“C2P”: from Content-based Environments to Policy-based Infrastructure

Richard Marciano

Sustainable Archives & Library Technologies (SALT) lab director
Data Intensive Cyber Environment (DICE) group
&
Professor @ SILS / UNC Chapel Hill
Chief Scientist for Persistent Archives and Digital Preservation @ RENCI

marciano@unc.edu

http://www.DiceResearch.org
JISC: Digital Preservation Policies Study

- Oct. 30, 2008
  http://www.jisc.ac.uk/publications/publications/jiscpolicyfinalreport.aspx
- **Last decade:** harnessing digital content and electronic services
- **In recent years:** priority on developing e-strategies and infrastructure to underpin electronic access
- **In moving forward:** long-term access is heavily dependent on digital preservation strategies being in place and underpinned by relevant policy and procedures
Lifecycle Approach to Managing Records
Life-cycle Approach to the Management and Preservation of e-mail Messages

• 1999 experiment w. NARA using postings from Usenet groups
• Challenge: manage 1 million e-mail messages
  • We used Usenet messages w.o. attachments: 2.5 GB
  • RFC 1036 (interchange of network news messages – not RFC 2822 e-mail standard)
  • XML images of the e-mail records based on an e-mail DTD
Does anyone know where I can get passenger lists for ships that transported Spaniards to Cuba circa 1860's? Any help would be appreciated.

Thanks,
Cheryl Sanchez-Sivers

Casivers@aol.com

Subject: Passenger Lists for Ships from Spain To Cuba
From: Casivers@aol.com
Newsgroups: soc.genealogy.hispanic
Organization: RootsWeb Genealogical Data Cooperative
Lines: 7
Message-ID: <2376321.36f6de03@aol.com>
Content-Type: text/plain; charset=US-ASCII
Content-Transfer-Encoding: 7bit
X-Trace: bl1.rootsweb.com 922148437 3147 127.0.0.1 (23 Mar 1999)
X-Complaints-To: usenet@rootsweb.com
NNTP-Posting-Date: 23 Mar 1999 00:02:37 GMT
Xref: news.sdsc.edu soc.genealogy.hispanic:3156
DTD for E-mail messages

```xml
<!ELEMENT rfc1036_mesg (headers, body)>  
<!ELEMENT headers (required_headers, optional_headers, other_headers)>  
<!ELEMENT body #PCDATA>  
<!ELEMENT required_headers (From, Date, Newsgroups, Subject, Message-ID, Path)>  
<!ELEMENT optional_headers (Folloup-To?, Expires?, Reply-To?, Sender?, References?, Control?, Distribution?, Keywords?, Summary?, Approved?, Lines?, Xref?, Organization?)>  
<!ELEMENT other_headers other+>  

<!-- 6 required header keywords -->  
<!ELEMENT From #PCDATA>  
<!ATTLIST From seqno CDATA #REQUIRED>  
<!ELEMENT Date #PCDATA>  
<!ATTLIST Date seqno CDATA #REQUIRED>  
<!ELEMENT Newsgroups #PCDATA>  
<!ATTLIST Newsgroups seqno CDATA #REQUIRED>  
<!ELEMENT Subject #PCDATA>  
<!ATTLIST Subject seqno CDATA #REQUIRED>  
<!ELEMENT Message-ID #PCDATA>  
<!ATTLIST Message-ID seqno CDATA #REQUIRED>  
<!ELEMENT Path #PCDATA>  
<!ATTLIST Path seqno CDATA #REQUIRED>  
```
<!-- 13 optional header keywords -->

