Building the Archives of the Future: Self-Describing Records

Kenneth Thibodeau
Director, Electronic Records Archives Program
National Archives and Records Administration
July 18, 2001
The Electronic Records Archives Vision

• Overcome technological obsolescence in a way that preserves demonstrably authentic records.
• Build a dynamic solution that incorporates the expectation of continuing change in information technology and in the records it produces.
• Find ways to take advantage of continuing progress in information technology in order to maintain and improve both performance and customer service.
Critical Challenge

• Proven methods for preserving digital information across generations of technology are limited to the simplest formats

• Available methods are increasingly inadequate

• The market has not delivered solutions.
How will we develop the Electronic Records Archives?

NARA ERA System

Electronic Records Archives Framework

Information Technology Architecture for Persistent Digital Collections
ERA Infrastructure Concept

SCALABLE
Gb/sec Internet
Grid Security
Distributed Processing
Mediation among Systems
Distributed, redundant Storage
Infrastructure Independence

Workbench
NARA User
Records Creator
Public User
Repository
Trusted Repository
NARA User
Government User
Records Creator
Digital Library
Workbench
ERA Infrastructure

- In NARA (using NARANET)
  - Archival workstations for staff
  - Reference workstations for researchers
- On the National Information Infrastructure, under NARA’s control
  - ERA Ingest & Distribution portal (Internet & Media)
  - POP repositories (Normal, Trusted, Special)
  - Affiliated Archives
- On the National Information Infrastructure
  - Agency systems with access to NARA portal
  - Digital Libraries
  - Public Users
NARA Partnerships

• **Open Archival Information System** (OAIS) Reference Model
  – NASA, Consultative Committee on Space Data Systems

• **Distributed Object Computation Testbed** (DOCT)

• **National Partnership for Advanced Computational Infrastructure** (NPACI)
  – National Science Foundation

• **Presidential Electronic Records Processing Operational System** (PERPOS)
  – Army Research Laboratory, Georgia Tech Research Institute

• **Archivist’s Workbench**
  – NHPRC Grant to San Diego Supercomputer Center

• **International research on Permanent Authentic Records in Electronic Systems** (InterPARES)
  – 7 international, multidisciplinary research teams, 10 national archives
ERA Functional Model
An Open Archival Information System Implementation

Producer

Submission Information Packages

OAIS
Archival Information Packages

Result sets
orders
queries

Dissemination Information Packages

Consumer
Information Management Architecture for Persistent Object Preservation

**Ingest Services**
- Relationships Between Concepts
- Knowledge Repository for Rules
- Rules - KQL

**Management**
- Knowledge Repository for Rules

**Access Services**
- Knowledge or Topic-Based Query / Browse

**Knowledge**
- (Topic Maps / Buckets / Model-based Access)

**Information**
- Attributes Semantics
- Information Repository
- Attribute-based Query

**Data**
- Fields Containers Folders
- Storage (Replicas, Persistent IDs)
- Feature-based Query

**.sendKeys**
- MCAT/HDF
- XML DTD
- XTMDTD
- SDLIP
- Grids

**MCAT/HDF**
- XML DTD
- XTMDTD
- SDLIP
- Grids

**Field**
- Fields Containers Folders
- Storage (Replicas, Persistent IDs)
- Feature-based Query

**Data Handling System - SRB / FTP / HTTP**
- (Data Handling System - SRB / FTP / HTTP)
ERA: Archival Components Concept

Distributed Security Infrastructure

Distributed Storage & Object Management

Accessioning Workbench
- Accession
- Verify
- Wmp & Containerize
- Describe

Archival Repository
- Collection
- Collection
- Collection
- Metadata

Reference Workbench
- Query
- Rebuild
- Present

Mediation of Information

Archival Research Catalog

Order Fulfillment System

Records Schedules

Tapes

Disks

Internet
ERA Processes

**ACCESSIONING**
- Electronic Records Collection
- Determine applicable models
- Apply models to collection
- Store Meta-data
- Store Collection Contents

**REFERENCE**
- Request for records
- Retrieve collection meta-data
- Rebuild collection structure
- Put content into collection
Accept an Accession?

- What should the agency have transferred?
- What did the agency say it transferred?
- What was transferred?
SELF-DESCRIBING

• Records
• Files
• Series
• Record Systems
Persistent Object Method

- Characterize significant properties of the things that are to be preserved.
- Express these properties in formal models.
- Encapsulate objects in metadata defined in the models.
- Use software “mediators” to enable future technologies to interpret the models and metadata:
  - to rebuild and repopulate collections
  - to re-present the records
  - support information discovery and delivery.
E-mail: Groupwise view

From: Yigal Arens <arens@ISI.EDU>
To:  
Subject: Announcing DG Online, the magazine of digital government research

Announcing DG Online, the magazine of digital government research.

