1-1-1998

The Heinz Electronic Library Interactive On-Line System (HELIOS): An Update

Gabrielle Michalek
Carnegie Mellon University, gm1l@andrew.cmu.edu

Published In
Public-Access Computer Systems Review, 6-19.

Galloway, Edward A. and Gabrielle V. Michalek. "The Heinz Electronic Library Interactive On-line System (HELIOS): An Update." *The Public-Access Computer Systems Review* 9, no. 1 (1998).

1.0 Introduction

In February 1994 Carnegie Mellon University (CMU) Libraries embarked on an ambitious project to convert approximately one million pages of the congressional papers of Pennsylvania Senator John Heinz into digital format. [1] Innovative image-management and text-retrieval software created at CMU provides the ability to search and retrieve these papers. Named in memory of the late Senator, the Heinz Electronic Library Interactive On-line System (HELIOS) currently allows researchers to browse, search, view, and print over 434,000 digital images from the collection. Accessible through the campus network and the Internet, HELIOS dramatically increases depth of indexing and quality of retrieval beyond that which archiving resources have traditionally allowed. Since January 1998, HELIOS can be accessed on the Internet at <URL:http://heinz1.library.cmu.edu/HELIOS>.

In order to stimulate the exploration and use of the Heinz papers by a much broader community of users than is expected with a paper-based archive, the University Libraries proposed to digitize the congressional papers. Over one million dollars was donated by the Heinz Family Foundation, Heinz Company Foundation, and the Heinz Endowments to support the establishment of the H. John Heinz III Archives and the digitization project. In addition to the Heinz gift, CMU and the CLARITECH Corporation contributed over $700,000 in resources, including permanent full-time staff salaries, archival equipment, and office rental space for a processing facility. Heinz assistance has made it possible to advance the techniques of digital preservation and access for archival collections.

Our goal has been to develop a digital archive to serve as a model for the archival profession and research community. Traditionally, archives have faced several challenges, including: 1. the creation of good finding aids and indexes for large archival collections; 2. the provision of effective retrieval of information from paper archives in spite of the inherent diversity and size of these unique materials; and 3. the difficulty of offering broad public access to archives since they represent resources that the researcher must visit in order to use effectively.

2.0 H. John Heinz III Congressional Collection

Shortly after the tragic death of Senator Heinz in 1991, the family placed his congressional papers at Carnegie Mellon University to serve as the research centerpiece for the H. John Heinz III School of Public Policy and Management. After a rented archival facility near campus was prepared, the collection was transferred from its storage facility in Harmarville, Pennsylvania to the rental facility.

The collection was completely processed by October 1996. It consisted of the Heinz House of Representatives Papers (54.9 linear feet) and the Heinz Senate Papers (670.6 linear feet). The Heinz Archives staff followed conventional processing methods to arrange and describe the papers while applying fundamental preservation techniques to the original material to ensure its longevity. In addition to providing access to the physical documents, the Heinz Archives provide electronic access to the most important series and subseries in the record groups.

In addition to documenting Heinz's tenure as a three-term member of the U.S. House of Representatives (1971-1976), the papers focus on his fifteen-year Senate career (1977-1991). Senator Heinz earned a national reputation based on his work on retirement and aging issues, international trade and finance, and environmental issues. The Heinz papers present a rich and valuable source of information about the professional life of John Heinz in the U.S. Congress and the social and political concerns of the nation during the senator's tenure. The Heinz Archives will aid scholars in understanding the senator's contributions to national policy and allow current public-policy makers to build upon his accomplishments and unfinished work.

3.0 HELIOS Team Members

The HELIOS project is composed of three umbrella units representing several disciplines. Each unit brings its own expertise to the project, resulting in major contributions to the design, creation, and implementation of HELIOS.

3.1 Laboratory For Computational Linguistics

Regarded as one of the world's premier laboratories for advanced work in the field, CMU's Laboratory for Computational Linguistics (LCL) focuses its research efforts on information management and retrieval issues. LCL researchers have developed efficient methods to analyze and extract language using computers, and this natural language processing (NLP) research is the basis of the CLARIT software.

3.2 CLARITECH Corporation

The CLARITECH Corporation, a CMU spin-off company recently purchased by JustSystem of Japan, has improved and marketed LCL's NLP technology, dubbing it CLARIT. CLARITECH's primary contribution to the HELIOS project is system

architecture and interface design. It is responsible for incorporating elements of the CLARIT system designed in the LCL into HELIOS and for creating four graphical user interfaces for the system.