<!ELEMENT Followup-To #PCDATA>  
<!ATTLIST Followup-To seqno CDATA #REQUIRED>  

<!ELEMENT Expires #PCDATA>  
<!ATTLIST Expires seqno CDATA #REQUIRED>  

<!ELEMENT Reply-To #PCDATA>  
<!ATTLIST Reply-To seqno CDATA #REQUIRED>  

<!ELEMENT Sender #PCDATA>  
<!ATTLIST Sender seqno CDATA #REQUIRED>  

<!ELEMENT References #PCDATA>  
<!ATTLIST References seqno CDATA #REQUIRED>  

<!ELEMENT Control #PCDATA>  
<!ATTLIST Control seqno CDATA #REQUIRED>  

<!ELEMENT Distribution #PCDATA>  
<!ATTLIST Distribution seqno CDATA #REQUIRED>  

<!ELEMENT Keywords #PCDATA>  
<!ATTLIST Keywords seqno CDATA #REQUIRED>  

<!ELEMENT Summary #PCDATA>  
<!ATTLIST Summary seqno CDATA #REQUIRED>  

<!ELEMENT Approved #PCDATA>  
<!ATTLIST Approved seqno CDATA #REQUIRED>  

<!ELEMENT Lines #PCDATA>  
<!ATTLIST Lines seqno CDATA #REQUIRED>  

<!ELEMENT Xref #PCDATA>  
<!ATTLIST Xref seqno CDATA #REQUIRED>  

<!ELEMENT Organization #PCDATA>  
<!ATTLIST Organization seqno CDATA #REQUIRED>  

<!-- other header keywords -->

<!ELEMENT other #PCDATA>  
<!ATTLIST other keyword CDATA #REQUIRED  
seqno CDATA #REQUIRED>
<table>
<thead>
<tr>
<th>Collection Size:</th>
<th>2.52 GB</th>
</tr>
</thead>
<tbody>
<tr>
<td># files:</td>
<td>1 Million</td>
</tr>
<tr>
<td># records/file:</td>
<td>6 req., 13 opt., variable other</td>
</tr>
<tr>
<td>Time to Assemble collection: Newsgroup→Collection.raw</td>
<td>12 hours</td>
</tr>
<tr>
<td>Time to convert: Collection.raw→Collection.xml</td>
<td>1 h 39 min</td>
</tr>
<tr>
<td>Time to archive: Collection.raw→HPSS</td>
<td>1 h 02 min</td>
</tr>
<tr>
<td>Time to archive: Collection.xml→HPSS</td>
<td>1 h 29 min</td>
</tr>
<tr>
<td>TOTAL TIME FOR assembling-tagging-archiving</td>
<td>16 hours</td>
</tr>
</tbody>
</table>

- 40 million e-mails can be processed in 3 weeks
- (10 Mbit Ethernet, old SGI server, tape-based archive writing at 2MB/s)
- technology now runs at 160MB/s (80 times faster) – could probably be done in ½ day

State of North Carolina (Steve Burbeck): Parse speed: about a 250 MB/hour on a Thinkpad T40
Recent Collaborations on Preservation
(NARA, NHPRC, LOC, NSF, IMLS)

**NARA:** 1998-2007, *NARA - U Md, GTech, SLAC, UC Berkeley*
- Transcontinental Persistent Archive Prototype based on data grids.

**IP2:** 2002-2006, *NHPRC/SSHRC/NSF - UBC and others*
- InterPARES 2 collaboration with UBC on infrastructure independence

**PERM:** 2002-2004, *NHPRC - Michigan, SDSC*
- Preservation of records from an RMA. Interoperability across RMAs.

**LoC:** 2003-2004, *LoC - SDSC, LOC*
- Evaluation of use of SRB for storing American Memory collections

**ICAP:** 2003-2006, *NHPRC - UCSD,UCLA,SDSC*
- Exploring the ability to compare versions of records, run historical queries

