Digital Libraries, Stanford, and Google

A brief early history

Gio Wiederhold, Stanford University, 12 October 2018

Using electronic communication and computing to gain access to information kept in libraries was already the goal of Memex, a device envisaged by Vannevar Bush in 1945 in an article "As We May think". Memex sketches made it look like an augmented desk, looking a bit like the 256-word LGP-30 computer I got to use in 1958. Memex was to be able to retrieve documents from afar and store them for the researcher. No automated help in selecting documents was envisaged.

In 1965, Ted Nelson envisaged linking of documents in his Xanadu project, and coined the name Hypertext.

A variety of primitive, often proprietary and isolated network arrived in the early 1970ties, enabling access to remote documents The Defense department's ARPANET had the FTP protocol for transferring files from or to remote locations. Prof. Joshua Lederberg, Nobel laureate in Genetics at Stanford, having had access to the ARPAnet, was impressed by the capabilities he foresaw in using computer-mediated communication, and published his views in the Nov. 1978 issue of the Proc. of the IEEE. Digital Communications and the Conduct of Science: the New Literacy. It till took time for technology to match that vision.

Such access would benefit research in general and led to a need for broader support than the military sponsors of the ARPAnet were able to provide. In 1986 the National Science Foundation (NSF) established the NSFnet for its researchers. Starting in 1983 Arpanet technology had been scaled thorough the introduction of Domain Name Servers and the IPv4 protocol by Paul Mockapetris at USC's ISI research laboratory and John Postel at SRI, formerly the Stanford Research Institute. Around 1987 the strict military nodes were extracted out of the ARPAnet into a protected Milnet subnetwork, allowing by 1989 to switch NSF-sponsored services to be supported by a public ARPAnet. With a new governance the result become the Internet, which now also permitted commercial use. While major services incur charges, these were absorbed by participating local suppliers and institutions. The public started expecting that Internet services should be free.

> A personal note: Free use encourages overuse, a process known as The Tragedy of the Commons. I'd rather pay a few pennies for the messages I write and send, and not receive thousands of spam messages from businesses that exploit free facilities. To receive free services from social media companies, participants are willing to provide their private data, and often private data of their on-line friends as well. Subsequently they complain about loss of privacy. [Gio Wiederhold]

In 1989 Tim Berners Lee at CERN proposed a system for sharing physics document which led to the Hypertext Transfer Protocol (HTTP). It was made generally available in 1996, initiating the World-Wide-Web. To represent documents with device independent formatting, tables, and images the the Hypertext Markup Protocol (HTML) was specified in 1991, expanding IBM's SGML document presentation format. These capabilities allowed rapid sharing of documents by other physicists at Illinois and the Stanford Linear Accelerator Center (SLAC). In 1993 Marc Andreessen and Eric Bina at Illinois provided the Mosaic browser to display HTML-formatted pages, which became Netscape.

In 1991 Maria Zemankova of NSF had sponsored a workshop on integrating prior research on text understanding, hypertext, and networking. Ed Fox and Michael Lesk wrote a report "Electronic Libraries" based on that meeting, motivating two more workshops in July 1992 at NSF and in December 1992 at Xerox PARC on Digital Libraries.

With the Internet resources broadened and several projects and companies provided help in finding relevant material. Those systems used prior user patterns, internal ontologies, and rules to determine relevant information by match terms entered by the searcher. The rapid increase of available material often overwhelmed them. Military researchers were equally frustrated. At ARPA's ISTO office, Gio Wiederhold (on leave from Stanford CSD and EE 1992-1995) initiated the Intelligent Integration of Information (I3) program. In addition to serving direct military needs it also envisaged integrating data from public sources.

An interagency Digital Library Initiative (DLI) was established in 1993 and then supported by 3 Federal agencies.

1. (D)ARPA –Brian Boesch (CSTO) and Gio Wiederhold
2. NSF – Larry Rosenberg, later Su-Shing Chen and then Stephen Griffin, who managed the DLI efforts from 1998 onwards
3. NASA – Paul Hunter, Milt Halem, and Eugene Miya (abt 16%)

A call for proposals was issued and half a dozen were selected for substantial DLI funding. Several showed good results, but the most visible outcome was the founding of Google.

Two Stanford Computer Science Department students, Sergey Brin and Larry Page invented the page rank algorithm. It prioritizes for display web pages using the pattern of links on the Internet itself. Their research was supported by Professors Rajiv Motwani, Jeffrey Ullman, and Terry Winograd. In 1996 Larry Page and Sergey Brin, then PhD students in Stanford CSD, working on the Digital Library project, needed a large amount of diskspace to test their Pagerank™ algorithm on actual world-wide-web data. At that time 4 GigaByte hard disks were the largest available, so they assembled 10 of these drives into this low-cost cabinet assembled with Leggos. In September 1998 Sergey and Larry left Stanford and founded Goggle.

In Nov 1999, Google Inc, by then operating one of the primary search engines on the web, provided replacement storage capacity to the Digital Library project so that we could move this original storage assembly into Stanford's Computer history displays. That original Google Storage is now on exhibit on the lower floor of the Jen-Hsu Huang center in the Stanford Engineering quad.

As of September 2000, Google, now located in Mountain View, operated 5000 PCs for information searching and web crawling, using the LINUX operating system.

Digital Libraries and Google are now part of most public services and professional societies. Digital content has replaced much of Stanford's paper libraries.

========== o ========== o ==========