DG Online <http://www.dgrc.org/dg-online/>.

DG Online: The Magazine for Digital Government Research is the new online quarterly of dg.o (DigitalGovernment.Org), a national consortium of government agencies, computer science researchers, the IT industry, and civic organizations concerned with improving online government operations and services.

DG Online presents the latest developments in advanced computer and IT research for Digital Government along with news and viewpoints on the most important DG issues:

* Cooperation among federal, state, and local government agencies
* Privacy and security
* Universal access--bridging the digital divide
* Streamlined delivery and tracking of public services
* Electronic voting, taxes, the Census, and other sensitive online data collection
* User-friendly interfaces

Date: 3/28/01 11:07PM
E-mail: text editor view

From: Ayigal Arens <arens@isi.edu>
Date: 3/28/01 11:06PM
Subject: Announcement of DG Online, the magazine of digital government research

Announcing DG Online, the magazine of digital government research.

DG Online: The Magazine for Digital Government Research is the new online quarterly of DigitalGovernment.Org, a national consortium of government agencies, computer science researchers, the IT industry, and civic organizations concerned with improving online government operations and services. DG Online presents the latest developments in advanced computer and IT research for Digital Government along with news and viewpoints on the most important DG issues.

Cooperation among federal, state, and local government agencies, privacy and security, universal access—bridging the digital divide, streamlined delivery and tracking of public services, electronic voting, taxes, the Census, and other sensitive online data collection, user-friendly interfaces, commercial and economic incentives for digital government.

And if you are interested in digital government, please consider attending dgo 2001. Join the National Science Foundation as they gather federal researchers, government agencies, industry leaders, and users to begin creating the Digital Government of the 21st Century.

Monday, May 21 -- Wednesday, May 23, Los Angeles, California.

http://www.dgrc.org/dgo2001/dgo2001, sponsored by the National Science Foundation, is the first national conference to bring together top computer science researchers, government agencies, e-commerce, the software industry, and members of the public with the aim of making online government systems available to citizens. Sessions will address vital technical, social, and economic issues in digital government for the new millennium.

For more information, please send email to dgo2001@dgrc.org. -- Ayigal Arens, Ph: 310-448-8766, Director, 4676 Admiralty Way, Fax: 310-822-0751, Intelligent Systems Division, Marina del Rey, CA 90292. mailto:arens@isi.edu http://www.isi.edu/arens
Message-Id: <p05010429b6e7e938e0b4@[10.2.68.205]>
X-Organization: USC/Information Sciences Institute
X-Phone: (310) 822-1511 ext. 766
X-Fax: (310) 822-0751
To: Readers:
From: Yigal Arens <arens@ISI.EDU>
Subject: Announcing DG Online, the magazine of
digital government
  research
Content-Type: multipart/alternative;
boundary=""---------
1226283781==_ma=""---------
|--
""---------
1226283781==_ma=""---------
Content-Type: text/plain; charset="us-ascii";
format="flowed"

DG Online <http://www.dgrc.org/dg-online/>.

DG Online: The Magazine for Digital Government
Research is the new online
<Message-Id> p05010429b6e7e938e0b4@[10.2.68.205] </Message-Id>
<X-Organization>USC/Information Sciences Institute</X-Organization>
<X-Phone>(310) 822-1511 ext. 766</X-Phone>
<X-Fax>(310) 822-0751</X-Fax>
<Date>Wed, 28 Mar 2001 22:22:30 -0500</Date>
<To>Readers:;</To>
<From>Yigal Arens {arens@ISI.EDU}</From>
<Subject>Announcing DG Online, the magazine of digital government research</Subject>
<Content-Type: multipart/alternative; boundary="============_-1226283781==_ma============"
--------=_1226283781==_maеляyahfya

<Content-Type: text/plain; charset="us-ascii"; format="flowed">
</Content-Body>
<Message_Body>DG Online {http://www.dgrc.org/dg-online/}.
....
Structure of E-mail Message
aka: Document Type Definition

<!ELEMENT Email_Message (Header, Message Body, Attachment*)>
  <!ELEMENT Header (Internal Header, External Header)>
    <!ELEMENT Internal Header (Message_Id, X-Organization, X-Phone, X-Fax)>
      <!ELEMENT Message_Id>
      <!ELEMENT X-Organization>
      <!ELEMENT X-Phone>
      <!ELEMENT X-Fax>
    <!ELEMENT External Header (Date, To, From, Subject)>
      <!ELEMENT Date (Weekday, Day_of_Month, Month, Time)>
        <!ELEMENT To (#PCDATA)+>
        <!ELEMENT From (#PCDATA)>
        <!ELEMENT Subject (#PCDATA)>*
      <!ELEMENT Message_Body (#PCDATA)>*
REPORT OF INDEPENDENT ACCOUNTANTS

To the Board of Directors and Stockholders of Great Plains Software, Inc.

