Cyber-archeology
by Ross A. Wirth, Ph.D.
2004

Note: this document was prepared as part of a case study using cyber-archeology to investigate the digital archive left behind by those involved in an organizational change initiative. Since parts of the original paper have been cut from this document, there may be some disjointed sections here.

Introduction to Cyber-archeology

Man has historically generated material objects that have significance within the culture of the time. Years later, archeologists can study the artifacts and draw conclusions about the society that no longer exists. Similarly sociologists study cultural artifacts in current use as a way to better understand the values and beliefs operating within the social organization. In modern society, knowledge plays an increasing role as an enabler of increased productivity and at the same time is generating new artifacts that define the culture of the day. Increasingly many of the artifacts that help define us are being created in a virtual form and stored on the various intranets and content management systems of today's organizations. These documents support daily operations within the organization having replaced much of the paper documentation that was used in the past. These online knowledge systems also contain many of the records that were previously stored in remote records retention centers. However, a critical difference exists between the current electronic storage and the earlier paper documents, a difference beyond being virtual versus physical. This is the ease which all individuals and project teams can establish virtual documents for their activities. Such ease also extends to the open availability of information previously hidden within departmental file cabinets. Now, everybody is expected to use the content management system for storage of important documents and access limits are established only for confidential documents. Any document not limited to the document creator and key user is openly available for any organizational member to read, which they do as evidenced by the access logs in the content management system. Collectively, the material being stored on the intranet and within content management systems defines what the organization believes and values for retention. The material is in most cases freely available to all members of the organization with a taxonomy that both enables finding specific information and discovery as someone surfs from one document to another through the virtual connections on the intranet and the within the structure of the folders in the content management system.

Through the stored material, an organizational culture is documented within the intranet and corporate content management system that includes some material from the older network servers. However, this cultural documentation is not explicitly known having been built over time by a number of people. Any master plan for building the electronic structure would only be for a small piece of the entire storage system or limited to the navigation structure and templates used. The ability to control all the content itself, especially within a content management system is very limited. Further, some of the documents making up the cultural records were purposefully placed while others may be remnants of past efforts that have been forgotten and not purged. Still other documents may have been deposited by individuals or teams, incidental to their jobs and outside any formal document review process. Some of these latter documents may even be contrary to corporate doctrine having been created by a local organizational silo or change effort that has not yet reached a critical mass to overtake the larger, established culture. This material provides a wealth of information for cyber-archeology, which is the study of cultural artifacts within a virtual medium (Jones, 1997).

Cyber-archeology can be used in two ways. First is to show how the organizational culture is captured within the digital archives stored on the intranet and within the content management system. The second is to document the culture that is implicitly being communicated to employees in the organization being studied, specifically that related to leading organizational change. While most communication programs are seen as a single shot effort, the retention of that effort continues to communicate long after the original purpose. In the past, information transfer at an employee meeting was limited to those attending the meeting and limited material passed on to those who were absent. Now, the entire slide presentation and accompanying speech is available for all to see even years later. The message communicated continues to communicate to new employees who have only now joined the company. Knowledge of the culture that is inadvertently being communicated needs to be identified to prevent possible disconnects with the current desires to shape the culture today.

Hodge (2000), Lyman and Kahle (1998), and Jones (1997) discussed the extent to which our culture is increasingly being captured in digital form in addition to physical artifacts. While most of the concern focuses on the fragile nature of electronic artifacts these authors also stress the importance of such artifacts for full understanding of a culture where increasing interaction is taking place in virtual mediums whether directly via email and forum discussions or indirectly through document repositories and web surfing. This body of research addresses both sides of the situation. First is the attention given to long-term retention of electronic artifacts so they can be studied by future researchers. The other aspect of their studies involves understanding how culture is being documented in our electronic repositories or on the various knowledge systems in use.

When examined at the surface a corporate intranet and content management system has unknown breadth and depth of material. While the breadth of the classification taxonomy is evident through top-level hyperlinks it is impossible to easily probe the full depth of all possible hyperlinks due to the inordinate amount of time that would be required to investigate all possible connections. The frequent use of hyperlinks between parts of the intranet and content management system also creates search redundancy when attempting to track all leads. To keep the search manageable the researcher must follow links until such time as further linkages were not suggestive of finding content related to the search. Therefore, one assumption that has to be made in cyber-archeology is that the individuals establishing the hyperlinks and taxonomy do so based upon some consistency or direct relationship between the pages, documents, and folders being linked. While there is the possibility that a document or page might be overlooked with this search methodology it is unlikely that such an artifact would be critical due to the importance given the document by the individual storing it. This is probably a safe assumption to make since the study focuses on cultural artifacts as evident to employees who are likely to encounter them as they surf the intranet and review material in the content management system. Any document or web page that is not easily found is also a cultural artifact that is limited in its ability to adequately describe the dominant culture within the organization. A further assumption that is present in all case studies is the ability of the researcher to identify the essential elements that are pertinent to the study and not overlook their importance or misinterpret them relative to other parts of the research.

