The Heinz Electronic Library Interactive On-Line System: A Multi-Disciplinary Approach

Gabrielle Michalek
Carnegie Mellon University, gm11@andrew.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/lib_science

Part of the Library and Information Science Commons

Published In
Congressional Papers Conference, Portland, Main, 109-116.
PROCEEDINGS

Edited under the direction of Gregory P. Gallant, Director, Margaret Chase Smith Library and William E. Brown, Jr., Head of Archives and Special Collections, University of Miami (FL)

Sponsored by

Northwood University
Margaret Chase Smith Library
The Heinz Electronic-Library Interactive On-line System: An Multidisciplinary Approach

Gabrielle V. Michalek
University Archivist, Carnegie Mellon University

INTRODUCTION

This year Carnegie Mellon University began work to develop a fully electronic historical archive of the papers of the late Senator H. John Heinz III. The project is named HELIOS, the Heinz Electronic-Library Interactive On-line System, in honor of the late Senator. Carnegie Mellon received over one million dollars from the Heinz Family Foundation, Heinz Company Foundation and Heinz Endowments in support of the HELIOS project. The grant supports the establishment of the H. John Heinz III Archives where the Congressional papers of the late Senator are being preserved and digitized to make them electronically accessible to students, faculty, public-policy professionals and researchers. Edward Galloway, the Heinz Archivist, will describe the Heinz Archives and the HELIOS project in his presentation. The present paper will concentrate on the members of the HELIOS project team and their intellectual contributions with special emphasis on Carnegie Mellon's multidisciplinary approach to creating an electronic archives.

Imagine two scenarios that depict the retrieval of archival information. In the first, Captain Kirk of Star Trek fame walks into the library of the U.S.S. Enterprise and presses the solitary button to the ship's library computer system. "Check the linguistic banks for the term rejacket and other related terms," he asks. Instantly the computer responds with the precise information that Captain Kirk is seeking. The information is presented in a manner that can be understood by the layman. Kirk, confident that he knows everything he needs to know about the situation at hand returns to the bridge and takes the proper course of action to save the Enterprise and the Federation from the insidious rule of the Klingon Empire.

Now imagine a quite different scenario. A phone rings in the office of the University Archivist. "Hello, this is Joe from the Office of University Automation. I was hoping you could come over here and help us with a problem we are having. We are running out of space, and would like to clear some things out. The problem is we do not want to throw anything away unless we are sure it does not contain any information of value. It would be a big help if you could give us some advice." The archivist, pleased to make a contribution to the university, arrives at the warehouse facility and soon discovers thousands of magnetic tapes, floppy disks, and hard drives in all sizes and formats. In addition, there are computer printouts loaded with coded information, as well as computer manuals and system logs that were intended to accompany a wide variety of software, hardware, and operating systems. The archivist learns that the oldest material in the collection is a set of thirty year old 7-track mag tapes that have not been inventoried.
and that the most recent materials are the weekly campus-wide system backups done on 8 mm cartridge. The rest of the collection represents tapes with little or no supporting documentation, documentation without the correlating tapes, tapes created on hardware that no longer exists, and manuals for software that have not been seen on campus for decades.

While those of us who manage collections of information surely envy the unreal situation depicted in the first scenario, current reality is far closer to that of the second. The question archivists face is how can we provide access to the precise information that our patrons are looking for and provide this information in a clear, concise, and timely manner? How, too, can we preserve this information so that archivists of the 21st century do not find themselves facing the same dilemma of their 20th century counterparts?

At Carnegie Mellon, we have concluded that resolving the challenges of managing and preserving an electronic archives requires the support of several disciplines. Using a multidisciplinary approach to address these issues offers several advantages. First, it distributes the burden of managing a complex automation system across several disciplines which may be better equipped at solving some of the problems encountered. Second, it allows us to take advantage of what each discipline has to contribute to the project while minimizing any biases that a discipline may have. For example, computer scientists bring an understanding of the hardware and software that make up computing systems. They also possess an intrinsic desire to utilize new technology to create these systems. Since computer scientists, however, prefer to look forward rather than to the past; the result is frequently a lack of recognition of the need for preservation. The archivist, on the other hand, sees preservation as a fundamental responsibility of the profession, because the archivist must be concerned that the collection he/she is responsible for be available to researchers in perpetuity. The archivist may be reticent to try new technology out of fear that the technology may fail or disappear altogether. Bringing these two professionals together, and encouraging them to share their professional concerns, may well influence the computer scientist to begin to consider issues like long-term preservation when designing new computer systems. Now that the discipline's concern for preservation is acknowledged and addressed, the archivist will begin to consider new technological approaches to resolving an archival concern.

