16bit application that doesn't work on a 32bit OS
  - operating system that only runs on a specific processor chipset
  - obsolete or incompatible backup tape cartridges
  - obsolete and unreadable 'archive' formats

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **The Digital Preservation Problem**

- 3. Lack of adequate metadata
  - failure to locate specific information due to lack of descriptive information,
  - inability to render and read the information due to lack of technical information,
  - inability to attribute meaning or value to the information due to the lack of contextual information,
  - inability to verify authenticity of information

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

# Can't We Just Buy Something to Deal With This?

- 1. The Digital Preservation Problem
- 2. Products and Technologies
- 3. Digital Preservation Initiatives
- 4. Digital Archives Framework
- 5. Resources and Links

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Can't We Just Buy Something to Deal With This?

- Technology, by itself, will not solve the digital preservation problem
- However, at the hardware and software layers, some of the basic technical components of the digital preservation solution are already in place
  - Electronic Information Management (EIM)
  - Enterprise Content Management (ECM)
  - Information Lifecycle Management (ILM)

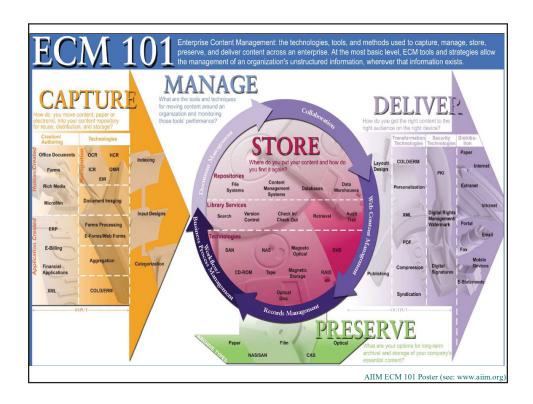
Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

# Enterprise Content Management (ECM) Industry

- Market size: big and growing
  - META Group estimates \$10 Billion by 2004 (includes web CMS market)
  - DOCULABS estimates \$4.5 Billion by 2005
  - ❖ IDC estimates \$8.3 Billion by 2006
- Major Players: ECM/Documentum, Hummingbird, IBM/Tarian, Filenet, OpenText, Vignette/TRIM, Stellent, Interwoven
- Major vendors' strategy:
  - ❖ be all things to all customers
  - as long as they're willing to pay six figures
- ECM wildcard vendor: Microsoft?

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)



# However, the 'E' in ECM doesn't really exist

- "We continue to see the majority of vendors selling point solutions targeting specific departmental problems due to the difficulty of selling and implementing enterprise-wide projects.
- As far as enterprise-wide deployment, until many of the common elements required are embedded within corporations' infrastructure it is difficult to envision how broad deployments will become the norm."
  - ❖ 'ECM in 2004' AIIM E-Doc Magazine (Nov/Dec 2003)

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

# However, the 'E' in ECM doesn't really exist

- Digital preservation initiatives will require enterprise-wide deployments
  - to leverage resources and expertise
  - to standardize on file formats, metadata and storage platforms

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

# Is ECM's 'Preserve' More Than Just Storage Technologies?

- Storage Media
  - Magnetic
  - Optical
  - ❖ Flash/Solid State
  - Holographic
  - ❖ Film
- ❖ File Systems
- Network Attached Storage
- Storage Area Networks
- \* RAID
- Grid

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### 'Preservation' or 'Storage' Technologies?

- Archiving features
  - ❖ 'Archival' is not a noun!
- Backup, Replication and Mirroring
- File System Virtualization
- Content Adressed Storage
- Data Grid technology
- Data Service tiers
- Repository Management Systems
- Open Source?
  - Dspace (www.dspace.org)
  - Jhove (http://hul.harvard.edu/jhove/)

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## Technology Alone Will Not Solve The Digital Preservation Problem

- no single vendor currently provides comprehensive solutions which fully address all of the preservation problem areas across the technological layers
- Vendors and products cannot address critical organizational issues such as:
  - ownership, responsibilities, policies, procedures, capacity, resources, funding, collobaration
- Business, user and compliance requirements will evolve over time
  - These must be mapped to the evolving technology layer:
    - \* requirements management, enterprise architecture
  - These moving targets must be managed

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## **Technology Alone Will Not Solve The Digital Preservation Problem**

- IT industry is cyclical and volatile
  - digital preservation solutions must provide permanence and consistency
- Vendors and COTS products are critical components but:
  - ❖ Beware the FUS! (Full Universal Solution)
  - Beware of anything that is 'future proof'

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## If Digital Preservation Is Not a Product, What is It?