Prior to the content management system, documents were previously stored on the network servers. Because this storage medium did not have internal linkages it was necessary to drill down through the folder structure to find the electronic documents stored there. Without a predetermined taxonomy, the folder structure would have emerged over time as the need arose to store and share additional documents. This created difficulty in understanding the contents of network servers since the folder names are limited by both system imposed character length and naming conventions understood by those using the information stored there. With the advent of the content management system, much of this material was moved while retaining its prior departmental organization, which creates the possibility of inadvertently overlooking pertinent documents because of the way they were previously stored. Therefore an assumption must be made that any material missed in a systematic drill-down into the content management system's folder structure would be captured in a keyword search or be inconsequential to cultural understanding by not having employed the keyword within the document. It is also assumed that any material determined to be important for retention would have been specifically moved to the corporate content management system in the time the corporate repository has been available. Because the network servers are now in the corporate backwaters of online information they are increasingly being phased out in favor of other systems that are both easier to manage and more user friendly for cross-organization sharing. For the most part these network servers are increasingly being used for data storage and not for text or presentation documents. By not searching the network servers, it is assumed that few documents would be overlooked or be of an important nature.

The content management system is normally a stand alone system that is entered from the corporate intranet. Navigation in the content management system is primarily through a hierarchical taxonomy with few hyperlinks that cross the folder structure. While having a folder structure similar to that found on the network servers this content management system does permit long text to be used for document and folder names thereby adding understanding to the contents.

Definition of Terms

Cultural artifacts are products of human activity that are essential to the activity itself or representative of the beliefs and values in use. While artifacts are normally thought of as physical objects they are increasingly being created in a virtual form through computer mediated communication and storage of documents in electronic formats.

An intranet is the collection of pages or documents that are linked together using internet technology but secured behind a firewall that limits access to approved organizational members.

A content management system is a user-oriented database of documents that employs advanced access and security features beyond what is used on network servers. The content can include documents, data, images, or other information that can be stored in an electronic format.

Network servers are the file repositories that are connected to the Local Area Network (LAN) and are accessible to everyone who connects to the network and has sufficient security.

Knowledge systems are used in this paper to collectively represent the intranet and the primary, corporate content management system.

Taxonomy is the organization of information that may take the form of a hierarchical classification system or connections between related information.

Research using Cyber-archeology

Cultural values and beliefs should be reflected in the documents created by that culture both in what is recorded and the terminology used. The following questions can be used to guide the investigation.

- What do the cultural artifacts contained in the knowledge systems say about the attitudes and beliefs?
- Is there a dominant model of organization evident in the digital archive?
- Do the artifacts show consistency over time or are there recognizable changes?
- Do all the cultural artifacts originate within the formal organization, i.e., top management, corporate communications, corporate planning, or the human resource department?

Significance of Cyber-archeology

Cyber-archeology is a relatively new field of study that is growing as researchers become aware of the magnitude of information now being stored electronically and easily accessible to the researcher. Within the stored documents and files on the intranet and content management system (the knowledge systems of the organization) is a cultural record that was purposefully retained and often forgotten over time. Yet some permanence exists for these cultural artifacts until such time as they are identified for deletion. Until then they are accessible to the majority of organizational members and continue to communicate a cultural message. One area of learning is the extent that conflicting messages are evident in the cultural record. Knowledge of such conflict can better prepare those who are attempting to influence the culture in a desired direction.

However, the purposeful identification of the culture being communicated within the digital archives also permits the destruction of parts of the culture that are not consistent with the intended message today thereby destroying some piece of history that will not be available to future researchers. Future cyber-archeologists will see this as willful destruction of cultural artifacts, which brings up an interesting topic of historical correctness versus business purpose. While there has always been destruction of cultural artifacts the destruction today is much more thorough when the cultural artifact being destroyed is virtual and does not leave recognizable physical elements behind after its destruction.

Cyber-archeology – Summary

Lyman and Kahle (1998, p. 1) introduced their article by saying that “our cultural heritage is now taking digital form.” While physical artifacts can be destroyed and stories distorted over time, an electronic document can be effectively destroyed with a single keystroke. If not recovered quickly, that piece of history could be lost forever if not previously converted to hard copy and retained. Further adding to the unique nature of electronic documents is the ease at which they can be produced thereby adding to the sheer number of documents created. This ease of creating and altering documents also translates to the non-permanent nature of electronic documents. No document or web page is ever declared to be in final form, only an acceptable form until further need of alteration is realized.

Hodge (2000) addressed the issue of how to decide what is to be retained for future reference. Without such action, that piece of the cultural history could be lost to future researchers. However, the mere act of archiving digital artifacts requires a judgment call on the future value of the document. Even in the formal attempts to document the internet, only the home page of a web site is being archived. The bulk of the information published to the internet is then subject to electronic deletion or being altered into a sufficiently different content representation over time. Conversely, attempts could be made to archive everything, but this creates a content management nightmare of how the material should be organized for storage and retrieval.