In the past, archivists did not actively participate in, or tended to play a peripheral role, in the design and creation of information management systems. Teams established to build these systems typically did not include an archivist. However, by becoming a team member the archivist contributes an understanding of the issues and questions that are important from an archival perspective. Furthermore, if the archivist does not make a contribution to this activity, a competing unit interested in information management will. There is no guarantee that the competing unit will value the same concerns as the archivist. The consequence is the chaos portrayed in scenario two. It is precisely because of the reality of scenario two that archivists must play a central role in the development of information management systems. The HELIOS project provides a prime example.
HELIOS Project Team

The HELIOS project is comprised of three umbrella units that each represent several disciplines. The Laboratory for Computational Linguistics (LCL) combines the fields of computer science and linguistics. CLARITECH Corporation incorporates the work of the LCL and brings with it additional areas within computer science. The University Libraries is represented by Library Administration, and the departments of Library Automation and the University Archives. While bringing its own expertise to the creative process, each one of these umbrella units is making major contributions to the design, creation, and implementation of HELIOS. The three principal investigators are involved in other research of this nature: Dr. Charles Lowry, University Librarian; Dr. David Evans, Director of the Laboratory for Computational Linguistics; and Michael Horowitz, Senior Systems Architect, CLARITECH Corporation.

Carnegie Mellon University Computing Environment

A fundamental component of the project is Carnegie Mellon University itself. The University 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 for a processing facility.

A further Carnegie Mellon contribution is the computing infrastructure of the campus. Carnegie Mellon has ethernet bandwidth access to the NSFnet national backbone, an internet data network physically located on the campus, which allows for the high speed transfer of information in and out of the network. Therefore, archivists at Carnegie Mellon can provide full-text electronic images of archival documents to users effectively and efficiently. In addition, the entire campus is laid with fiber optic cable necessary for distributive computing which enables a user to access information from remote locations. As a result, researchers will gain access to the information in the Heinz Archives without traveling to Pittsburgh.

Laboratory For Computational Linguistics

The Carnegie Mellon University Laboratory for Computational Linguistics' fundamental contribution is the use of linguistics and its application in information retrieval. The staff of the LCL are the scientists for the project. They are responsible for the original research and development for the foundation version of the CLARIT software. The commercial version of CLARIT will serve as the 'search engine' for HELIOS and is based on Natural Language Processing (NLP) which stems from work done in the fields of computer science, artificial intelligence and linguistics. Researchers at the LCL interested in Artificial Intelligence and Computer Science developed efficient methods using computers to analyze and extract language. Natural language is human language, as
spoken such as English, French and German. Natural language processing allows humans to interact with a computer system using their natural language as opposed to reacting to menus or prompts. In this way patrons can make better use of the system's full capability with only a general understanding of the system. The research done at the LCL became the model on which the CLARIT software is based. In addition, the LCL will enhance the optical character recognition (OCR) process in which electronic images are transferred to machine readable text.

Dr. David Evans, Director of the Laboratory for Computational Linguistics, is responsible for much of the research in the LCL. He is also the Head of CLARITECH Corporation, a spin-off company affiliated with the University.

CLARITECH CORPORATION

CLARITECH Corporation's fundamental contribution is system design. The CLARITECH staff are the engineers of the project. They are responsible for incorporating elements of the CLARIT system designed in the LCL into HELIOS, for developing three graphical interfaces for the system, and for supporting the development of new CLARIT tools for HELIOS throughout the duration of the project. The activities performed by the Corporation are primarily based on the work done by computer scientists, such as Michael Horowitz who is responsible for much of CLARITECH's work on the HELIOS project. CLARITECH's contributions are in the areas of understanding and applying the computer programming used to write the software enhancements for the CLARIT system, and of knowledge of available hardware and software and how they should be configured in the system architecture. Most importantly, CLARITECH recognizes the need of readiness to experiment with new technologies and to apply them to different situations.

Carnegie Mellon University (CMU) Libraries

Three different units within the University Libraries play a major role in the multidisciplinary functioning of the project.

