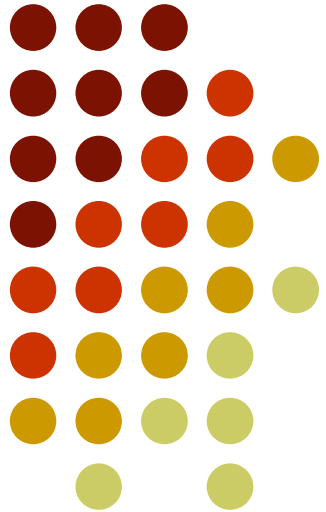
Continuing Persistence:

The Persistent Archives Test-bed (PAT) Project at SLAC in 2005 – 2006: A Progress Report



Basic records description

- SLAC Large Detector (SLD) for the SLAC Linear Collider (SLC) 1983-1988
- Early and prolific user of world-wide web
- No further need to keep data confidential
- Many types of electronic documents
- Meet US Department of Energy (DOE)/National Archives (NARA) criteria for retention

Basic records description

- News items and Hypertext News
- Publications and Technical Notes in a variety of formats
- Presentations in PowerPoint, PDF, and Postscript
- Web pages in HTML format
- Graphics in Postscript, Encapsulated Postscript, GIF and JPEG formats

Progress in 2005-2006

- Web Crawl Analysis
- Metadata Skeleton / Scheme development
- SLD Collection Arrangement
- Next Steps

Web Crawl Analysis

ITERATIVE PROCESS

- **Round 1**: Difficulties/issues encountered
  - Massive crawl: *Mother Lode* and *Monstrum Ingens*
  - ***Mother Lode***
    - Preserved endangered electronic records
    - Serves as a foundation and basis for subsequent work: can be mined as we iterate crawling
    - Absolutely necessary first step

Web Crawl Analysis

- ***Monstrum Ingens***
  - Too much information,
  - Too little useful organization
  - **Benefit**: Made us have to think about what we really want / need…
    - ***Had:*** Series descriptions based on archival appraisal of SLD records.
    - ***Needed:*** the same information, ***but***
      - Arranged hierarchically
      - Linked to NARA/DOE research records control schedule (our target)

Web Crawl Analysis

- Made us have to think about what we really want/need…(cont'd)
  - ***Had:*** all of the SLD electronic records (maybe?)
  - ***Needed:*** a way to know precisely what we had gathered in the crawl
  - Analyzed original crawl, sorted by urls
    - Were all links captured? Which ones weren't? why not?
    - Created a script to parse webpage for URLS and compare the URLs with the crawl result. If the URL isn't in the list, capture the URL along with the file

Web Crawl Analysis

- **Round 2**:
  - Ran a second, tightly targeted crawl
  - Used freeware tool: HTTrack
  - Crawled only one records series in the hierarchy (7: Committee Reports)
  - Uploaded crawl result to SRB at SDSC
  - Now ready to attempt metadata extraction/injection
- **Major epiphany**: the crawl is PART of the archival process, not outside of it

Metadata Development

- Parallel activity: constructing metadata scheme
  - Compatible with NARA LCDRG (Life-Cycle Data Requirements Guide)
  - Informed by current best practices :
    - Dublin Core
    - Arizona Model
    - PREMIS (PREservation Metadata Implementation Strategies – OCLC) issued 2005—not studied in depth
    - Hodge, et al.
    - Discussed metadata attributes with collaborators

Metadata Development

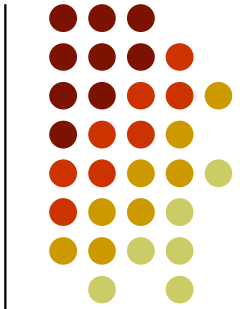
- Evolving as the crawl analysis progresses
- Two levels of metadata
  - Collection level
  - Item level
- Two main categories of metadata
  - **Injected** – externally applied, manually or automatically
  - **Extracted** – automatically pulled out of the content of the electronic records

http://www.slac.stanford.edu/history/projects/MetadataSchem7.html

Metadata Development

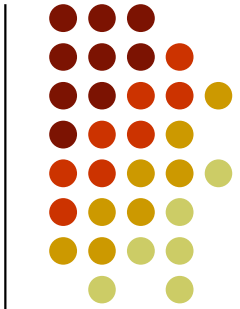
- Six sub-categories of metadata
  - 1. slac.gov – NARA/DOE required attributes
    - slac.gov.recordgroup
    - slac.gov.agency
    - slac.gov.referenceby
    - slac.gov.schedule
    - slac.gov.series
    - slac.gov.description
    - slac.gov.retention