**A&W:** 2000-2003, *NHPRC - SDSC*
- Methodologies for preservation & access of software-dependent electronic records

**DIGARCH:** 2005-2007, *NSF - UCTV,Berkeley,UCSD Libraries,SDSC*
- Preservation of video workflows

**eLegislature:** 2005-2007, *NSF - Minnesota, SDSC*
- Preserving the records of the e-Legislature

**VanMAP:** 2005-2006, *UBC - UBC,Vancouver*
- Preserving the GIS records of the city of Vancouver

**eLegacy:** 2006-2008, *NHPRC - California*
- Preserving the geospatial data of the state of California

**T-RACES:** 2006-2008, *IMLS - UCHRI,SDSC*
- California’s redlining archives testbed

**PAT:** 2004-2007, *NHPRC - Mi,Mn,Ke,Oh,Slac,SDSC*
- Demonstration of a cost-effective system for preserving electronic records.
PAT: Persistent Archives Testbed

OBES e-mail
“All-In-One” system w.
- 300,000+ messages / 96 folders / 1.1 GB

Local Storage Resources

- Kentucky Grid Brick
- Michigan Grid Brick
- Minnesota Grid Brick
- Ohio Grid Brick
- SLAC Storage

Metadata Catalog (Oracle)

Archival Storage (HPSS, Sam-QFS)

Shared Preservation Environment

Archive

MCAT
NARA - TPAP
Transcontinental Persistent Archive Prototype

Federation of Seven Independent Data Grids

Extensible Environment, can federate with additional research and education sites. Each data grid uses different vendor products.

Document Collections Summary

1973-1974 Telegrams Transferred to NARA

State Department Collections

- US-UK Communications on Iraq, Summer 2002
- The President’s Emergency Plan for AIDS Relief
- Katrina Document Collection
- The Future of Iraq Project
- Kissinger Telephone Transcripts
- FOIA Released Documents
- Allegations of Drug Trafficking in L.A.
- Argentine Declassification Project (1975-1984)
- State Chile Declassification Project Tranche I (1972-1975)
- State Chile Declassification Project Tranche II (1968-1972)
- State Chile Declassification Project Tranche III (1970-1991)
- CIA Creation Documents
- El Salvador Churchwomen Documents
- Guatemala Collection
- Reuel Wallenberg
- El Salvador Collection
- Amelia Earhart Collection
- International Agreements Collection
- International Traffic in Arms Regulations (ITARs)

Argentina Declassification Press Releases/Statements

- 03/20/2002 State Department Press Release

Chile Declassification Press Releases/Statements

- 06/30/1999 State Department Press Release
- 10/08/1999 State Department Press Release
- 11/13/2000 State Department Press Statement
- 06/30/1999 White House Press Statement
- 10/08/1999 White House Press Statement
### Amelia Earhart Collection (63 records)