3.3 Carnegie Mellon University Libraries

Three different units within the Carnegie Mellon University Libraries play a major role in the interdisciplinary functioning of the project. The Library administration is responsible for providing the leadership function for HELIOS as well as the fiscal management of the project. The Department of Library Information Technology is responsible for supporting and maintaining the HELIOS client/server system. The University Archives facilitates the interdisciplinary teamwork of the project. The Heinz Archives, a unit within the University Archives, is responsible for establishing control over the collection, appraising and processing the original Heinz papers, creating a finding aid to the collection, providing reference service, disseminating and cataloging the collection via OCLC, and preserving the original collection in perpetuity.

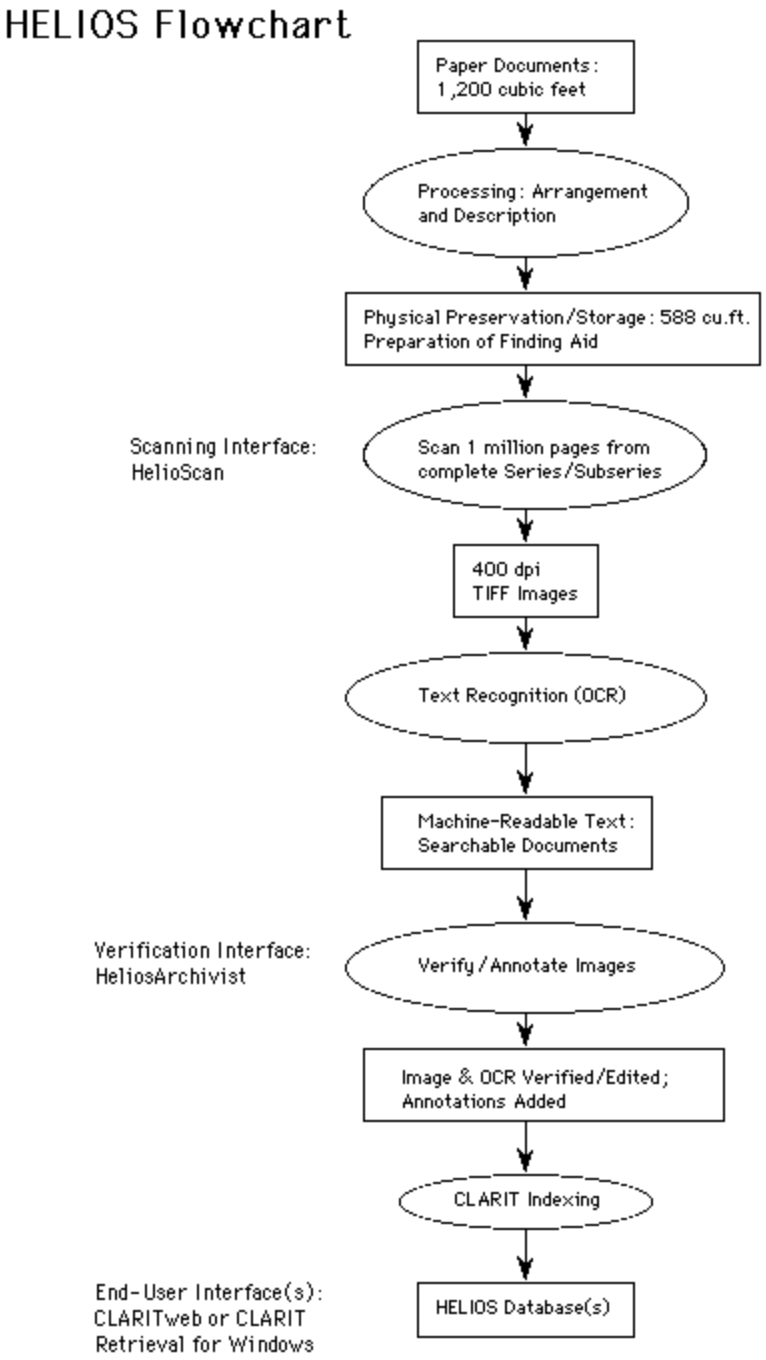
Working together, the University and Heinz Archives are responsible for developing four interface specifications, testing the interfaces before release, scanning the original material into electronic format, verifying the quality of the images, performing additional organizational tasks, creating annotations and links to other parts of the electronic collection, conducting user protocol testing, and training other library staff to use the system.

CMU archivists were principally involved in creating system specifications. Working diligently and closely with software designers was an important role necessary to guarantee that archival specifications were understood and clearly communicated. Moreover, we incorporated preservation issues into our plan, design, system architecture, and resource allocation to ensure the portability and interoperability of the source files.

4.0 HELIOS Processing

To create the growing HELIOS database, processed documents are scanned, converted to ASCII form via OCR (Optical Character Recognition) software, verified and annotated, and then indexed using the CLARIT natural language processing software. The project has developed four graphical user interfaces: a scanning interface (HelioScan), a verification interface (HeliosArchivist), a Windows-based user interface (CLARIT Retrieval for Windows), and a Web-based user interface (CLARITweb). (See Figure 1.)

