Everyone is a Curator: Human-Assisted Preservation for ORE Aggregations

Frank McCown
Harding University
Computer Science Dept.
Searcy, Arkansas, USA

Michael L. Nelson
Old Dominion University
Computer Science Dept.
Norfolk, Virginia, USA

Herbert Van de Sompel
Digital Library Research & Prototyping Team
Los Alamos National Lab
Los Alamos, NM, USA

DigCCurr 2009
Chapel Hill, NC
April 3, 2009

OAI-ORE work is supported by the Andrew W. Mellon Foundation, Microsoft Corporation, and the National Science Foundation (IIS-0430906).
Outline

• Aggregations on the Web
• Object Reuse and Exchange (ORE)
• Web Infrastructure (WI)
• Prototype system: ReMember
• Future work
Aggregations
click here to add a description

32 photos | 3 views | Add a comment

items are from between 29 Feb 2008 & 01 Mar 2008.
March/April 2009

Vol. 15 No. 3/4
doi:10.1045/dlib.magazine
ISSN: 1082-9873

Table of Contents

EDITORIAL
Some Results from the 2008 D-Lib Magazine Survey
by Bonita Wilson, CNRI

LETTERS
To the Editor

OPINION
What's Wrong with Citation Counts?
José H. Canós Cerdá and Manuel Llavador Campos, Technical University of Valencia, Spain; and Eduardo Mena Nieto, University of Zaragoza, Spain
doi:10.1045/march2009-canos

ARTICLES
Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images
Tracy Powell and Gordon Paynter, National Library of New Zealand
doi:10.1045/march2009-powell

How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs
by Rose Holley, National Library of Australia
doi:10.1045/march2009-holley
Aggregation spans multiple web sites
Aggregation composed of multiple representations
Robot’s view of web aggregations

Web sites

Web pages

Web resources

Why GDC matters more than ever
1 hour, 22 minutes ago
With the recession looming over
Digital Preservation (Library of Congress)
Home page for the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress, a collaborative effort dedicated ...
About the Program - Sustainability of Digital Formats - Partners
www.digitalpreservation.gov/ - 15k - Cached - Similar pages

Sustainability of Digital Formats: Planning for Library of ...
May 21, 2007 ... The Digital Formats Web site provides information about digital content formats.
www.digitalpreservation.gov/formats/ - 9k - Cached - Similar pages
More results from www.digitalpreservation.gov »

Digital preservation - Wikipedia, the free encyclopedia
Feb 23, 2009 ... Digital preservation is the management of digital information over time. Preservation of digital information is widely considered to require ...
en.wikipedia.org/wiki/Digital_preservation - 82k - Cached - Similar pages

Digital Preservation
DIGARCH funds cutting-edge research in digital preservation with emphasis on digital repository models; tools, technologies, and processes; ...
www.diglib.org/preserve.htm - 32k - Cached - Similar pages

Digital Preservation Management Resources
The Inter-university Consortium for Political and Social Research (ICPSR) is pleased to be the host institution for the Digital Preservation Management ...
www.icpsr.umich.edu/dpm/ - 5k - Cached - Similar pages

Digital Collection Management [OCLC - Products and Services]
A central registry of digital preservation masters. Search for, and find, digitally preserved
OAI-ORE solution

• OAI-ORE provides standards to
  – Describe an aggregation of web resources
  – Provide boundaries for an aggregation
  – State the relationships between aggregated resources

• OAI-ORE uses existing web standards and technologies

• Example application: improve web search
OAI-ORE solution

• Resource Map (ReM) describes the aggregation
  – Has a URI (new resource)
  – May express relationships and properties pertaining to all aggregated resources
  – Contains metadata about the ReM (e.g., who created it, when it was last modified, etc.)
  – Machine readable (serialized as Atom XML, RDF/XML, RDFa, n3, turtle, and other RDF serialization formats)
Example ReM

http://www.openarchives.org/ore/1.0/primer.html
Atom XML ReM Serialization Example

```xml
<?xml version="1.0" encoding="UTF-8"?>
<atom:entry
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:atom="http://www.w3.org/2005/Atom"
    xmlns:ore="http://www.openarchives.org/ore/terms/"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">

    <!-- About the Aggregation for the ArXiv document -->

    <!-- This Atom entry describes a specific ORE Aggregation -->
        rel="http://www.openarchives.org/ore/terms/describes"/>
    <atom:category term="http://www.openarchives.org/ore/terms/Aggregation"
        scheme="http://www.openarchives.org/ore/terms/" label="Aggregation"/>
    <!-- The Aggregation aggregates ... -->
        rel="http://www.openarchives.org/ore/terms/aggregates"
        title="[astro-ph/0601007] Parametrization of K-essence and Its Kinetic Term"
        type="text/html"/>
        rel="http://www.openarchives.org/ore/terms/aggregates"
        title="Parametrization of K-essence and Its Kinetic Term" type="application/pdf"
        hreflang="en"/>
        rel="http://www.openarchives.org/ore/terms/aggregates"
        title="Parametrization of K-essence and Its Kinetic Term" type="application/postscript"
        hreflang="en"/>

    <!-- Metadata about the Aggregation: title and authors -->
    <atom:title>Parametrization of K-essence and Its Kinetic Term</atom:title>
    <atom:author>
        <atom:name>Hui Li</atom:name>
        <atom:email>lihui@somewhere.cn</atom:email>
    </atom:author>
</atom:entry>
```