Metadata Development



- Six sub-categories of metadata
  - **2. slac.creator** – all flavors of creators
    - **slac.creator.organization**
    - **slac.creator.division**
    - **slac.creator.group**
    - **slac.creator.person**
    - **slac.creator.owner**

Metadata Development



- Six sub-categories of metadata
  - **3. slac.description**
    - **slac.description.type**
    - **slac.description.by**
    - **slac.description.date**
    - **slac.description.remarks**
    - **slac.description.local**
    - **slac.description.webplatform**
    - **slac.description.format**
    - **slac.description.filesize**

Metadata Development

- Six sub-categories of metadata
  - **4.** slac.identifier – attributes that identify this copy of the electronic entity
    - **slac.identifier.storagelocation**
    - **slac.identifier.persistent**
    - **[others may be developed…]**

Metadata Development

- Six sub-categories of metadata
  - **5. slac.capture** – attributes that detail how the electronic entity was gathered for archiving
    - **slac.capture.tool**
    - **slac.capture.settings**
    - **slac.capture.sitemap**
    - **slac.capture.date**
    - **slac.capture.contact**
    - **slac.capture.remarks**

Metadata Development

- Six sub-categories of metadata
  - **6.** slac.date – date of archived entity, rather than of any processing/handling of entity
    - **slac.date.begun**
    - **slac.date.modified**

SLD Collection Arrangement

- Arranged descriptions hierarchically:
  - 30 descriptions → 5 series:
    - 1B1a Administrative records
    - 1B8 Computer code documentation
    - 1B9a Technical documents
    - 1B10 Supporting technical information
    - 1B13a Evaluated or summarized data
  - Based on relevant DOE Records Control Schedule (RCS) items

SLD Collection Arrangement

- Linked to NARA DOE Records Control Schedule for Research and Development Records

http://www.slac.stanford.edu/history/projects/SLDERecsV5.htm

- Analyzed how electronic records series relate to SLD paper records:
  - Duplicates?
  - Supplements?
  - Entirely new/different content?

Next steps … Crawling

- Automate further analysis?
  - Comparing what we have crawled with the records descriptions, to see how completely the crawl captured the desired sites.
  - Part automatic and part manual
- Why are we taking from a web crawl rather than from the machine?
  - **Benefit:** will pick up the linked information.
  - **Drawback:** has limitations/boundaries (dynamic pages)

Next Steps… Collection Arrangement

- Trial run on Committee Reports Series
  - Upload to SRB (done)
  - Try out PAWN tool (Producer Archive Workflow Network – UMd) (beginning in April 2006)
  - Transfer electronically to NARA ERA
  - Evaluate results
- Replicate process with a second SLD records series… and a third series…

Next Steps… Metadata

- Trial run on Committee Reports Series
  - Upload to SRB (done)
  - Automate injection of metadata (with GaTech tools – beginning in April 2006)
  - Automate extraction of metadata (“ “ “ “ “)
  - Evaluate results
- Develop crawl parameters metadata that could possibly be generalized across several crawl tools?
- Look in-depth at PREMIS

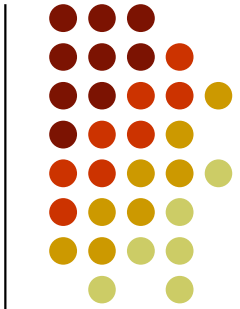
Next Steps… Beyond PAT

- Establish Electronic Records archiving program at SLAC
  - Institutional commitment
  - Financial support
- Who is an Archival IT professional?
  - What type of background?
  - What kind of position description
  - What sort of pay scale/compensation?
  - How and where recruited?

Next Steps… Beyond PAT

- Archival Primer for IT professionals (?)
  - NARA ERM Guidance on the Web (http://www.archives.gov/records-mgmt/initiatives/erm-guidance.html), Fast-Track Guidance Products
    - Preliminary Planning for Electronic Recordkeeping: Checklist for IT Staff
    - Preliminary Planning for Electronic Recordkeeping: Checklist for RM Staff

Conclusion



- All SLAC work products for the PAT project are online, at

  http://www.slac.stanford.edu/history/projects.shtml
- Home page for entire PAT project is

  http://www.sdsc.edu/PAT/
- My email address:

  jmdeken@slac.stanford.edu