Shared Infrastructure Preservation Models

PI: Michael L. Nelson

Co-PI: Johan Bollen

Old Dominion University

www.cs.odu.edu/~{mln,jbollen}

DIGARCH PI Meeting

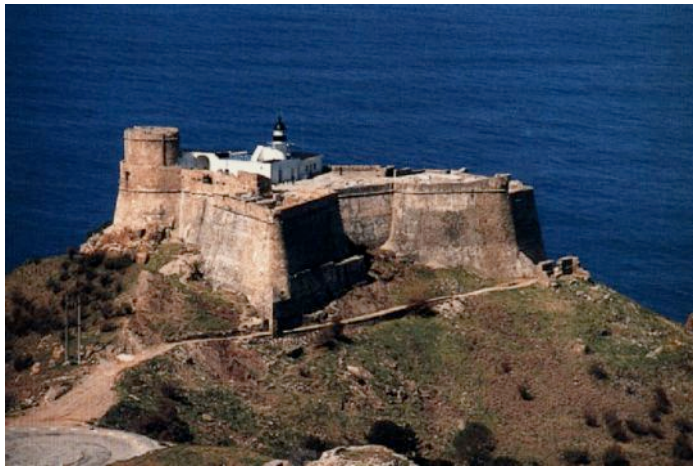
dg.o 2005 Meeting

Atlanta, GA May 17 2005

Preservation: Fortress Model

Five Easy Steps for Preservation:

1. Get a lot of $
2. Buy a lot of disks, machines, tapes, etc.
3. Hire an army of staff
4. Load a small amount of data
5. “Look upon my archive ye Mighty, and despair!”



Ex. $14M NDIIPP Partnerships
http://www.digitalpreservation.gov/about/pr_093004.html

image from: http://www.itunisie.com/tourisme/excursion/tabarka/images/fort.jpg

How Long is Forever?

- Average human life span (from: http://www.che.uc.edu/acs/archives/cintacs/vol39no5/vol39no5.html)
  - female: 78
  - male: 77
- Average Fortune 500 company lifespan: (from: http://www.businessweek.com/chapter/degeus.htm)
  - 40 - 50 years
- Universities & Research Institutes?
    - Wang Institute http://csdl.computer.org/comp/mags/co/1989/05/r5078abs.htm
    - ICASE http://www.icase.edu/
    - Marycrest International U. http://www.mcrest.edu/
    - Ambassador U. http://www.ambassador.edu/
- U.S. Government agency or institution?
  - Federal Laboratory Reforms
    - http://clinton2.nara.gov/WH/EOP/OSTP/NSTC/html/pdd5status.html
  - NASA Zero Base Review
    - http://www.nasawatch.com/archives/2005/04/calvert_looks_a.html
    - http://www.hq.nasa.gov/office/pao/97budget/statement.txt
  - U.S. Military BRAC
    - http://www.globalsecurity.org/military/facility/brac.htm
    - http://www.defenselink.mil/brac/

Preservation: P2P* Model

(= Peter-to-Paul, Ponzi-Pyramid)*

Three Easy Steps for Preservation:

1. Convince a user to give up 100 MB of local storage for 3 years
2. Guarantee the user 10 MB of storage in perpetuity
3. Add more users!

Charles Ponzi

Examples: Intermemory, Free Haven, Freenet, PAST

image from: http://www.innovationodyssey.com/images/Ponzi.jpg

Shared Infrastructure Preservation Models

- Something between “Fortress” and “P2P”?
  - fewer “heroes”
  - limited $ resources
  - increase sustainability by leveraging software/protocols/environments
- Study feasibility of exporting institutional repository contents with:
  - SMTP (email)
  - IP multicasting
  - NNTP (Usenet News)

OAI-PMH Data Model / Complex Objects

resource

OAI-PMH identifier
= entry point to all records pertaining to the resource

item

metadata pertaining to the resource

Dublin Core metadata	MARCXML metadata	MPEG-21 DIDL	METS

records

simple	more expressive	highly expressive	highly expressive

Ideas first presented in Van de Sompel, Nelson, Lagoze & Warner,
http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html

SI: SMTP

- Instrument sendmail / procmail to:
  - attach (by-ref or by-value) individual records
  - baseURLs of the institution's repositories
  - feed results to an institution's harvester
- Premise: discover repositories based on members' access patterns