- Anticipating and analysing technological obsolescence and threats to information integrity over the long-term
  - readability, usability, authenticity
- Devising strategies to circumvent these risks and to keep digital information usable across generations of technology
  - ❖ Technical: migration, emulation, lockss, evolution
  - Organizational: responsibilities, capacity, funding

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## If Digital Preservation Is Not a Product, What is It?

- Implementing systems, methodologies and procedures to put the strategies into action
  - will include storage, backup and 'archiving' technologies
  - \* repository architectures, file formats, metadata

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## So...Somebody is Looking After This Then, Right?

- 1. The Digital Preservation Problem
- 2. Products and Technologies
- 3. Digital Preservation Initiatives
- 4. Digital Archives Framework
- 5. Resources and Links

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **Noteworthy Collaborations**

- U.S. National Digital Information and Infrastructure Preservation Program (NDIIPP)
  - www.digitalpreservation.gov
- U.K. Digital Preservation Coalition (DPC)
  - www.dpconline.org
- International Internet Preservation Consortium (NIPC)
  - http://netpreserve.org

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

### **Noteworthy Collaborations**

- E.C. Electronic Resource Preservation and Access Network (ERPANET)
  - www.erpanet.org
- International Research on Preserving Authentic Records in Electronic Systems (InterPARES Project)
  - www.interpares.org

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### Noteworthy Institutional Projects: U.S. N.A.R.A. Electronic Records Archive

- Prototype design phase
- ERA requirements:
  - persistent, authenticity, scalable (volume and file formats), modular
  - all six elements of a record must be preserved:
    - content, context, structure, authenticity, presentation, behaviour

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## Noteworthy Institutional Projects: U.S. N.A.R.A. Electronic Records Archive

- each e-record to be stored in both native format and technology-independent format
- Broken into components which are converted to XML
- Packaged with a set of instructions that specificy how to reassemble the record components to reconstitute all six required elements
  - e.g., when an MS Excel spreadsheet is reconstituted it will not be running in Excel, but it will look more or less like it is running in Excel (presentation) and it will still show formulas and equations in mouse-over boxes (behaviour)

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Noteworthy Institutional Projects: U.K. N.A. Digital Preservation Services

- Digital Archives
  - Long-term secure storage for 'born-digital' public records and their metadata
    - documents, emails, websites, sound, video, databases
  - Operational in April 2003
  - Conforms to U.K. e-Government Interoperability Framework
  - Scalable, open (Java, XML)
  - ❖ Tape library-based system
  - Scalable to 100 terabytes

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## Noteworthy Institutional Projects: U.K. N.A. Digital Preservation Services

- Central Government Websites Archive
  - partnership with Internet Archive
  - began September 2003
  - web archiving consideration: breadth, frequency and depth
  - \* representative samples of U.K. Government websites
  - 54 different sites, seven million web pages
  - adding 225,000 new pages every week
  - websites that change rapidly, harvested once a week
  - websites that have relatively static content, harvested every six months

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Noteworthy Institutional Projects: U.K. N.A. Digital Preservation Services

#### PRONOM

- Database of information about technical dependencies of electronic records and file formats, including hardware, software and operating systems
- Launched February 2004
- Currently holds details of about 550 file formats, 250 software products, 100 vendors
- ❖ Further development:
  - ❖ File format identification tools
  - ❖ File format conversion tools

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## Noteworthy Institutional Projects: Victoria Electronic Records Strategy

- Public Records Office of Victoria (Australia)
- VERS Digital Archive
  - ❖ Pilot in Department of Infrastructure
  - Currently ingesting e-mail and word-processing documents
  - VERS Encapsulated Object (VEO)
    - ❖ Convert records to TIFF image or PDF
    - Converted file and original bitstream are bundled with XML metadata
    - ❖ Digital Signature is attached

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### But, What's Happening in Canada, eh?