Figure 1. Helios Project Flowchart



The original paper documents were processed based on the typical arrangement scheme of a congressional collection. During the last year of document processing, we began the digital conversion process, taking entire folder contents from completed series and subseries and digitizing them at 400 dots per inch (dpi). We converted the image files to ASCII text on a nightly basis using the TextBridge OCR software package. After the images and text were verified, with additional notes and organization added, the text was indexed by CLARIT, resulting in the searchable Heinz Senate Papers database.

4.1 Scanning

After one year of initial testing and development, we commenced the full-scale scanning of material in January 1996. By this time almost the entire Senate collection was processed, so the Heinz Archivist selected several series indicated by survey results as being of interest to local researchers. Since we are attempting to digitize the record of Senator Heinz's congressional life and activities, we are scanning the complete contents of the folders within a series and subseries.

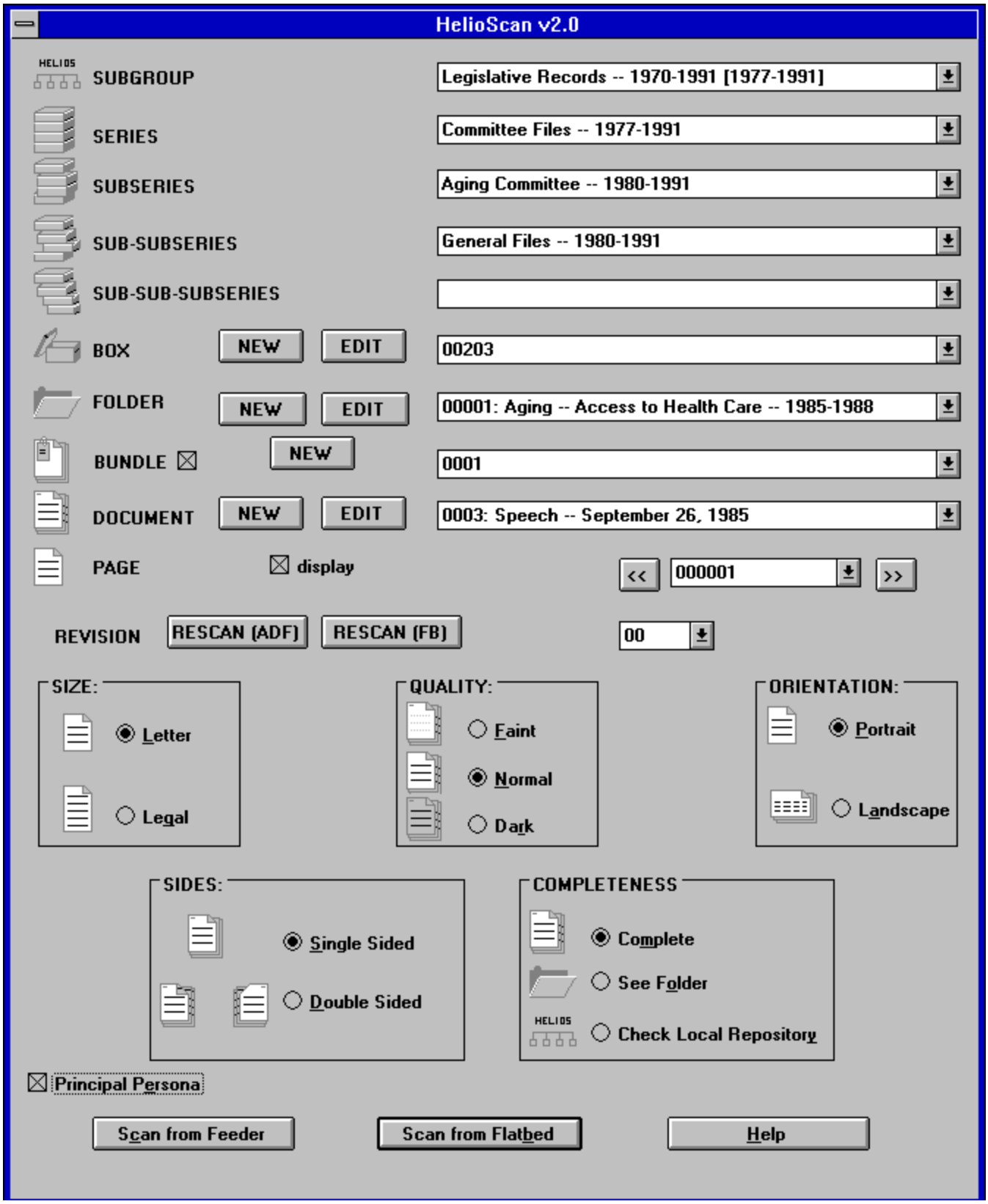
With the use of two 586 Dell PCs running Windows for Workgroups 3.1.1, 20" color monitors (1600 x 1200 pixels), and two high-end Fujitsu scanners, we are creating 400 dpi bitonal TIFF (Tagged Image File Format) images. The images are compressed using CCITT Group IV, an international compression standard, and backed up on 4 mm digital data storage tapes on a nightly basis. The scanning workstations are networked to our Sun Sparcstation server (UNIX) where the data is stored.

