SLAC-PUB-10857
December 2004

Archiving SLD Records in SRB:
The Persistent Archives Test-bed (PAT) Project at SLAC in 2004

J. M. Deken, A. Hasan
Stanford Linear Accelerator Center, Stanford University, Stanford, California 94025

Abstract

Report on the first year of SLAC's participation in the collaboration to test the NARA prototype persistent archives' ability to perform the functions of accessioning, arrangement, description, preservation and access on the electronic records of the SLD (SLAC Large Detector) collaboration.

Work supported in part by National Archives and Records Administration – National Partnership for Advanced Computational Infrastructure Collaboration on Persistent Archives, and by Department Of Energy Contract number DE-AC02-76SF00515

Project Background

The National Partnership for Advanced Computational Infrastructure (NPACI), a National Science Foundation-sponsored program, has been collaborating for some time with the United States' National Archives and Records Administration (NARA) on the development of prototype persistent archives. The NPACI partners formally involved in the collaboration include research staff of the Data and Knowledge Systems group at the San Diego Supercomputer Center (SDSC), the University of California at Berkeley, Stanford, and the University of Maryland, among others.[1]

In 2004, the Stanford Linear Accelerator Center (SLAC) joined this collaboration to participate in the Persistent Archives Test-bed (PAT) Project. The SLAC Archives and History Office is serving as a test site for the PAT Project for the automation of archival processes. SLAC's task has been to provide a set of related web-based permanent records series already appraised for permanent retention according to the Department of Energy's Research and Development Records Schedule (N1-434-96-9)[2]. SLAC's PAT project goal is to test the NARA prototype persistent archives' ability to perform the functions of accessioning, arrangement, description, preservation and access.

Records selected for PAT

The candidate records at SLAC have been generated by the SLD (SLAC Large Detector) for the SLC (SLAC Linear Collider) Collaboration, a group of about 150 physicists from many Universities and Laboratories who proposed, built, ran, and analyzed data from the SLD detector at SLAC from 1983 through 1998.[3] The records of this group consist of:

- 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

These records have been selected for the PAT because they meet the project requirements of:

- Permanent retention. The SLD meets Department of Energy criteria for a Level I project, and its substantive records are therefore scheduled for permanent retention.
- Project completion. The SLD is a completed project (as of June, 1998), which means that the collaboration data and results are no longer embargoed, are ready for

[1] Moore, Reagan, et al. NARA Persistent Archives NPACI Collaboration Project Proposed Research Agenda. July 9, 2003, SDSC Technical Report 2003-02r.
[2] United States Department of Energy Research and Development Records Retention Schedule. (http://www-it.hr.doe.gov/records/doe_rd2.htm, 10-20/98)
[3] See SPIRES Experiments Database entry for the SLD collaboration, at: http://www.slac.stanford.edu/spires/find/experiments/www2?ee=SLAC-SLC-SLD

transfer to NARA, and can be freely accessed by the public.[4] Since the PAT is a prototype that will be widely disseminated and examined, it is important that the data and records used to populate it are free from access restrictions.

Plan of Work

The first year of PAT project work at SLAC has been devoted to completing a web crawl through the SLD web site, assessing the crawled materials, and developing and populating metadata elements for the captured web materials.

Web crawl

Web crawl technology developed by UCSD-SDSC has been used to crawl the SLAC web site, starting with the Uniform Resource Locators (URLs) for the main pages for the SLAC Large Detector (SLD) collaboration (http://www-sldnt.slac.stanford.edu and http://www-sld.slac.stanford.edu).

Two major problems have been encountered with the web crawl:

1. The crawler broadened its search out too widely in a horizontal fashion, harvesting web pages linked to those linked to the SLD pages, many of which were not of interest to the present project. Solving this problem has required a manual review of the over 1100 pages harvested to identify the 112 URL roots of interest to the project at hand.
2. One issue that has arisen during the analysis is that the crawler appears to have been stopped from harvesting certain branches of the SLD web site because it encountered "nocrawl" messages, placed in various spots on the site to block outside web search engines from exploring those portions. When the crawler encountered a robot.txt file, it halted harvesting below that file. Some of the pages beneath such robot.txt files are of interest to this project, but were not automatically gathered by the robot. The solution to this particular problem is still under investigation. It appears that a method needs to be developed to allow creating agency/entity web crawlers to access all portions of their own sites, even those which are not open to the crawlers instigated by outside search engines.

Assessing the crawled materials

Root URLs of interest to the project were identified by the application of standard archival processing techniques to the content of the SLD collaboration web site, viewed through a web browser on a desktop personal computer.

[4] SLAC is obliged to retire permanent research records 2 years after the collaboration, experiment or project that created them has completed work. The retired permanent records are required to be transferred to NARA 28 years after their retirement.

Metadata Development

See table at end of paper for metadata elements developed to date, along with their working definitions.

Issues and Items to be Addressed

SLAC is now about to store the SLD web pages in a SLAC-installed MCAT-enabled SRB server. The reason for this is to allow us to play around with a few things (such as stored location in the SRB namespace, who can read/write to them what extra metadata to store etc).

