1-1-1995

The Heinz Electronic Library Interactive On-Line System (HELIOS): Building A Digital Archive Using Imaging, OCR, and Natural Language Processing Technologies

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

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

The Public-Access Computer Systems Review

Galloway, Edward A., and Gabrielle V. Michalek. "The Heinz Electronic Library Interactive Online System (HELIOS): Building a Digital Archive Using Imaging, OCR, and Natural Language Processing Technologies." *The Public-Access Computer Systems Review* 6, no. 4 (1995).

1.0 Introduction

In February 1994, Carnegie Mellon University (CMU) embarked on an ambitious project to convert one million pages of the congressional papers of Senator John Heinz (R-PA) into digital format and to provide access to these papers through innovative information retrieval software developed at CMU. Named in memory of the late Senator, the Heinz Electronic Library Interactive Online System (HELIOS) supports full-page digital images and it utilizes natural language processing (NLP) technology to search large quantities of unstructured text. HELIOS will allow researchers to access the Heinz papers through the campus network as well as through the Internet.

Over one million dollars was donated by the Heinz Family Foundation, Heinz Company Foundation, and Heinz Endowments to support the establishment of the H. John Heinz III Archives and the digitization project. Heinz assistance has made it possible to advance the principles of digital preservation and access for archival collections. In addition to the Heinz gift, CMU has committed an additional $450,000 in matching resources to the project. These resources primarily come in the form of permanent full-time staff salaries, archival equipment, and rental of a processing facility.

Our goal is to develop a digital archive that will serve as a model for the archival profession. We expect to create an archival information technology environment that dramatically increases the depth of indexing and the quality of retrieval beyond what archiving resources have traditionally allowed.

To create the HELIOS database, documents are scanned, converted to ASCII form via OCR, verified and organized, and indexed using the CLARIT natural language processing software. The project will develop three graphical user interfaces in a Microsoft Windows environment: a scanning interface, an archivist/verification interface, and an end-user interface.

HELIOS represents a significant breakthrough technology that has the potential to transform the work of archivists by helping them to overcome the significant challenges they face, including an inability to: 1. create good finding aids and indexes for paper archives that provide deep access to collections, 2. provide effective retrieval from paper archives due to the inherent diversity and size of these one-of-a-kind files, and 3. offer broad public access to archives because they represent resources that the researcher must visit in order to use effectively.

Archivists have resisted the use of information technology because they lack appropriate tools to automatically process large amounts of text for retrieval. HELIOS will offer such a tool.

Clearly, there are many problems yet to be solved in the management and preservation of digital archives, but it is CMU's intention to work with the larger archival and library community to help establish standard practices for digitizing paper archives and to develop the information management tools to give scholars and students state-of-the-art access to them.

2.0 H. John Heinz III Congressional Collection

Shortly after the tragic death of Senator Heinz in 1991, the family placed the congressional papers at Carnegie Mellon University to serve as the research centerpiece for the Heinz Graduate School of Public Policy and Management. CMU spent $70,000 to prepare an archival facility near campus in anticipation of receiving the collection. Upon completion of the space, the collection was transferred from its storage facility in Harmarville, Pennsylvania to CMU.

In addition to documenting Heinz' tenure as a three-term member of the U.S. House of Representatives, the papers focus on his fifteen-year Senate career. Senator Heinz earned a national reputation based on his work on retirement and aging concerns, health care, international trade and finance, human development, 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.

The H. John Heinz III Archives will provide both traditional and electronic access to the papers. The Heinz Archives staff uses conventional processing methods to arrange and describe the papers while applying fundamental preservation techniques to the original material to ensure longevity. The material is housed in proper environmental conditions suitable for long-term preservation. To date, over 500 of the original 1,200 cubic feet of material has been processed. At the completion of processing, the collection should comprise approximately 650 to 700 cubic feet of material. The HELIOS project will provide electronic access to the most important series and subseries in the collection.

3.0 HELIOS Team Members

The HELIOS project is comprised of three umbrella units representing several disciplines. While bringing its own expertise to the creative process, each unit is making major contributions to the design, creation, and implementation of HELIOS.