CMU Libraries Administration

CMU Library Administration is responsible for providing the leadership function for HELIOS as well as the fiscal management of the project. The multidisciplinary approach to HELIOS makes the role of Library Administration vital to the successful completion of the project. Charles Lowry, University Librarian, provides leadership by seeing that decisions are made that lead to effective and efficient results. He introduces a librarian's appreciation for information management combined with a global view of direction for information networks.
CMU Libraries Department of Automation

The CMU Department of Library Automation, is responsible for specifying and evaluating the three HELIOS interfaces; integrating HELIOS with the University Libraries Information System (LIS), so that LIS will act as a gateway to HELIOS, maintaining the HELIOS system, training users and documenting the system.

The department has had substantial experience creating, maintaining and providing instruction for the use of very large full-text databases. Project Mercury, launched in 1986 by the CMU Libraries, was the first step toward the creation of an electronic library capable of delivering large bodies of electronic text over the Internet, or to use the popular term, information superhighway." Researchers gain access to these images through the Library Information System (LIS), also developed by the Library Automation Department. However, the major contribution made by Library Automation is in the area of distributive computing. Distributive computing eliminates the need for a mainframe computer by dispersing some computer functions such as processing and storage. Eliminating the need for a mainframe simplifies system maintenance and makes a large scale project such as HELIOS economically feasible.

Representing Library Automation in the HELIOS project is Denise Troll, Head of Research and Development for Library Automation. Mrs. Troll's background is in Rhetoric, not Computer Science, and she is personally responsible for testing and evaluating software. She does so by means of protocol and user transaction log analyses, tools more commonly used in Rhetoric. These tools help the system designers understand the needs and expectations of the system users. This information is then used to modify the system design so that it becomes more useful to the users. These tools will not only identify design flaws in the system, but alert system trainers to where the user will encounter problems. Training methods can then be prepared in anticipation of these human snags. Staff members from the University and Heinz Archives are participating in preliminary protocol analyses conducted by Mrs. Troll.

CMU University and Heinz Archives

The third unit from the University Libraries that is participating in the HELIOS project comprises the University and Heinz Archives. The Heinz Archives is a unit within the University Archives. The Archives are responsible for the appraisal and processing of the original Heinz Collection, scanning the original material into electronic images, transferring the electronic images to machine readable text (ASCII format), cataloging the collection into OCLC, providing reference services for the collection, and preserving the original collection in perpetuity. The major contributions of the archivists fall into several categories: facilitating the multidisciplinary teamwork, establishing control over the collection, providing context for the collection, providing quality reference, dissemination of information via informational networks such as OCLC, and long-term presentation.
Facilitating teamwork provides an interesting and new approach for the archivist. Traditionally, archivists have received collections into their archives and have been solely responsible for their presentation and access. The archivist's job was done with very little input from other disciplines. For a project that involves the design, creation and maintenance of an electronic records system, the archivist not only needed to become part of a complex team, but needed to facilitate those relationships. The archivist needs to help identify team members from other disciplines with related concerns and work with these professionals to ensure that the archival concerns were being addressed in the project. In this type of work environment open communication is essential. This necessitated the use of a great many meetings and work groups where ideas could be shared.

The second major contribution of the archives, establishing control, is traditionally accomplished by creating inventories at the record group, series, and occasionally, the sub-series level of the collection. For the HELIOS project the archivists provide expertise on how the inventory process works. Portions of this process have been automated in the scanning interface to serve as one of the information management tools.

For the HELIOS project approximately one million pages will be scanned into electronic format. The most important contribution of the archivist is to place each record in the context of its creation, use and potential use. This must be done while maintaining the integrity of the content of the original record. Integrity of the records is done through the use of a quality control process whereby the archivist visually inspects and compares the scanned images and the ASCII version with the original documents. The simplest way of providing context is by allowing the user to access the surrounding records by navigating backwards and forwards sequentially through the documents.

Another way in which the records will be placed in context is through the use of virtual links. Records will be intellectually linked in a number of ways. The first is to link a record to other similar records in the collection. For example, an archivist processing a congressional collection may identify a large series of materials that pertain to health care for the elderly in a collection of Governmental Affairs Committee papers. After processing several hundred more boxes of materials, the archivist discovers another series of materials dealing with background information on social security in America. Contained within that series is a subseries of materials pertaining to the social security program paying for health care for the elderly. The archivist should recognize this relationship in information and create an intellectual link between the two parts of the collection. Virtual links will also be used to link a record to the report or group of documents that it was originally part of, to the file folder level, series level and record group level, by linking it to the OCLC record.