#### Standard Results List

**Collection Searched:** Amelia Earhart Collection

Your Query "" matched 63 out of 63.
50 documents displayed.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>THE GOVERNMENT ANNOUNCED THAT ERMINE AND MR. MURPHY DIED IMMEDIATELY AFTER A TRAGIC AIRCRAFT ACCIDENT&lt;br&gt;<strong>Date:</strong> 04-19-1937&lt;br&gt;<strong>To:</strong> EDWARD R. KEICHAR From: FREDERICK AAMOHL&lt;br&gt;<strong>Document Type:</strong> LETTER Length: 2 pp.&lt;br&gt;<strong>Collections:</strong> Amelia Earhart Collection</td>
</tr>
<tr>
<td>2</td>
<td>FORWARDING INQUIRY TO DEPARTMENT OF INTERIOR&lt;br&gt;<strong>Date:</strong> 02-15-1977&lt;br&gt;<strong>To:</strong> THE HONORABLE HAROLD S. HINSON From: KEMPSON &amp; HINSON&lt;br&gt;<strong>Document Type:</strong> LETTER Length: 1 pp.&lt;br&gt;<strong>Collections:</strong> Amelia Earhart Collection</td>
</tr>
<tr>
<td>3</td>
<td>FILE OF AMELIA EARHART OPEN TO PUBLIC&lt;br&gt;<strong>Date:</strong> 11-30-1976&lt;br&gt;<strong>To:</strong> DEAN KOPPEN From: FREDERICK MANDAHL&lt;br&gt;<strong>Document Type:</strong> LETTER Length: 2 pp.&lt;br&gt;<strong>Collections:</strong> Amelia Earhart Collection</td>
</tr>
<tr>
<td>4</td>
<td>VISIT OF MRS. MORRISSEY, MRS. AMELIA EARHART PUTNAM'S SISTER, CAME TO SEE ME&lt;br&gt;<strong>Date:</strong> 05-21-1976&lt;br&gt;<strong>To:</strong> FOR THE FILE From: FREDERICK MANDAHL&lt;br&gt;<strong>Document Type:</strong> MEMORANDUM Length: 1 pp.&lt;br&gt;<strong>Collections:</strong> Amelia Earhart Collection</td>
</tr>
<tr>
<td>5</td>
<td>COMMENT ONLY ON THE STATE DEPARTMENT ASPECTS&lt;br&gt;<strong>Date:</strong> 08-27-1969&lt;br&gt;<strong>To:</strong> HONORABLE GEORGE MURPHY From: WILLIAM S. MACOMBER&lt;br&gt;<strong>Document Type:</strong> LETTER Length: 1 pp.&lt;br&gt;<strong>Collections:</strong> Amelia Earhart Collection</td>
</tr>
<tr>
<td>6</td>
<td>THANKS FOR YOUR HELPFUL RESPONSE TO MY INQUIRY&lt;br&gt;<strong>Date:</strong> 05-27-1968&lt;br&gt;<strong>To:</strong> DR. RICHARDSON O'DOUGALL From: RICHARD J. LEOPOLD, CHAIRMAN&lt;br&gt;<strong>Document Type:</strong> LETTER Length: 2 pp.&lt;br&gt;<strong>Collections:</strong> Amelia Earhart Collection</td>
</tr>
</tbody>
</table>
iRODS Starting Point: empty
<table>
<thead>
<tr>
<th>Name</th>
<th>Resource</th>
<th>Size</th>
<th>Date Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>RG 001</td>
<td>Records of the War Labor Policies Board</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 002</td>
<td>Records of the National War Labor Board (World War I)</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 003</td>
<td>Records of the U.S. Housing Corporation</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 004</td>
<td>Records of the U.S. Food Administration</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 005</td>
<td>Records of the U.S. Grain Corporation</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 006</td>
<td>Records of the U.S. Sugar Exportation Board, Inc.</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 007</td>
<td>Records of the Bureau of Enrolled and Plant Quarantine</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 008</td>
<td>Records of the Bureau of Agricultural Engineering</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 009</td>
<td>Records of the National Recovery Administration</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 010</td>
<td>Records of the National Commission on Law Observance and Enforcement</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 011</td>
<td>General Records of the United States Government</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 012</td>
<td>Records of the Office of Education</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 013</td>
<td>Records of the National Mediation Board</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 014</td>
<td>Records of the U.S. Railroad Administration</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 015</td>
<td>Records of the Department of Veterans Affairs</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 016</td>
<td>Records of the Office of the Secretary of Agriculture</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 017</td>
<td>Records of the Bureau of Animal Industry</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 018</td>
<td>Records of the Army Air Forces</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 019</td>
<td>Records of the Bureau of Ships</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 020</td>
<td>Records of the Office of the Special Adviser to the President on Foreign Trade</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 021</td>
<td>Records of District Courts of the United States</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 022</td>
<td>Records of the U.S. Fish and Wildlife Service</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 023</td>
<td>Records of the Coast and Geodetic Survey</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 024</td>
<td>Records of the Bureau of Naval Personnel</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 025</td>
<td>Records of the National Labor Relations Board</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 026</td>
<td>Records of the U.S. Coast Guard</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
<tr>
<td>RG 027</td>
<td>Records of the Weather Bureau</td>
<td>July 8, 2008, 8:41 pm</td>
<td></td>
</tr>
</tbody>
</table>
Record 00001.pdf with Metadata