- ❖ InterPARES Project (1999 2006)
- NRC-CISTI Digital Archives Research Consultation (Sept 2004)
- National Data Archive
- Canadian Initiative on Digital Libraries (CIDL)
- Forum of the National Consultation on Access to Scientific Research Data (Nov 2004)

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

### But, What's Happening in Canada, eh?

- Treasury Board: Chief Information Officer Branch
- Library and Archives of Canada
  - Information Management Capacity Check
    - MCC Lifecycle Element #5: Maintenance, Protection, and Preservation
    - \* "The extent to which the long-term usability and safeguarding of information is ensured."
  - Digital Collections 'catalytic initiative'

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Alberta Government Services' Information Management Framework

- A critical component in this framework is the need to address digital preservation issues.
  - \* "6.3 Ministries, in cooperation with the Provincial Archives, must ensure that information created by government that is of permanent and enduring value is preserved."

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

# Alberta Government Services' Information Management Framework

- Recognition that currently no single government agency has the full capacity to proceed on its own to address its digital preservation issues.
- Multiple organizations
  - Government Services, Provincial Archives of Alberta, Office of the Corporate CIO, CIO Council, Alberta Corporate Service Centre, Information and Privacy Commissioner
- Multiple sets of expertise
  - IM professionals, records management, librarians, web, IT, security, knowledge management, freedom of information and privacy

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Towards an Alberta Digital Preservation Strategy

- one day facilitated strategic planning session (March 15, 2004)
- brought together key players from across the Government to discuss the nature and extent of the issues more deeply and to determine the appropriate courses of action that should be taken.

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, (a)rtefactual

### Towards an Alberta Digital Preservation Strategy

- Recommendation: Pilot Project / Study
  - results can be adopted by other departments and agencies across the government who are searching for solutions to their own digital preservation issues.
  - Ultimately, the results are expected to contribute to a government-wide digital preservation strategy that can be applied consistently across the Government of Alberta
  - ❖ January March 2004

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Preserving Electronic Information Assets for Future Generations

- 1. The Digital Preservation Problem
- 2. Products and Technologies
- 3. Digital Preservation Initiatives
- 4. Digital Archives Framework
  - framework vs. reference model vs. architecture
  - digital archives vs. digital libraries vs. digital repositories
- 5. Resources and Links

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## How to Avoid a Digital Dark Age in Your Organization

- Reach out and collaborate with stakeholders, partners, vendors, experts
- Make a formal commitment to the ongoing stewardship of the digital information as part of long-term organizational responsibilities and budgetary obligations
  - Business case: 'public good', long-term 'access', storage management efficiency, 'nightmare scenarios'

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## How to Avoid a Digital Dark Age in Your Organization

- Implement digital preservation strategies as an integrated part of the organizational information management program/framework
  - Appraisal and retention scheduling
  - Classification, access and retrieval

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## How to Avoid a Digital Dark Age in Your Organization

- 4. Implement digital preservation strategies as an integrated part of the organizational ICT framework
  - Enterprise Architecture
  - Enterprise Data and Metadata Architecture
  - Enterprise Storage Architecture
- Implement systems, technologies, procedures, and responsibilities to manage and monitor the on-going accessibility, usability and authenticity of digital information
  - adopt/adapt/develop a digital archives framework

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **Digital Archives Framework**

- Digital Repository Architecture
- File Formats
- Metadata
- Custodial vs. Post-Custodial vs. Both

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

## Reference Model for an *Open*Archival Information System (OAIS)

- international ISO standard 14721 (January 2002)
- comprehensive logical model describing all of the functions of a digital archives
- without addressing specific technologies or archiving techniques
- archival information system: "an organization of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community."

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

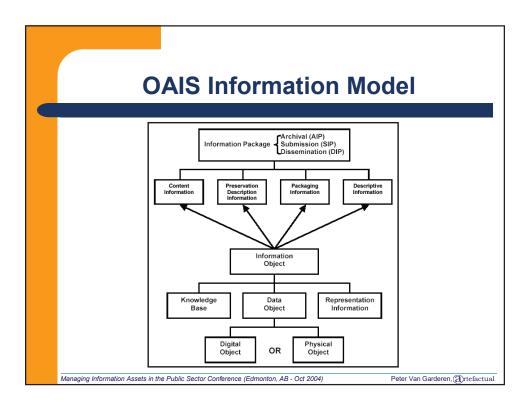
## Reference Model for an *Open*Archival Information System (OAIS)