http://www.openarchives.org/ore/1.0/primer.html#Serialization
Black hat: http://img.webpronews.com/securitypronews/110705blackhat.jpg
Virus image: http://polarboing.com/images/topics/misc/story.computer.virus_1137794805.jpg
Web Infrastructure
WI capture of web page
Warrick: Crawling the crawlers to reconstruct entire web sites

http://warrick.cs.odu.edu/
Web $\xrightarrow{Pull}$ WI
Pre-ORE
Pre-ORE

ORE
Can we harness the web community to curate ReMs?

- Anyone can create a ReM
- ReM’s aggregated resources may change
- Aggregated resources may go 404
- Aggregated resources may move location
- Humans are better at resolving these problems than computers
Many people doing a little work = curation on a grand scale
Original URL is 404:
ReMember Prototype

Resource Not Found

Digital Archiving - Developing Policy And Best Practice Guidelines At The National Library Of Australia
author: Gatenby, Pam

Please help us find a current version of the missing resource above using the search engine of your choice. Click on a search engine below to begin your search. Use any keywords you think useful to perform your search. We suggest the title of the resource, the author, or any other metadata that is available.

1. Google
2. Yahoo
3. Live Search

Did you find a current version of the missing resource?  Yes  No  I'm not sure

Done
None of these results are a proper replacement.
See if this is the correct resource
Digital Archiving — Developing Policy And Best Practice Guidelines At The National Library Of Australia

By Pam Gatenby, National Library of Australia

1. Introduction

The main focus of my presentation is the business model the National Library of Australia has put in place to archive digital on-line publications. I’ll explain the strategic framework and rationale for the approach we are currently taking; the processes we follow and then outline areas we are addressing to ensure cost-effective and workable solutions to providing long-term access to digital publications.

Like most national libraries, the National Library of Australia has responsibility for collecting, preserving and making available the published national output. It accepts that this responsibility extends to digital publications in all forms.

We have been archiving Internet Australian publications since 1996 when the PANDORA project was established. PANDORA stands for Preserving and Accessing Networked Documentary Resources of Australia. Through this project we have been working at two levels to develop our strategy for providing long-term access to Australian on-line publications — at the conceptual level to establish the standards, technical, and business requirements for a durable, robust archive; and at the practical level to actually start building an archive of titles. This now numbers 450 titles which includes some very complex ones which have provided us with significant technical challenges, and many which are gathered on an on-going basis.

At the Library we attach great importance and value to the "hands-on" approach and to collaborative effort. After an initial period of intensive discussion and grappling with issues back in 1995, we decided that the only way to make progress was to actually start doing something in order to learn through experience and to develop our ideas further. The progress that the PANDORA project has made since that time is a tribute to the intelligence, perseverance and dedication of a small team of staff who have worked closely together on many difficult issues. PANDORA was launched based on a set of basic software tools developed in-house, and a modest level of staffing funded from recurrent budget. This approach has had some inherent operational frustrations but has allowed us to test and refine our theoretical principles.

2. Strategic framework for digital archiving implementation

The Library’s overall approach to managing the archiving and preservation of digital publications and to providing access to them, takes place within a
What if Resource is found on “live” Web?

1. Update ReM to point to new location
2. Push live version to WebCite

Digital Archiving — Developing Policy And Best Library Of Australia

By Pam Gatenby, National Library of Australia

1. Introduction

The main focus of my presentation is the business model the National Library of Australia's Strategic Framework and rationale for the approach we are currently taking to ensure cost-effective and workable solutions to providing long-term access to digitised materials. Like most national libraries, the National Library of Australia has responsibility for its output. It accepts that this responsibility extends to digital publications in all forms.

We have been archiving Internet Australian publications since 1996 when the PAN Accessing Networked Documentary Resources of Australia project was funded to promote access to Australian scholarly and educational content.
Wiki for revision control
Resource Map Visualization

Key:

1. ◦ Resource is archived for the first time.
2. ● Resource has been significantly changed and re-archived.
3. ○ Resource has changed insignificantly.
4. ▲ Resource has moved to a new URL.
5. ● Resource is no longer accessible on the live Web.
Future Work

• Waiting for ORE to catch on (Version 1.0 just released Oct 2008)
• Will user community adopt ReMember?
• How do we prevent or detect malicious alterations?
• Allow public to add and remove Resources from ReMs?
Thank You
Frank McCown
fmccown@harding.edu
http://www.harding.edu/fmccown/

OAI-ORE work is supported by the Andrew W. Mellon Foundation, Microsoft Corporation, and the National Science Foundation (IIS-0430906).
Denver Broncos
ReM
Microsoft-Yahoo Merger ReM