Our choice to use 400 dpi resolution coupled with bitonal scanning was driven by three factors: 1. 400 dpi was the highest true resolution offered when we began scanning; 2. the material in the archives consists almost exclusively of typical documents with black ink on white paper; and 3. a higher resolution yields better OCR performance. As far as digital preservation goes, the faithfulness of our image to the original document is directly proportional to the scanning resolution. Unlike compression algorithms for gray scale or color scanning, compression algorithms for binary images do not lead to the loss of any data and can thus be used to reduce the storage size of our images without compromising their faithfulness to the original.

Because the scanning procedure represents the most crucial aspect of the project, we designed a scanning interface, dubbed HelioScan, to facilitate the rapid scanning of documents while capturing essential contextual information (metadata) for the user and archives staff. [2] On average, the scanning operators scan approximately 1,200 pages per day. As of May 1, 1998, we have scanned over 560,000 pages from the Heinz Senate Papers.

HelioScan is structured to imitate a standard archival collection arrangement scheme, organizing documents into subgroups, series, subseries, and smaller units. (See Figure 2.) The scanning operator selects the appropriate level using drop-down menus. The operator then enters the box and folder number as well as the folder title and date. HelioScan was also designed to capture "bundles"; that is, groups of documents within folders originally fastened together by paper clips, staples, or rubber bands. These bundles, which often reflect inherent meaning, are more difficult to depict to an online user; however, doing so is important because it gives the user the same context as if he or she were physically examining the material.

Figure 2. HelioScan v2.0


HelioScan v2.0
HELIOS
SUBGROUP
Legislative Records -- 1970-1991 [1977-1991]
SERIES
Committee Files -- 1977-1991
SUBSERIES
Aging Committee -- 1980-1991
SUB-SUBSERIES
General Files -- 1980-1991
SUB-SUB-SUBSERIES
BOX
NEW
EDIT
00203
FOLDER
NEW
EDIT
00001: Aging -- Access to Health Care -- 1985-1988
BUNDLE
NEW
0001
DOCUMENT
NEW
EDIT
0003: Speech -- September 26, 1985
PAGE
display
<<
000001
>>
REVISION
RESCAN (ADF)
RESCAN (FB)
00
SIZE:
Letter
Legal
QUALITY:
Faint
Normal
Dark
ORIENTATION:
Portrait
Landscape
SIDES:
Single Sided
Double Sided
COMPLETENESS
Complete
See Folder
HELIOS
Check Local Repository
Principal Persona
Scan from Feeder
Scan from Flatbed
Help

The document feature allows the operator to choose from a prepared list of document types, such as correspondence, memoranda, speeches, and notes, and to assign a corresponding date. There are two reasons for doing this. First, tagging this kind of data will enable a user to restrict a search to a specified document type. Second, most archival documents do not have distinct titles. To overcome this problem and to generate a useful description of the retrieved document for the user, the document type and date can be offered as the title. Providing this kind of "fielded" information is vital for access to the material and to maintain contextual accuracy.

In addition to capturing the contextual information or metadata, this interface was developed to take into account the unique characteristics of archival documents. Prior to scanning a document, the operator must specify page size; brightness and contrast levels; whether the document is single- or double-sided; orientation of the page; and the scanner source (flatbed or automatic document feeder). When a page is scanned, it appears in an image viewer adjacent to the scanning interface, allowing the operator to determine the success of the scan and to rescan if necessary. Each scanning session is logged to monitor quality control and record scanning performance.

4.2 Optical Character Recognition

In order for the system to provide innovative searching capabilities, the images must be converted to machine-readable format. This text recognition process, commonly referred to as optical character recognition, produces a standard ASCII text file. An off-the-shelf package called TextBridge, a Xerox Imaging Systems product, is used for OCR conversion. We run the OCR program in batch mode at night to economize staff time and computer usage.

4.3 Verification

Once a complete series or subseries has been digitized and text files have been produced, members of the archives staff utilize a verification interface, dubbed HeliosArchivist, that supports image and text verification, annotation, and organization. We are using three Dell 586 PCs running Windows NT 4.0 with 20" color monitors (1280 x 1024 pixels) to perform the verification tasks. Using a client/server system, the PCs are networked to the server.

