Minutes
CHM Software Preservation Group
March 29, 2007

by Paul McJones

Attending

Lee Courtney	Al Kossow	Bernard Peuto
Bob Fraley	Paul McJones	Mike Powell
Henry Gladney	Randy Neff	Bob Sanguedolce
Kathe Gust	Paula Newman	Len Shustek
Philip Gust	Eric Petrich	Dick Toepfer

Status

Bob Goldberg, who worked with Robert Dewar on Spitbol in the 1970s and who already did a public-domain release Spitbol 360, is interested in working on categorizing and curating the Snobol-related content on the Arizona DVD. He knew Ralph Griswold, and other principals from the Snobol world. He notes there are many executable files as well as source file, and they are potentially runnable on emulators.

Al Kossow hopes to make available under some kind of hobbyist license the software for the 16-bit HP 21xx series. He's making progress getting access to the Xerox SIGMA software. MIT now has a copy of Multics sources from Bull; the actual announcement is still pending.

Paula Newman: Paula volunteers as a friend of a public library, and notices a number of computer books being thrown away; is there some way to cross-check whether CHM would like some for the permanent collection?

Bernard Peuto:

- We're coming down the home stretch on a standardized license for Microsoft to donate historic software. We're working with IBM on a similar license for specific software: the System R relational database management research system. We're also working with Xerox PARC re the Alto source code.
- Another note: there will be an NLS anniversary in 2008; we're considering what kind of observance would be possible.
- We will keep the monthly schedule as publicized but if the the agenda does not justify a meeting it will be cancelled generally a week before the planned date and status and available presentations will be postponed to the next meeting.

Henry Gladney: Preserving Digital Information

Henry recently published a book on digital preservation, and gave us an overview of the book, and of its core methodology that he calls "Trustworthy Digital Objects (TDOs)" from these slides: http://home.pacbell.net/hgladney/PDIslides.pdf .

Readers might be surprised by differences from other work in this area:

- TDO methodology focuses on information representation, in contrast to the management of digital repositories;
- Much discussed preservation difficulties are shown not to be difficulties after all by application of early 20th-century epistemology

Questions

Bernard: What is the competition?
Henry: NDIIPP at the Library of Congress (http://www.digitalpreservation.gov), ERA at NARA (http://www.digitalpreservation.gov), PLANETS in Europe (http://www.planets-project.eu), work at National Archives of Australia (http://www.naa.gov.au/recordkeeping/preservation/digital), and several national projects.

Len: The Library of Congress is spending tens of millions of dollars.What are they spending it on?
Henry: MDIPP -- they are spending the money on creating collections, but Henry believes their technical program has problems -- see Digital Preservation in a National Context: Questions and Views of an NDIIPP Outsider, *D-Lib Magazine* 13(1/2), January 2007.

Dick: What about Google's book scanning project?
Henry: Am focussing on the archiving process, not the digitization process.

Bob Fraley: Repository Proposal

Observations

1. After being away from the group for a long time, he was surprised that there is still lots of discussion about what to do, tools, etc., rather than on actual collection/preservation work.
2. We should pick some basic tools now, things that can be built on, and layer other tools on them later.
3. If we bring in a tool now, how do we know it will still be usable in ten years? It seems that we should be using very simple, easy to maintain tools.

Proposal

Select a software configuration management system (CMS), such as Subversion. Check the artifacts into the CMS. Various museum/cataloging systems could be used over time, and could point into the artifacts stored within the CMS. If you need to make derived or updated versions, they could be checked back into the CMS, or stored in a separate system, as appropriate.

The basic idea is a separation of concerns: use the CMS for permanent, reliable, versioned storage of artifacts.

Questions

Paula N: Would this deal with access control?
Answer: Yes.

Len: Does a CMS really deal with long-term archival storage? A CMS lets you make a new version.

Eric: A version-control system tracks changes to documents. Since the software repository would need to contain stuff that never changes (plus metadata that could easily change), I'd want to feel confident that the version-control software would allow us to keep stuff

forever. (Administrators usually have the ability to remove files completely, while other users can "retire" files that are no longer part of the "top of tree.")

Bob F.: We definitely wouldn't use all the features.

Paula N.: It would be important to enforce ability to change some things (e.g., correct metadata), but not others ("original" artifacts).

Al: He's working on a design for digital artifact storage.

Originally used a directory tree representing vendor/model/artifact (see bitsavers.org) stored on a RAID5. As this tree evolved in structure based on "editorial decisions", it made it very difficult to keep track of which portions had a current backup. So the new scheme is "log" based: there is logically an append-only structure, where each addition to this log is a record with some descriptive XML and the data itself. Various index structures can be built that point into this. If a record needs to be "updated", a new record is written; it can link back to the previous version. If the index is reorganized, that doesn't affect an existing pointer to an item. Also, the entire old index could be maintained for the historical record. Al is thinking about using DSpace with LOCKSS (Lots Of Copies Keep Stuff Safe) on top.

Bernard: The DAM Book describes a digital access management scheme (for digital photography) that uses a very similar approach to Al's scheme. It describes one additional idea: the "side-car", which allows adding metadata to an item without modifying the original. There is fairly inexpensive software for this: iVue Media Pro [recently licensed by Microsoft], and several public-domain equivalents.

Len: In principle, would you have to scan all log entries to find the "current" version of an item?
Various: Yes, but in fact you keep an index; if it is damaged, can rebuild with a full scan.

Len: What is the size of the records in your log?
Al: It could be a DVD, a tape image, ...

Henry: There are about 120 content management/digital library systems currently. (Of the systems from universities, Henry likes Fedora and Greenstone. It would be a bad idea to build yet another.

Al: What's wrong with starting now with my immutable time-sequenced log stored in a regular file system?

Bob F.: A CMS would track relationships while people created various derived versions of the original files.

Henry Gladney: SNOBOL project

Two goals: creating durable annotated collection of materials related to the SNOBOL language; and to set up and test a practical environment (using Greenstone) that can be used by the participants of the SNOBOL project. Currently Henry is climbing the steep curve of Linux system management.

See attached "Snobol Collection.pdf" for a copy of the slides Henry presented, or follow this link: http://mail.computerhistory.org/pipermail/scc_active/attachments/20070323/1b02fd60/SnobolCollection-0001.ppt

Upcoming Meetings (2007)

Day	Date	Time	Conf Room
Wednesday	April 25	5:30 pm – 7:30 pm	Hopper
Wednesday	May 16	5:30 pm – 7:30 pm	Hopper.
Wednesday	**June 27**	5:30 pm – 7:30 pm	Hopper **CHANGED**
Wednesday	July 18	5:30 pm – 7:30 pm	Hopper
Wednesday	August 15	5:30 pm – 7:30 pm	Hopper
Wednesday	September 26	5:30 pm – 7:30 pm	Hopper
Wednesday	October 17	5:30 pm – 7:30 pm,	Hopper
Wednesday	November 14	5:30 pm – 7:30 pm	Hopper.

(no December Meeting)