Another contribution of the archivist is the identification and understanding of user groups. The archivists are working with Mrs. Troll on the user study of the HELIOS pro-
ject. In this study, readily identified archives users such as historians and other archivists, will be identified and studied. However, an important aspect of the project is to identify and study groups of potential users, including high school students and public policy makers. In addition to identifying who our users are or should be, this study is aimed at understanding how the use of an electronic records system effects the users' work. Some of the questions that need to be addressed include: does having information in electronic format allow the user to work more efficiently? Based on user needs and expectations, can we modify the system to make the system work more efficiently? Should we be developing new educational tools that will help patrons use the system more efficiently? Other than the work of Paul Conway and Karen Dawley Paul, in The Documentation of Congress, very little has been done in the way of user studies for archives. Much more research in this area needs to be accomplished. Having archivists as an integral part of the HELIOS project team has made it possible to have the need for this type of research acknowledged and integrated into the HELIOS research agenda.

The archivists on this project have contributed their subject expertise as well. Prior to coming to Carnegie Mellon, Edward Galloway worked at the Senator Tower Archives at Southwestern University. There he became familiar with congressional collections and the type of information contained in them. Now with two congressional collections under his belt, he is well equipped to provide expert reference service to our patrons. His reference services will be enhanced through the use of the automated system. A major benefit of the system is that it utilizes its natural language processing component to extract noun phrases which provide conceptual views of the collection. The archivist can use his/her expertise to generate many different views of the collection depending upon the nature of the research.

Another archival concern that has been incorporated into the design and implementation of the HELIOS database is the need for long-term presentation. The archivists are presenting the more traditional paper collection, but, since they also insisted that presentation of the electronic collection was a requirement, this concern was addressed in the system design. The necessity for presenting the electronic collection presented some interesting challenges for the HELIOS project team. As recommended by the chief archivist, the system design had to conform to as many formal, industry or de facto standards as possible. For example, scanned images are placed in TIFF files and OCR'ed into ASCII text. This will ensure the ability to migrate information as the system evolves.

**Advantage/Disadvantages To Approach**

Creating an electronic archives is a relatively new experience. Because there is no well-established model for this type of project available to us, we have been inventing a process as we proceed. Using teamwork as an approach to problem-solving has proven to have some disadvantages as well as advantages. The advantages far outweigh the former, however.
Combining professionals from a diverse body of disciplines has allowed the sharing of information and exchange of ideas, thus increasing the number of possible solutions to any problem. Creating a team has spread the perceived responsibility for the problem. For example, providing access and presenting archival materials is now considered an undertaking of the three units within the university, and not just the problem of the archives. This broadening perspective has also led the team members to gain an appreciation for the work of each unit, which in turn breaks down professional barriers. Because the archives has allied itself with other more powerful disciplines, such as Computer Science, the archivist may witness an increase in powerbase. The archivist is no longer viewed as the keeper of old stuff, but as a professional capable of making a valuable contribution to the future of information management.

The disadvantages lie mainly in the realm of communication. Obviously when working on a problem as complex as creating an electronic archives with many individuals involved in the process there are going to be communication breakdowns. Keeping the lines of communication open is both time consuming and requires a great deal of energy. However, the reward of a system that satisfies the needs of its users justifies those expenditures.

**Conclusion**

For the HELIOS project, we have constructed an multidisciplinary team to meet the challenge of creating an electronic archives. Financial support for the project came from the private Heinz Foundations, Carnegie Mellon University, and a spin-off company CLARITECH Corporation. The university also partnered by providing the necessary infrastructure. The disciplines represented in the project include: Computer Science, Artificial Intelligence, Linguistics, Rhetoric, Librarianship, and Archives Management. Clearly, Computer Science plays a significant role in the project through its involvement in the development of new software and creation of system architecture specially designed for HELIOS. Aspects of Computer Science were combined with Artificial Intelligence and Linguistics to create the search engine for HELIOS called CLARIT. From the discipline of Rhetoric, we have borrowed the tools for testing the system such as protocol analysis. Leadership for the project flows from the University Libraries. As we have seen, professionals from other disciplines are better equipped to meet some of the challenges posed by creating an information management system. While each of these disciplines is making important contributions to the project, the archivist's role must be central. When the professional concerns of the archivist, such as access, context and presentation, are not integrated into the design of the system, the result is a system that does not meet the needs of the archivists or the archives users. When the archivist exchanges his/her professional understanding with those of other disciplines, the archivist will begin to make a contribution to longterm solutions to electronic records management.