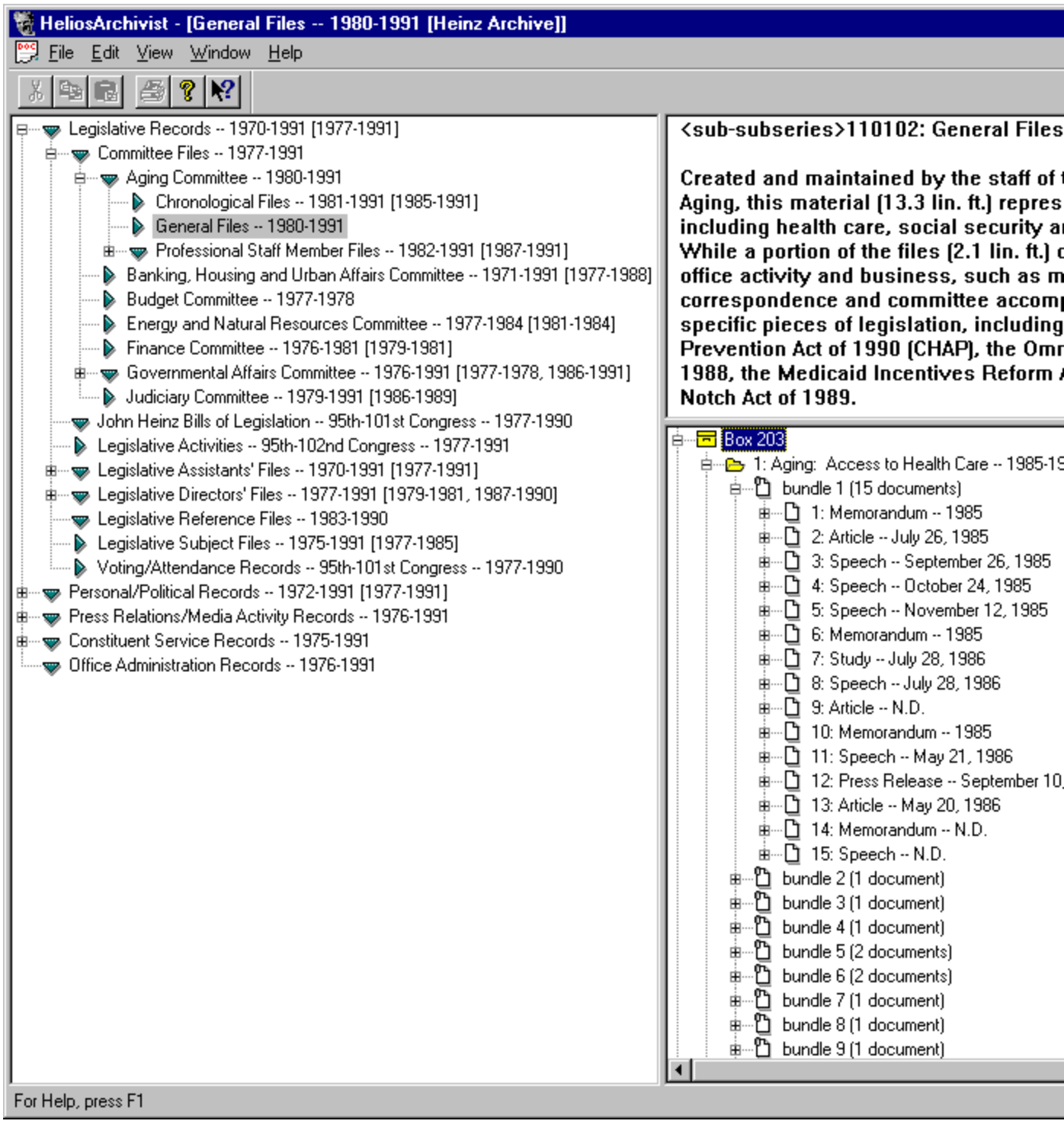
Since December 1996, graduate students and the Heinz Archivist have performed the majority of the verification tasks. As of May 1, 1998, we have verified over 495,000 images.

HeliosArchivist provides the ability to browse the hierarchical contents of the archives as well as move through the contents of a folder page by page, document by document. HeliosArchivist displays each page image and its associated ASCII text to the verifier, and it enables the verifier to:

- Verify the quality of the page image against the original page itself.

- Schedule pages for rescanning.
- Check and correct the attributes associated with each page (e.g., document type).
- Evaluate the quality of the OCR conversion for each page.
- Perform minimal editing of the converted ASCII text, perhaps keying in sections that were not converted by the OCR process, such as handwritten notes.
- Mark pages with serious OCR conversion problems so that they can be keyboarded by a typist at a later date.
- Add notations at the folder and/or document level.
- Perform other organizational tasks, including the reordering of pages, documents, or folders. (See Figures 3 and 4.)