In our opinion, the consolidated financial statements accompanying index present fairly, in all material respects, the financial position of Great Plains Software, Inc. and its subsidiaries at May 31, 1999 and 1998, and the results of their operations and their cash flows for each of the three years in the period ended May 31, 1999, in conformity with generally accepted accounting principles. In addition, the financial statement schedules listed in the accompanying index present fairly, in all material respects, the information set forth thereon with the related consolidated financial statements and financial statement schedules are the responsibility of management; our responsibility is to express an opinion on the statements and financial statement schedules based on our audits. We conducted our audits of these statements in accordance with generally accepted auditing standards, which require that we obtain reasonable assurance about whether the financial statements are free of material misstatement. An audit includes examining, on a test basis, evidence supporting the amounts and disclosures in the financial statements, and evaluating the overall financial statements. Our audits provide a reasonable basis for the opinion expressed above.

/s/ PricewaterhouseCoopers LLP
PricewaterhouseCoopers LLP
Minneapolis, Minnesota
June 25, 1999
REPORT OF INDEPENDENT ACCOUNTANTS

To the Board of Directors and Stockholders of Great Plains Software, Inc.

In our opinion, the consolidated financial statements listed in the accompanying index present fairly, in all material respects, the financial position of Great Plains Software, Inc. and its subsidiaries at May 31, 1999 and 1998, and the results of their operations and their cash flows for each of the three years in the period ended May 31, 1999, in conformity with generally accepted accounting principles. In addition, in our opinion, the financial statement schedules listed in the accompanying index present fairly, in all material respects, the information set forth therein when read in conjunction with the related consolidated financial statements. These financial statements and financial statement schedules are the responsibility of the Company's management; our responsibility is to express an opinion on these financial statements and financial statement schedules based on our audits. We conducted our audits of these statements in accordance with generally accepted auditing standards, which require that we plan and perform the audit to obtain reasonable assurance about whether the financial statements are free of material misstatement. An audit includes examining, on a test basis, evidence supporting the amounts and disclosures in the financial statements, assessing the accounting principles used and significant estimates made by management, and evaluating the overall financial statement presentation. We believe that our audits provide a reasonable basis for the opinion expressed above.

/s/ PricewaterhouseCoopers LLP
<table>
<thead>
<tr>
<th>Structure</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>[COMMENT]</td>
<td>SECTION: AccountantReport</td>
</tr>
<tr>
<td>ITEM</td>
<td>PricewaterhouseCoopers LLP</td>
</tr>
<tr>
<td>ITEM</td>
<td>/s/ PricewaterhouseCoopers LLP</td>
</tr>
<tr>
<td>ITEM</td>
<td>accountant_sReport.independent</td>
</tr>
<tr>
<td>ITEM</td>
<td>/s/ PricewaterhouseCoopers LLP</td>
</tr>
<tr>
<td>ITEM</td>
<td>accountant_sReport.accountantSignature</td>
</tr>
<tr>
<td>ITEM</td>
<td>Minneapolis</td>
</tr>
<tr>
<td>ITEM</td>
<td>Minnesota</td>
</tr>
<tr>
<td>ITEM</td>
<td>accountantSignature.city</td>
</tr>
<tr>
<td>ITEM</td>
<td>accountantSignature.state</td>
</tr>
<tr>
<td>ITEM</td>
<td>REPORT OF INDEPENDENT ACCOUNTANTS</td>
</tr>
<tr>
<td>ITEM</td>
<td>accountant_sReport.titleOfAccountantsReport</td>
</tr>
<tr>
<td>ITEM</td>
<td>To the Board of Directors and Stockholders of Great Plains Software,</td>
</tr>
<tr>
<td>ITEM</td>
<td>Inc.</td>
</tr>
<tr>
<td>ITEM</td>
<td>accountant_sReport.addressee</td>
</tr>
<tr>
<td>ITEM</td>
<td>June 25, 1999</td>
</tr>
<tr>
<td>ITEM</td>
<td>reportDate.date</td>
</tr>
<tr>
<td>ITEM</td>
<td>Unqualified</td>
</tr>
<tr>
<td>ITEM</td>
<td>accountant_sReport.typeOfOpinion</td>
</tr>
<tr>
<td>ITEM</td>
<td>US GAAP</td>
</tr>
<tr>
<td>ITEM</td>
<td>reportingMethod.generallyAcceptedAccountingPrinciples</td>
</tr>
<tr>
<td>ITEM</td>
<td>In our opinion, the consolidated financial statements listed in the</td>
</tr>
<tr>
<td>ITEM</td>
<td>accompanying index present fairly, in all material respects, the</td>
</tr>
<tr>
<td>ITEM</td>
<td>f...</td>
</tr>
<tr>
<td>ITEM</td>
<td>SECTION: BalanceSheet</td>
</tr>
<tr>
<td>ITEM</td>
<td>26983</td>
</tr>
<tr>
<td>ITEM</td>
<td>BS-01</td>
</tr>
<tr>
<td>ITEM</td>
<td>cashAndCashEquivalents.cashEquivalents</td>
</tr>
<tr>
<td>ITEM</td>
<td>1999-05-31</td>
</tr>
<tr>
<td>ITEM</td>
<td>18197</td>
</tr>
<tr>
<td>ITEM</td>
<td>96700</td>
</tr>
<tr>
<td>ITEM</td>
<td>48721</td>
</tr>
<tr>
<td>ITEM</td>
<td>12593</td>
</tr>
<tr>
<td>ITEM</td>
<td>8790</td>
</tr>
</tbody>
</table>
Structure expressed as Tree