3.1 Laboratory for Computational Linguistics

CMU's Laboratory for Computational Linguistics (LCL) focuses its research efforts on information retrieval issues. LCL researchers developed efficient methods to analyze and extract language using computers, and this NLP research is the basis of the CLARIT software.

3.2 CLARITECH Corporation

The CLARITECH Corporation, a CMU spin-off company, has improved and marketed LCL's NLP technology, dubbing it CLARIT. CLARITECH's primary contribution to the HELIOS project is system design. It is responsible for incorporating elements of the CLARIT system designed in the LCL into HELIOS, for creating three graphical user interfaces for the system, and for supporting the development of new CLARIT tools for HELIOS throughout the duration of the project.

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 developing the HELIOS client/server system, maintaining the system, training users, and documenting the 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 quality 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 three 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.

4.0 HELIOS Document Processing

In the HELIOS project, the original paper documents are processed based on the typical arrangement scheme of a congressional collection. After this is completed, document pages from complete series and subseries are scanned, resulting in 400 dpi image files. The image files are then converted to ASCII text files using the TextBridge OCR package. The images and text are verified and additional notes and organization added. The text is indexed by the CLARIT natural language processing software, resulting in the searchable HELIOS database.

4.1 Scanning

Once an entire series has been processed, the documents are transferred into electronic format. With the use of a 66 MHz 486 DX2 PC with a 2 GB hard drive, 20" monochrome monitor (1600 x 1200 pixels), and a high-end Fujitsu scanner, we are creating 400 dpi bitonal TIFF images. The images are compressed using CCITT Group IV, an international compression standard, and backed-up onto 4 mm digital data storage tapes.

Because the scanning procedure represents the most crucial aspect of the project, we designed a scanning interface to facilitate the rapid scanning of documents while capturing essential contextual information for the end-user.

The scanning interface imitates the standard archival collection arrangement, organizing documents into subgroups, series, subseries, and smaller units. Operators select the appropriate level using drop-down menus. The operator then enters the box and folder number as well as the folder title and date. The scanning interface was also designed to capture "bundles"; that is, groups of documents 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.

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 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, 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 quality levels; whether the document is single- or double-sided; orientation of the page; and the scanner source (flatbed or feeder). When a page is scanned, it appears in an image viewer, 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 (OCR)

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 (OCR), produces a standard ASCII text file. An off-the-shelf package called TextBridge, a Xerox Imaging Systems product, is used for OCR conversion. The HELIOS system designers have tweaked TextBridge to run the OCR process in batch mode at night to economize staff time and computer usage. They have also implemented a CLARIT tool to perform post-OCR correction, thereby boosting the accuracy of the OCR process to an even greater degree.

4.3 Verification

Once each page image has a corresponding text file, the archives staff will utilize an archivist/verification interface that is presently being designed to support image and text verification, annotation, and organization. To perform the majority of the verification tasks, operators will use a Sun SparcStation 20 workstation with an 18 GB external hard drive and 20" color monitor (1280 x 1024 pixels). This workstation will also serve as the HELIOS search engine server and file server. Two Sun SparcStation 5 workstations will be used to perform additional verification.

The archivist/verification interface will display each page image and its associated ASCII text to the operator, and it will:

- Enable the operator 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.
- Evaluate the quality of the OCR conversion for each page.

- Perform minimal editing of the converted ASCII text, perhaps keying in sections that did not respond to 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.
- Create links to other groups of documents.
- Perform other organizational tasks, including the reordering of pages or folders.

The Heinz Archives Assistant and graduate students will perform the majority of the verification tasks. For documents not scanned in their entirety, such as government publications, operators will note the availability of the complete report in a regional repository or provide some other explanation. For poor or skewed images, operators will indicate that the original page possessed these characteristics.

The Heinz Archivist will provide additional organizing, indexing, and annotations. Descriptive or structural notes will be added to various levels of the electronic collection. The Heinz Archivist will be able to organize the electronic archives along several different dimensions apart from its processed arrangement. For example, a taxonomy of related terms describing any level of the collection could be constructed. Each series will be linked to its appropriate inventory and description, and cross-reference notes will be established to link records to similar groups of material.