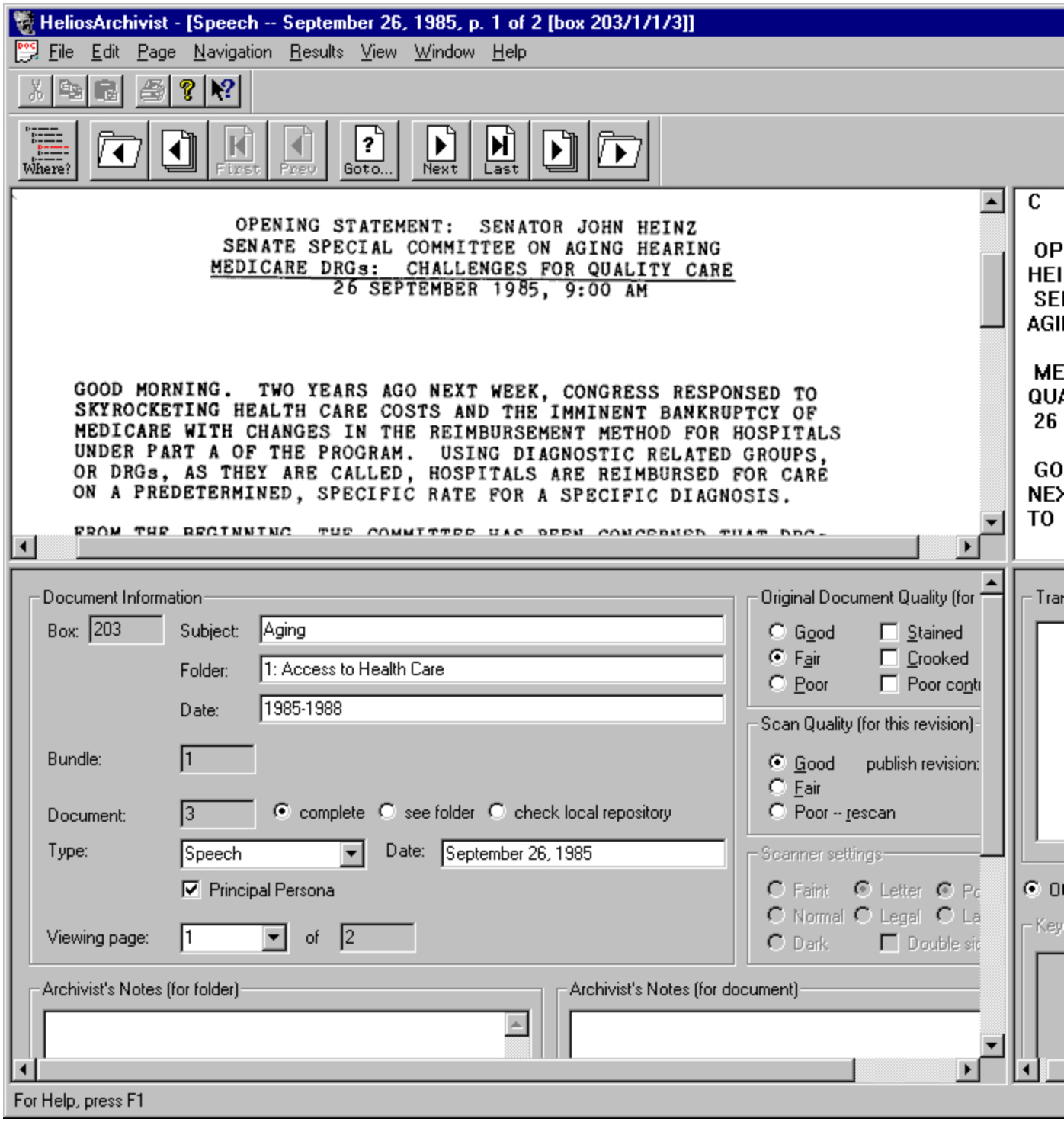
Figure 3. HeliosArchivist v2.0--Browser Window



Every page of a primary source (e.g., memos, speeches, correspondence) is verified; only the first page of a secondary source is reviewed (e.g., reports, articles). The verifiers are assigned a complete series or subseries to verify. For documents not scanned in their entirety, such as government publications, the verifiers note the availability of the complete report in a regional repository or alert the user to see the actual folder. Each document and corresponding image is rated as "Good," "Fair," or "Poor." We can also

indicate that the original document possessed one or more of the following characteristics: it was stained, crooked, and/or had poor contrast.

Figure 4. HeliosArchivist v2.0--Image Window and Document Report Card



5.0 HELIOS Server

Running Solaris 2.4, the Sun Sparcstation server (UNIX) is housed in CMU's Computer Center, a strictly climate-controlled environment not susceptible to damage from high humidity, rapid and extreme temperature fluctuations, contamination from airborne particulate matter, or power outages. In addition to storing the images and text, the server also stores the verification application, the CLARIT Retrieval for Windows application, the CLARIT index software, and the CLARIT server and CLARITweb applications.