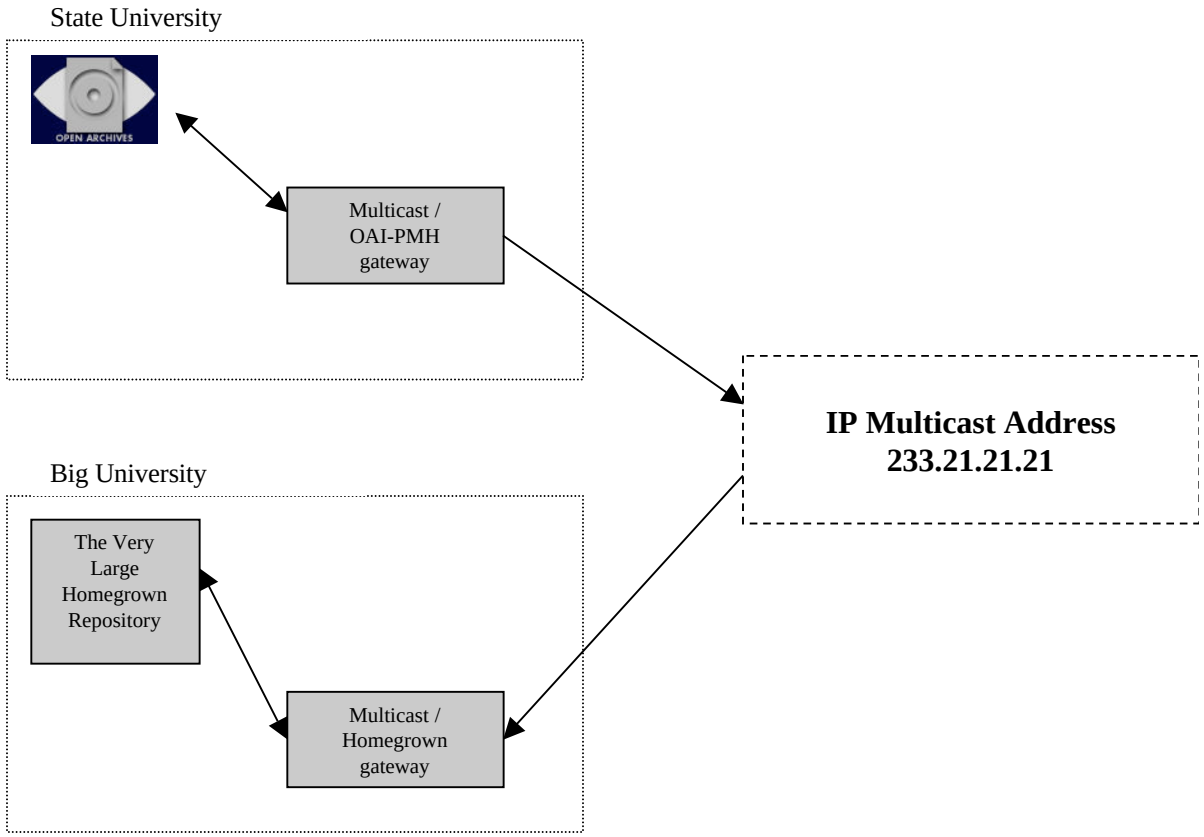
```
From cwild@notesmail.cs.odu.edu  Sun Sep  5 07:49:04 2004
Return-Path: <cwild@notesmail.cs.odu.edu>
Received: from notesmail.cs.odu.edu (notesmail.cs.odu.edu [128.82.4.18])
by cartero.cs.odu.edu (8.12.10/8.12.10) with ESMTP id i85BmlmV024367
for <fac@cs.odu.edu>; Sun, 5 Sep 2004 07:48:48 -0400 (EDT)
Subject: diagnostic exam
To: fac@cs.odu.edu
X-Mailer: Lotus Notes Release 5.0  March 30, 1999
Message-ID: <OFEC66F3F2.759C45D7-ON85256F06.00419627@cs.odu.edu>
From: cwild@notesmail.cs.odu.edu
Date: Sun, 5 Sep 2004 07:58:11 -0400
X-MIMETrack: Serialize by Router on lotus/ODUCS(Release 5.0.12  |February 13, 2003) at
09/05/2004 08:00:41 AM
MIME-Version: 1.0
Content-type: text/plain; charset=us-ascii
X-SBPass: GlobalNoBounce
X-SBClass: OK
Status: R
X-Status:
X-Keywords:
X-Remora: http://repo.state.edu/oai?verb=GetRecord&identifier=oai:repo.state.edu:202134
&metadataPrefix=oai_mets

Fall 2004 Diagnostic exam will be Saturday October 2nd 9 am to 5PM.

[rest of message deleted]
```