- current de facto standard in digital preservation
- being used around the world as a common point of reference to build further understanding, consensus, and interoperability.
- Major areas of study include:
  - developing "OAIS-compliant" repository architectures
  - establishing OAIS-related standards
  - adapting the general OAIS model to domain-specific implementations
  - fleshing out the metadata requirements of the OAIS information model
  - developing methods and protocols for encoding and exchanging archived information

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

#### **OAIS Functional Model** OAIS **Preservation Planning** Data Management 0 R 0 Ν D s Archival Ingest Access Ü Storage U C М E Е R Administration MANAGEMENT Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004) Peter Van Garderen, artefactual



### **OAIS Six Mandatory Responsiblities**

- 1. Negotiate for and accept appropriate information from information producers
- 2. Obtain sufficient control of the information in order to meet long-term preservation objectives
- 3. Determine the scope of the archive's user community

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

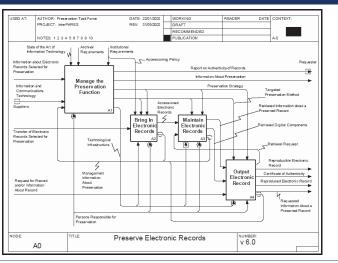
### **OAIS Six Mandatory Responsiblities**

- 4. Ensure that the preserved information is independently understandable to the user community
  - the information can be understood by users without the assistance of the information producer
- 5. Follow documented policies and procedures to ensure the information is preserved against all reasonable contingencies, and to enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the original
- 6. Make the preserved information available to the user

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

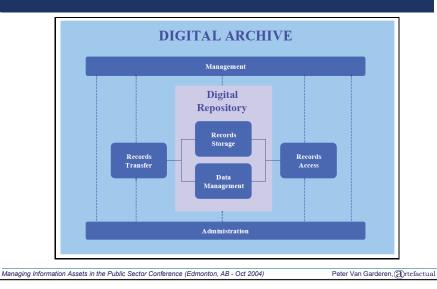
Peter Van Garderen, artefactual

### **InterPARES Preservation Model**



Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **VERS Digital Archive**



# Standard File Formats for Long-Term Preservation

- Simple Objects (e.g. word processing file, image file) vs.
- · Compound Objects (e.g. webpages) vs.
- Complex Objects (e.g. multi-tier business information systems, 'deep' websites)
- Also, structured data vs. documents
- Simple Objects
  - Text
  - Images
  - Sound
  - Moving images with Sound

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Standard File Formats for Long-Term Preservation

- It is in the vendors interest to lock users into proprietary formats (e.g. Microsoft) or to retain control over the definition of the formats (e.g. Adobe)
- Selecting File Formats for Long-Term Preservation
  - United Kingdom National Archives (June 19, 2003)
  - http://www.pro.gov.uk/about/preservation/digital/guidance/default. htm
- Primary considerations: open license, ubiquity, compression, scripts, and encryption
- LAC's File Format Guidelines

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

# **Standard File Formats** for Simple Digital Objects

- Text Documents: PDF (ISO 19005-1: PDF/Archival)
- Still Image: PNG, TIFF
  - JPEG2000? Adobe Digital Negative (DNG)? Picture Archiving and Sharing Standard (Pass) group?
- Audio: WAV or MPEG-1 layer 3 (MP3)
- Moving image: MPEG2

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

### **Metadata for Preservation**

- Dublin Core Metadata Initiative
- Metadata Encoding and Transmission Standard (METS)
- Preservation Metadata: Implementation Strategies (PREMIS)
- NISO Z39.87 Technical Metadata for Still Images

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

### **Metadata for Preservation**

- Archival Standards? (ISAD(G), RAD)
- ❖ ISO standard 23081 Metadata for Records
- InterPARES metadata schema repository
- Would you liked that Wrapped, Linked or Embedded?
  - \* Adobe Extensible Metadata Platform (XMP)
- Persistent Unique Identifiers

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

## Preserving Electronic Information Assets for Future Generations

- 1. The Digital Preservation Problem
- 2. Products and Technologies
- 3. Digital Preservation Initiatives
- 4. Digital Archives Frameworks
- 5. Resources and Links

Managing Information Assets in the Public Sector Conference (Edmonton, AB - Oct 2004)

Peter Van Garderen, artefactual

### www.library.cornell.edu/iris/tutorial/dpm/