To accommodate our rapidly growing database, we continue to add external drives to the Sparcstation; currently we have nine nine-gigabyte Seagate drives. To manage the external drives, we employ a virtual volume manager (Solstice Disk Suite). The source files that are indexed for the CLARIT server application comprise slightly more than the sum of the ASCII text files.

6.0 Searching HELIOS

Digital technology offers effective and innovative methods of providing access to and managing large bodies of heterogeneous material. However, converting material to digital format only for preservation purposes ". . . will add little value to the research process if it serves only as an alternative form of storage from which analog replicas are produced for use with conventional analytical methods." [3] Indeed, scholars will come to expect the coupling of digital research sources with the tools necessary to analyze them. Therefore, we wanted to equip researchers with traditional methods of searching and browsing an archive while adding new and robust electronic capabilities not available in the analog world.

6.1 NLP Search Engine

As previously mentioned, the HELIOS search engine (CLARIT) utilizes natural language processing technology. NLP stems from work done in the fields of computer science, artificial intelligence, and linguistics. Natural language is simply common, everyday language we use to speak and write. Natural language processing allows users to interact with a computer system, describing topics of interest using their own language as opposed to reacting to menus and prompts or using keyword and Boolean searching techniques. Consequently, they can make better use of the database with only a general knowledge of its contents.

As the HELIOS search engine, CLARIT supports more accurate, sensitive, and robust content-based indexing and retrieval than is possible with traditional "word-based" information retrieval technologies. Its indexing and retrieval capabilities are not based on locating individual words, but rather on extracting concepts that accurately characterize the content of documents. Combined with specialized statistical methods, CLARIT analyzes a query linguistically, comparing it with a similar linguistic analysis of the

actual documents in the database. We have applied CLARIT to the problem of managing compound documents (text and images) and the special requirements of archival material.

Why use NLP? Concrete disciplines, such as the medical and legal professions, often communicate and express ideas in rigorous terminology. But historians and other scholars, who use archives and historical material, approach their discipline with more imprecise language. This is why NLP technology has such promise for robust retrieval of archival material.

The Text Retrieval Conference (TREC) studies sponsored by the National Institute of Science and Technology (NIST) and the Department of Defense's Advanced Research Projects Agency (ARPA) have now demonstrated that CLARIT has a compelling advantage over traditional keyword and Boolean searching and retrieval. [4] Studies of keyword and Boolean retrieval systems have shown that they sometimes provide good precision and sometimes good recall, but never both together, and often neither. [5] The non-expert searcher (i.e., the average library user) has even less success. In addition, Boolean logic operators or special devices like adjacency and nesting are usually ignored by the general user who opts for single-term searches in hopes of getting the greatest number of retrieved items. They know from experience that they will do better by manually sifting the results and selecting relevant documents.

Efforts to enhance online records have improved recall at the expense of precision. Unless we find new tools, moving to full-text electronic access will only make matters worse. CMU believes that CLARIT is the "better mousetrap"--one that will be especially useful for accessing archival material.

6.2 CLARITweb and CLARIT Retrieval for Windows

The majority of users can access HELIOS on the Internet at <URL:http://heinz1.library.cmu.edu/HELIOS>. In addition to offering the ability to search portions of the collection, the site includes Web pages entitled: About the HELIOS Project; Frequently Asked Questions; CLARITweb User's Guide; and Restrictions on Using HELIOS.

Campus users can use a more robust searching interface called CLARIT Retrieval for Windows. It adds more powerful functions for constraining searches, displaying retrieved documents, editing queries, and more.

Either interface allows the user to select a database, submit a natural-language query (e.g., the global loss of biodiversity associated with the destruction of tropical rainforests and global warming), and review a list of retrieved documents ranked in order of their estimated relevance to the query. The "title" for each retrieved document is generated by several metadata fields: Document Type; Document Date; Subject; Folder Title; Folder Date. Once the image is displayed, the user can move through the list of retrieved documents or move forward or backward through any level of the collection (e.g., move

to the next or previous page of the document or the next or previous document in the folder).

Each interface offers tools to improve query results. The "enhance query" feature extracts related terminology from selected documents; these terms are generated "on the fly" by CLARIT. This feature allows the documents to describe themselves and eliminates the need for pre-existing indices. A second feature allows the researcher to edit the query by adjusting the weight of each search term. Another feature allows the researcher to use an existing page as an example query to locate more documents like it.

The user can limit or constrain a search to particular fields in the database. However, CLARITweb only allows specification of the following fields in a search: Document Type, Document Date, or Principal Persona (e.g., affiliation with John Heinz).