Case File

- Form
- Electronic Mail Message
- Report

**Header**
- Internal Header
  - Message ID
  - X-Organization
  - X_Phone
  - X_Fax
- External Header
  - Date
  - From
  - To
  - Subject
  - Day of Week
  - Day of Month
  - Month
  - Year
  - Time

**Message Body**

**Attachment**
- Digital File
  - File Name
  - File Format
  - File Content
Accept an Accession?

Accessioning Workbench

- Transferred Records
- ?
- Expected Records
- =
- Appraisal Report
- Records Schedule
- Transfer Documentation

Expected Records = Transferred Records?
How does ERA determine the dates of records?

• E-mail
  – All e-mail contains a field indicating the date it was sent. For the sender, that is the date of the record. ERA needs to search the date-sent fields.
    – *(Technology solution)*

• Attachments to e-mail messages
  – Attachments to a record are parts of that record. The date of the message is the date of the record.
    – *(Archival principle)*

• Records forwarded, via e-mail, for filing in a recordkeeping system
  – E-mail is used only to transmit a record to the system. The date of the attached record depends on the record
    – *(Archival principle)*
Defining models for electronic records
e.g. E-mail

- All E-mail
  - Groupwise mail
  - cc:mail
  - USENET mail
    - User defined fields
  - .....
Aggregation

Accessioning

Workbench

Persistent Records

Exceptions

Container

Container
Metadata
Aggregation: risk management
Risk Management: Multi-Valent Documents
ERA: Reference Process

Request for records → Retrieve collection meta-data → Rebuild collection structure → Put content into collection

Archival Repository:
- Collection
- Collection
- Collection
- Metadata

Reference Workbench:
- Query
- Rebuild
- Present

User
Process: **Check metadata for the series to identify relevant DTDs**

Reference Workbench

- **Request for Director’s E-mail**
- **Series Metadata**
  - Files
  - Records

Metadata

Repository

Electronic Mail Message

Form

Report

Folder

File Series

+++

###

Request for Director’s E-mail
Translate E-mail DTD to Relational Database Structure

Electronic Mail Message
- Header
- Body
- Attachment *

Electronic Mail Messages
- Message Identifier
  - From
  - Subject
  - Date
  - Body

Message Addressees
- Addressee Identifier
  - Message Identifier (FK)
  - Addressee Name

Message Carbon Copies
- Carbon Copy Identifier
  - Message Identifier (FK)
  - Addressee Name

Message Attachments
- Attachment Identifier
  - Message Identifier (FK)
  - Attachment

Reference Workbench
Process: Retrieve the records and place in the target structure

Reference Workbench

Metadata

Repository
Persistent Object Preservation

+ Aims at independence of technological infrastructure
  + Reduce threats to integrity and authenticity by minimizing changes over time.
+ Embeds changes in a comprehensive information management architecture designed for preservation
+ Inherently extensible
+ Facilitates use of future, advanced technologies, without requiring change in what is preserved.
  - Currently beyond state of the art of the art of information technology.
Self-describing Objects for Records Management

- Facilitate management, exchange, and disposition of records
  - explicitly identify the content of records, files, series,...
  - express how content is organized
  - allow the content to be stored once and used in different documents
  - separate, but link, management of content and presentation
  - capture the relationships among documents and collections of documents
  - and support multiple views of a collection of documents
  - all in plain language
Thank you.

For more information:

www.nara.gov/era