The issues that we will face with this coming task (based on previous experience) are:

1) How to present data to users?
   Is it sufficient to present the user with set of pages containing the names of the different SLD web pages and the user clicks on those?
   This will mean that the links within a particular SLD web page will be dead as they will point to the original sites which are no longer present.
   Or, do we have to manipulate the SLD web pages to make the link point to an SRB location allowing the user the look and feel of the whole set of pages.
2) How to relate SRB data to archiving metadata? It seems that at the moment it's best to keep the archiving metadata in a separate set of database tables outside of SRB. This allows the flexibility to change the attributes of the tables quickly without having to get a new version of SRB. This issue is relevant whilst the metadata is in the process of being defined.
3) How to make data invisible to users? We need some simple way allow the archivist to take a record offline for some reason (maybe to update metadata or to replace it or something). We need an attribute (perhaps it is part of the archiving metadata) that provides a code, the application displaying the data will not display pages that have a certain value, and instead they display the meaning of the code (maybe that can be configurable).

Once we have an understanding of how to do things at SLAC we can copy the data to a production server (we clearly need <50GB of space so it should not be difficult to find). But, that brings up another interesting question: how long does data stay on disk? The only way we will know that a user has finished with a record is that the record is untouched for some period of time (where some period of time may be a day or longer) then the record could be eligible for purging. Or, should there be some explicit log-on-log-off system so we know who is accessing which records and? One guess is that the 'Sinit' will be recorded in the SRB log file so what will be needed is an application that is able to query that and find out who's 'logged on'. This raises a further question of how do we know if/when someone has ‘logged off’?

SLAC DESCRIPTIVE METADATA ELEMENT NAME	DEFINITION
Record Group Number	NARA-assigned number designating cognizant/originating organization. At SLAC: usually 434 (DOE)
Agency	Entity responsible for making the resource available At SLAC: US Department of Energy
Reference provided by	Organization responsible for providing reference service on the subject record/resource. At SLAC: Before transfer to NARA, inquire with SLAC Archives and History Office, 2575 Sand Hill Road, MS82, Menlo Park CA 94025 (phone) 650-926-3091 (fax) 650-926-5371 (email) slacarc@slac.stanford.edu.
Organization	Entity responsible for making the resource available. At SLAC: Stanford Linear Accelerator Center
Division	At SLAC: Division of SLAC in which the creator works
Group	At SLAC: Group within SLAC division in which the creator worked when record/resource was created
Creator	Entity primarily responsible for making the content of the resource At SLAC: Person or collaboration that created the resource
Owner	Entity responsible for making contributions to the content of the resource At SLAC: Person primarily responsible for updating and maintaining the content of the resource during the maintenance and use phases of its life-cycle/continuum.
Root URL	A reference to a related resource At SLAC: root URL of the tree on which the subject record/ resource originally resided.
Series Designator	Unique identifier within a record group for a series At SLAC: will vary
DOE Schedule Item Number	Alpha-numeric reference to disposition schedule item relevant to the record/resource in question At SLAC: Usually item number from N1-434-96-9, Records Disposition Authority for Department of Energy Research and Development Records.
DOE Schedule Item Description	Narrative description of disposition schedule item relevant to the record/resource in question
DOE Retention	Retention period designated by DOE Schedule At SLAC: Usually this will be “Permanent”
Access Restriction Status	Limitations on use of record/resource designated in disposition schedule or by record/resource creator
Use Restriction Status	Indicates whether restriction status has been verified
Saved As	At SLAC: short file name assigned during web crawl that harvested the record/resource
Entry number	An unambiguous reference to the record/resource within a given context. At SLAC: unique identifier assigned by SLAC Archives
File location	At SLAC: original URL of the record/resource
Name	Name given to the record/resource At SLAC: title of the record series/electronic resource
Beginning date	A date in the event of the lifecycle of the resource At SLAC: date record/resource was originated
Last modification	At SLAC: date record/resource was last modified

SLAC DESCRIPTIVE METADATA ELEMENT NAME	DEFINITION
Description	An account of the content of the record/resource At SLAC: free text account of the record/resource
Description type	Indicates the level of description provided At SLAC: Series
Description author	Name of individual responsible for writing the Series Description
Description date	Date Series Description was completed
Copy status	Indicates what types of copies of the record/resource have been made. Choices are: Preservation; Reproduction; and Reference
Type of electronic entity	Nature or genre of the content of the resource At SLAC: Will use DCMI vocabulary terms (Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text)
Format	Physical or digital manifestation of the record/resource At SLAC: Will use MIME Internet Media Type designators
Filesize	Narrative description of record/resource size
Holdings Measurement Type	Quantitative metric used for the subject record/resource
Holdings Measurement Count	Number of units of the measurement metric the subject record/resource represents
Storage location	Name of the storage location of the record/resource
Storage media type	Storage media used for the record/resource
Remarks	Administrative notes for use by staff