4.4 NLP Search Engine

The HELIOS search engine utilizes natural language processing (NLP) 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 DOD's Advanced Research Projects Agency (ARPA) have now demonstrated that CLARIT has a compelling advantage over traditional keyword and Boolean searching and retrieval. [1] Studies of keyword and Boolean retrieval systems have shown that sometimes they provide good precision and sometimes good recall, but never both together, and often neither. [2] 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.

4.5 Prototype End-User Interface

One year ago, approximately 20,000 pages of archival material related to the work of Michael Lockerby, a Legislative Assistant (1977-1981) specializing in environmental issues and legislation, were used as a testbed for initial scanning and CLARIT retrieval.

This scanning project allowed us to assess the physical demands of scanning archival documents and to determine the strengths and weaknesses of off-the-shelf technology used in HELIOS. The resultant document database was very useful in conducting focus groups on the prototype HELIOS end-user interface, and it was also used for HELIOS demonstrations at other sites.

The prototype end-user interface has four major windows.

1. *Query Window.* The user enters a natural language search in this window, such as "the economic impact of environmental regulations and policies on the steel industry."
2. *CLARIT Results Window.* This window shows the list of retrieved pages. The pages are ranked in order of their estimated relevance to the query. The Change column indicates the upward or downward movement of pages based on a prior query search (e.g., +12 or -8), and an asterisk indicates that the system retrieved a new page. The Document Type and Date fields are used to generate a useful description of the retrieved page, such as "Memoranda--May 5, 1990" or "Speech--November 12, 1990."
3. *Document Window.* This window provides an ASCII version of the page with highlighting of retrieval terms associated with the query.
4. *Image Viewer.* This window displays the bitmapped image of the original page retrieved for the search.

The end-user will have the ability to move forward and backward through any level of the collection (e.g., move to the next or previous page, the next or previous document, or the next or previous folder). Since document types will be tagged, the user could restrict a search to a specified date and document type, such as correspondence, memoranda, or speeches.

CLARIT offers several tools to improve query results. One feature can extract related terminology from the actual 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 use an existing page as an example query to locate more documents like it. A third tool incorporates a hypertext feature to link other relevant portions of the collection together.

Unless specified, CLARIT searches the entire collection regardless of the arrangement and contextual framework imposed by the archivist. Therefore, it is crucial for the user to understand the context in which the retrieved documents were created. The HELIOS end-user interface will present the name of the folder from which a retrieved page originated, and it will allow the user to browse the inventory of any series as well as read the series descriptions. Therefore, the user interface will incorporate the traditional methods of performing archival research that maintain the context in which documents were created.

Much has already been learned from initial focus groups about what modifications are needed to make the end-user interface more user friendly. Users commented that some terminology was not easily understandable. For instance, they were not sure that "score" correctly expressed the meaning of the rank order or even needed to be shown. Terms like "augment" and "parse," common though they are in linguistics, are not clear to end-users. With respect to the structure of the interface, users want to save queries and reuse them.

Users also commented on what should be displayed in the list of retrieved titles. The initial focus groups discovered that users are confounded by a system that can retrieve a variable number of documents based on relevance scores. They tend to want "all" of the retrieved documents; however, this is probably an artifact of past experience in keyword and Boolean retrieval systems. Similarly, they expect the use of "not" to be Boolean in its effects, though this effect is not achieved when retrieval is only on nouns and noun phrases.

Upon completion of the initial version of the end-user interface, CMU will begin conducting formal user protocol testing to provide concrete data about how researchers actually approach the user interface, and to make changes as needed. In focus groups, users will often describe what they think they need, but in protocol testing they actually want something else. Additional tools and features will be built and introduced as users become more experienced with the system.

5.0 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 methods. 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 relevance 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 Mosaic. 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, instead turning to secondary sources of information. HELIOS will encourage these traditional users to conduct more archival research, and it will attract new types of users.

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

[2] 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, Carnegie Mellon University, 5000 Forbes Avenue, Hamburg Hall, Room 1506, Pittsburgh, PA 15213-3890. Internet: eg2d@andrew.cmu.edu.

Gabrielle V. Michalek, University Archivist, Carnegie Mellon University, 4825 Frew Street, 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: http://info.lib.uh.edu/pacsrev.html.

Copyright