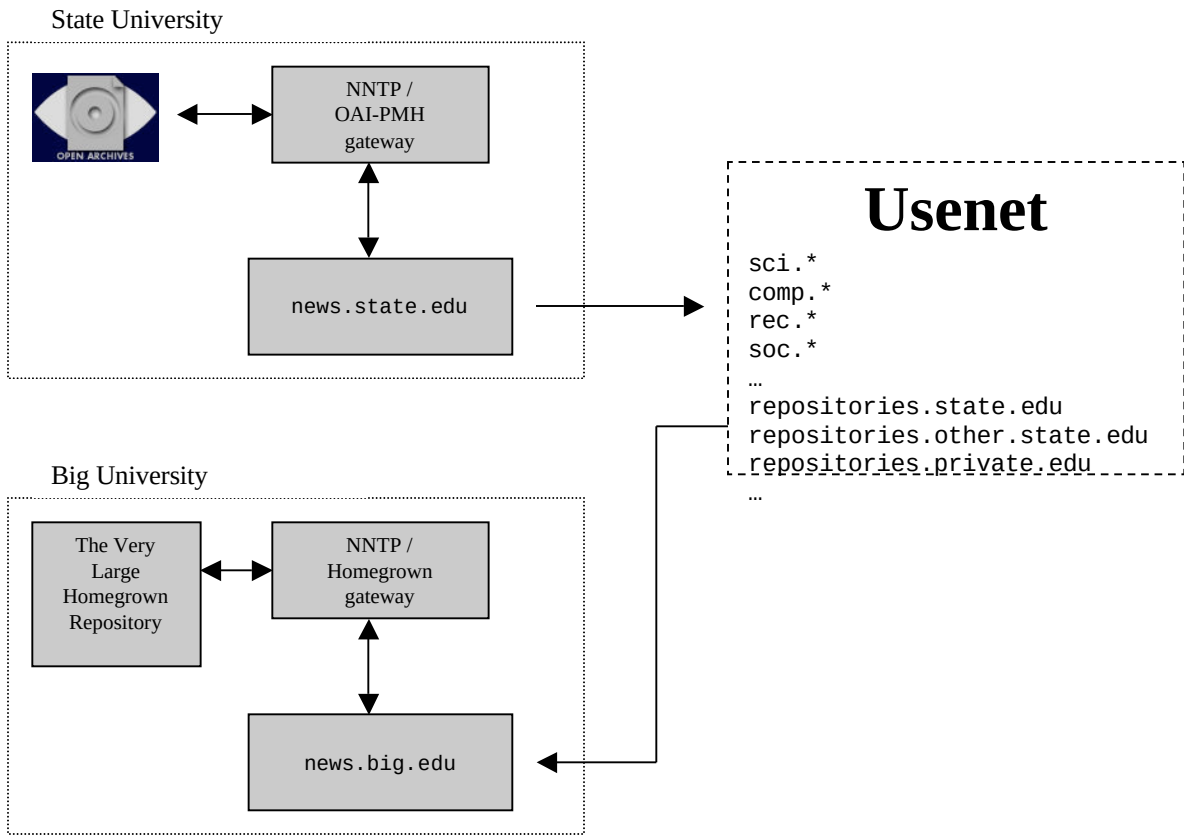
SI: IP Multicasting

- Harvest repository contents (or baseURL or Identify repsonse) and multicast to a well known addr / port
- Listen to the addr / port to discover new repositories (or contents)



SI: NNTP

- Associate a newsgroup with each repository baseURL
- Harvest contents from the repository and post as news messages
- Use NNTP to advertise new repositories
- Let Google Groups (or other Usenet services) archive the contents



Research Questions

- Repository discovery vs. exposing the repository contents
- OAI-PMH harvesting vs. other methods of distributing content
- Measuring / profiling:
  - replication
  - scalability
  - security
  - provenance

Project Management

Milestone	Date	Deliverable
1. simulate systems, analysis, requirements	months 0-3	exploratory report
2. build proof-of-concept systems	months 4-8	1-3 software packages
3. run tests in local testbed	month 9	software enhancements
4. run tests w/ remote partners	month 10	software enhancements
5. issue report, identify problem areas, future work	months 11-12	final report

Table 2. Milestones and Deliverables

Milestone	Evaluation Metric
1. simulate systems, analysis, requirements	Simulation data for upper bounds of 10k repositories, 100M records, and 1TB of data.
2. build proof-of-concept systems	Each system configurable in < 4 hours.
3. run tests in local testbed	Test with upper bounds of 10 repositories, 10k records and 1GB of data.
4. run tests w/ remote partners	Test with upper bounds of 3 repositories, 10k records and 1GB of data.

Table 3. Evaluation Metrics

2 Co-PIs, 2 graduate students, 1 undergraduate student