Unless otherwise specified, CLARIT searches the entire collection; therefore it is crucial for the user to be able to see the context in which the retrieved documents were created. Each HELIOS interface features a context button ("Where am I?") that creates a hierarchical view of the archives structure, placing the displayed document in its proper context. It presents the name and date of the folder from which a retrieved page originated, including the name of its subgroup, series, or smaller unit. It also allows the user to browse the inventory of any series as well as read the series descriptions. In this way the user interfaces incorporate the traditional methods of performing archival research that maintain the context in which documents were created.

Either interface allows the user to examine the metadata, archivist notes, quality assessments, and ASCII text. Captured for each image, the metadata display contextual information which includes Title, Document Type, Document Date, Bundle Number, Folder Number, Subject (if applicable), Folder Title, Folder Date, Box Number, Subgroup name, Series name, Subseries name, and more. In addition, the metadata show the assessment of the original document, the transcriptions of selected handwritten notes or failed OCR results, the archivist's notes at the folder and/or document level, the raw OCR output, and the indication of the availability of the complete document elsewhere.

6.3 Focus Groups and Protocol Testing

Shortly after the release of HELIOS on the World Wide Web, CMU Libraries would like to begin conducting formal user protocol testing to provide concrete data about how researchers actually approach and use the Web-based interface and to make changes as needed. In focus groups, users will often describe what they think they need, but protocol testing will show that they actually want something else.

7.0 Potential HELIOS Benefits

The HELIOS project team anticipates that the system will have a number of potential benefits:

- It will allow users to find archival information quickly and efficiently. Because of the overwhelming amount of material that is often present in congressional archives, research is often a result of an extremely time-consuming manual "hit or miss" research method. Using HELIOS, users will eliminate the need to wade through pages and pages of less significant material in search of those "golden nuggets." Scholars will be able to focus their efforts more on exploring new ideas, comparing and contrasting new relationships, and drawing conclusions, rather than on performing endless hours of manual research.
- It will provide uniform and consistent access to the collection in a way that is superior to the access provided by traditional finding aids.
- It will provide subject access across the entire record group, series, subseries, and folders, making the collection accessible in a variety of ways.
- New series, which in the past received little research attention due to unmanageable bulk or perceived irrelevance of folder titles, will be easily accessible.
- Many archives users simply do not have the time or money to travel to distant repositories to conduct research. Remote users will be able to access both HELIOS and a finding aid via the Internet using World Wide Web browsers such as Netscape. Consequently, the archives' location and operating hours will no longer be a concern.
- Many potential users of archives avoid them because of poor finding aids, excessive bulk, and time constraints, turning instead to secondary sources of information. HELIOS will encourage these traditional users to conduct more archival research, and it will attract new types of users.

8.0 Conclusion

By effectively utilizing imaging, OCR, and natural language processing technologies, the HELIOS project promises to dramatically transform the Heinz Archives' services by providing researchers with state-of-the-art electronic access to archival source materials. The HELIOS project is building a prototype of the digital archive of the future. It is to be hoped that it will be one of many similar projects that will make archival information instantly available to users across the globe, offering them advanced information retrieval capabilities that significantly enhance their research activities.

Notes

1. This article is a revised version of "The Heinz Electronic Library Interactive Online System (HELIOS): Building a Digital Archive Using Imaging, OCR, and Natural Language" <URL:http://info.lib.uh.edu/pr/v6/n4/gall6n4.html>, published in *Public-Access Computer Systems Review*, volume 6, number 4, 1995.

2. Metadata is data about data. The term refers to any data used to aid the identification, description, and location of networked electronic resources. Many different metadata formats exist, some quite simple in their description, others quite complex and rich.

3. Margaret Hedstrom, *Digital Preservation: a Time Bomb for Digital Libraries.* See <URL:http://www.uky.edu/~kiernan/DL/hedstrom.html>.

4. Donna Harmon, ed., *The Second Text REtrieval Conference (TREC-2)* (Washington, DC: Government Printing Office, 1994).

5. D.C. Blair and M.E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the ACM* 28 (March 1985): 289-299.

About the Authors

Edward A. Galloway, Heinz Archivist, H. John Heinz III Archives, Carnegie Mellon University, 5000 Forbes Avenue, Hamburg Hall, Room 2504B, Pittsburgh, PA 15213-3890. Internet: eg2d@andrew.cmu.edu.

Gabrielle V. Michalek, University Archivist, University Libraries, Carnegie Mellon University, 5000 Forbes Avenue, Hunt Library - University Archives, Pittsburgh, PA 15213-3890. Internet: gm1l@andrew.cmu.edu.

About the Journal

The World Wide Web home page for *The Public-Access Computer Systems Review* provides detailed information about the journal and access to all article files: <URL:http://info.lib.uh.edu/pacsrev.html>.

Copyright