



© 2002 The Charles Babbage Institute for the History of Information Technology
211 Andersen Library, 222 – 21st Avenue South, Minneapolis, MN 55455 USA

Digital Archives for the History of Software: The Allen Newell Collection and the Herbert A. Simon Collection

Corinna Schlombs
University of Bielefeld

Date published: 13 September 2002

Carnegie Mellon University Libraries maintain two full-text digital archives that render large portions of the Allen Newell Collection (<http://heinz1.library.cmu.edu/Newell>) and the Herbert A. Simon Collection (<http://diva.library.cmu.edu/Simon/>) directly available to online users. About eighty-five percent of the Simon collection and more than ninety percent of the Newell collection are selected for online publication, excluding those portions of the material from digitization that involve intellectual property rights and confidentiality issues, such as large correspondence files and student or tenure files. As of January 2002, the digital processing of the Newell collection neared completion with publication of about 145,000 scanned images; considerable extension of the Simon collection is expected for March 2002 with publication of an additional 100,000 images that will be added to the existing 56,000 images. Intended for diverse user communities, the archive sites contain archival databases equipped with a user interface for extensive full-text search and browse options. They also have links to biographical information and other background material on Newell and Simon. They are an outstanding source of information for users from high school students to professional scholars.

The archival databases of both online collections employ identical interactive platforms that allow users to display the archival documents directly online with Acrobat Reader. The displayed images still show physical qualities of the original documents such as style of handwriting or appearance of the paper. The user interface offers four ways of searching the collections and two ways of browsing them. Users can search the documents' full text as the originals have been scanned and processed with natural language recognition software. Alternatively,

the browsing functions still give users access to the physical and logical organization of the physical collections.

Recently, both archival databases have been moved to a new user platform with enhanced search options, the new integration of the Acrobat Reader interface, and more user-friendly design for more intuitive orientation of first-time users. Unfortunately, displaying a document still requires considerable time, from several seconds to several minutes, depending on a document's size. Amendments to the new navigation tools are desirable. Their current size and layout considerably decrease the monitor area that remains for the actual display of documents in Acrobat Reader, and when browsing the collections the only means for cursor navigation in exceedingly long document lists still is scrolling the side bar. Likewise, users can no longer directly switch from displaying a document that had been identified in a full-text search to the browse function, a former feature that allowed for the easy location of a document and viewing of its neighboring documents. Most importantly, the non-digitized archival holdings have been removed from the physical browsing lists, thus currently depriving researchers of information that formerly facilitated further investigations—for example, the identification of correspondence partners. However, the editors have made plans for reintegration of these materials in the database in near future.

Overall, the digital collections are intended to serve the needs of a diverse user community. The start pages not only give access to the archive databases and brief descriptions of the collections, they also include short biographical sketches and provide links to (auto)biographical writings and other selected papers. In this way, the collections address user groups such as high school or undergraduate students who seek more general information than the historical researcher. The historian, however, might wish additional information on the archival processing and the selection of material.

The online archives are part of a grant from the Institute of Museum and Library Services (IMLS) to the Carnegie Mellon University Libraries, the School of Computer Science at Carnegie Mellon, and the Carnegie Museum of Natural History. The project adheres to rigorous standards for preservation-quality digital imaging, and uses Machine Readable Catalog (MARC) records and the Encoded Archival Description (EAD) for content descriptions of books, monographs, journals and archival materials. Back up systems for emergencies such as data corruption or disk failure are provided, and the system is secured against hackers and intruders. The software was developed with an awareness that data will have to be migrated to new systems and technologies over time.

The digital archives provide highly convenient access to primary sources for researchers who are interested in the history of computing and artificial intelligence. They may often render costly and time-consuming archive visits

unnecessary, though for the time being scientific users will in some cases need to contact the archives directly in order to systematically include non-digitized material in their investigations. While preserving many qualities of physical records, the digitized collections open innovative ways of archival research by providing the full-text search function. They are an invaluable new research tool. In general, given the fragility of data due to insecurity and potential migration problems, one would hope that digital archives enhance but do not replace physical collections. Paper still proves to be our most reliable and permanent preservation material, and only future experiences will disclose which qualities of the originals inadvertently are missing from digital form.

Corinna Schlombs, “Digital Archives for the History of Software: The Allen Newell Collection and the Herbert A. Simon Collection,” *Iterations: An Interdisciplinary Journal of Software History* 1 (September 13, 2002): 1-3.