Metadata: 00001.pdf

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>To</td>
<td>EDWARD R. REICHAK</td>
</tr>
<tr>
<td>Subject</td>
<td>THE GOVERNMENT ASSUMES...</td>
</tr>
<tr>
<td>Original File</td>
<td><a href="http://foia.state.gov/documents">http://foia.state.gov/documents</a>...</td>
</tr>
<tr>
<td>Length</td>
<td>2 pp.</td>
</tr>
<tr>
<td>From</td>
<td>FREDRIK KANDAH</td>
</tr>
<tr>
<td>Document Type</td>
<td>LETTER</td>
</tr>
<tr>
<td>Date</td>
<td>04-19-1977</td>
</tr>
<tr>
<td>Collection</td>
<td>Amelia Earhart Collection</td>
</tr>
</tbody>
</table>

Resource | Size | Date Modified
---|---|---
pzoneResc | 129.33 KB | July 12, 2008, 5:32 pm
pzoneResc | 72.85 KB | July 12, 2008, 5:32 pm
pzoneResc | 65.01 KB | July 12, 2008, 5:32 pm
pzoneResc | 34.04 KB | July 12, 2008, 5:32 pm
pzoneResc | 69.75 KB | July 12, 2008, 5:32 pm
pzoneResc | 59.74 KB | July 12, 2008, 5:32 pm
pzoneResc | 44.04 KB | July 12, 2008, 5:32 pm
pzoneResc | 39.61 KB | July 12, 2008, 5:32 pm
pzoneResc | 82.38 KB | July 12, 2008, 5:32 pm
pzoneResc | 61.35 KB | July 12, 2008, 5:32 pm
pzoneResc | 33.19 KB | July 12, 2008, 5:32 pm
pzoneResc | 28.17 KB | July 12, 2008, 5:32 pm
pzoneResc | 19.98 KB | July 12, 2008, 5:32 pm
pzoneResc | 23.19 KB | July 12, 2008, 5:32 pm
pzoneResc | 17.74 KB | July 12, 2008, 5:32 pm
pzoneResc | 7.14 KB | July 12, 2008, 5:32 pm
pzoneResc | 7.51 KB | July 12, 2008, 5:32 pm
pzoneResc | 301.27 KB | July 12, 2008, 5:32 pm
pzoneResc | 120.86 KB | July 12, 2008, 5:32 pm
pzoneResc | 9.21 KB | July 12, 2008, 5:32 pm
pzoneResc | 172.36 KB | July 12, 2008, 5:32 pm
pzoneResc | 174.19 KB | July 12, 2008, 5:32 pm
pzoneResc | 237.33 KB | July 12, 2008, 5:32 pm

Displaying objects 1 - 50 of 63
DCAPE: Distributed Custodial Archival Preservation Environments

Purpose: http://dcape.org

Build a distributed production preservation environment that meets the needs of archival repositories for trusted archival preservation services

Distributed partnership of 35 participants across 11 institutions:

* STATES:
  - California
  - Kansas
  - Michigan
  - Kentucky
  - North Carolina: Kelly Eubank, David Minor
  - New York

* UNIVERSITIES:
  - Tufts University
  - West Virginia University
  - Carleton University (Geomatics and Cartographic Research Centre), Canada
  - U. of North Carolina @ Chapel Hill (SILS/RENCI) Richard Marciano, Reagan Moore

* CULTURAL ENTITIES:
  - Getty Research Institute
Preservation Services

• automated conversion of records to a required archival form (standard encoding format),
• automated extraction of descriptive metadata based on templates that describe the information structure within a class of records
• automated verification of assessment criteria (RLG/NARA TRAC - Drambora, etc.)
• automated ingestion of records through periodical synchronization with a submission source
• etc.
Rule-based Data Management

• Associate Rules with Combinations of:
  • Data Objects
  • Collections
  • User Groups
  • Storage Systems

• For Example:
  • Particular User Groups when Accessing a Particular Collection
For More Information

Richard Marciano & Reagan Moore
DICE Group
marciano@unc.edu & rwmoore@email.unc.edu

http://www.DiceResearch.org