In lieu of any formal approach, the document creator or steward has the responsibility to judge the future value of each document as part of routine maintenance of the knowledge base. Aiding this process of efficient storage and retrieval is the possible use of metadata to describe the contents of the document in a manner that aids future retrieval. However, this requires additional time to be spent when archiving the document, which is not likely when dealing with documents that have questionable future value. This problem of digital archiving is especially acute when trying to guess all the possible uses the document might have beyond its current use.

Even with the uncertainty surrounding what is being retained in the digital archives of an organization there has to be a realization that computer mediated communication is increasingly becoming the dominant form of communication. Jones (1997) described how anthropologists are increasingly turning to the digital archives for understanding how a culture operates. This branch of archaeology has been termed cyber-archaeology in recognition of the virtual form the cultural artifacts take. Just as archaeology studies the past by examining the material remains of past cultures, cyber-archaeology seeks to understand the past by examining the documents in use or archived in a content management system. While these digital archives are more recent than some of the physical remains often studied, there is the advantage of being able to date them better due to the technology itself. This, of course, assumes that the digital format is still in use and that current technology still exists for reading the information contained in the digital packet. An example of this problem is in retrieving information from old WordStar or VisiCalc files (Lyman & Kahle, 1998). Even if the file is still available, the researcher would need the necessary software and operating system to retrieve the information. Once retrieved to the monitor, the researcher would need the right printer configuration or other storage device for saving the information in another format. Each generation of technology permits some inter-generational transfer, but rarely more than one generation at a time.

In comparing cyber-archaeology to archaeology Jones (1997) drew a parallel between Middle-East tells and digital archives. Tells are large mounds of the remains of past civilizations. Scientific examination of this material gives great insight into the culture. Similarly, examination of the digital archives gives insight into past activities. One difference though is the ease at which the tells can be identified among other surroundings. Digital archives are effectively hidden from sight and are only found when stumbled across or a search is made.

Cyber-archeology – Experiences & Comments

In this investigation of latent content from a change initiative, the amount of material archived in the content management system exceeded expectations, but was also less organized than would have been expected. Numerous documents were misfiled, especially the speeches and slides prepared for the employee meetings, which were sometimes filed in the wrong year. Also confusing the retrieval of information was the inconsistency of document classification. In some cases material was filed by event year and at other times by the name of the speaker. The filing of draft copies also created some difficulty in that different drafts would be filed differently. Part of this difficulty is a function of the content management system's classification process, which is a single-path hierarchy. While such a taxonomy permits easy document classification it does not lend itself to easy retrieval along any other approach than the one used for the initial filing. Even though the content management system has the ability for cross-classification of documents, this feature was only seen used a few times. While keyword searches were used to find some documents the usefulness of search was greatly limited by finding a sufficiently small number of search responses that were relevant. For example, the search for "organizational change" brought up hundreds of documents that somehow dealt with organizational or change issues, but not organizational change. The issue of misclassification is something that future researchers should take into account when they plan to use digital archives as the basis of a case study. In general, this situation leads the researcher to conclude that little attention is being given to the storage nor the management of documents once stored. It is likely that such a situation would only worsen over time as still more documents are created and stored in the content management system.

On the other hand, the intranets are very well organized in their display of current information. However, as Hodge (2000) and Lyman and Kahle (1998) warned, there is great danger in losing historic records as currency of information is maintained. On the intranet, the only historic record was the maintenance of the duplicate set of copies of the internal employee magazine and one page devoted to company history. Some material that had been on the intranet may have been migrated to the content management system, but no process was identified for this in either naming convention or documented policy. Further, none of the documents retrieved from the content management system were html based documents leaving the assumption that most intranet web pages are lost to later researchers after they have served their usefulness on the intranet.

The digital archives then present an interesting contrast of very current, but shallow information found on the intranet web pages and the great depth of information found in the content management system that is likely to be unorganized. In this research, the information found on the content management system indicated there was to have been an active section on the intranet to support the change program. However, none of this material currently exists on the intranet. Yet, without having found evidence of the communication plan it is unlikely that the researcher may have known of such a plan in the first place. In this case, the significance of using a digital archive is that while some material may be lost, there is also the likelihood that other evidence of its existence may be found elsewhere. Similarly, while numerous draft copies of some documents creates some confusion and a bother to sort through, it also presents an interesting insight into the past process of how a particular employee communication message evolved from initial draft to final copy. And, while lack of management created a significant accumulation of documents in the content management system, it also provides the researcher more material for understanding the past than might have remained if the content were better managed and some material discarded.

One last comment on the adequacy of the digital archive to serve research purposes concerns the issue of not knowing how thorough the archived material might be. This involves the problem of what is kept for long-term retention and what is destroyed, not to be available to any researcher in the future. While paper-based documents can be destroyed and lost, electronic documents could easily be lost without efforts to retain them, but this brings up the issue of the criteria that should be used for retention. In some cases, seeing the evolution of a particular document provides insight into the development process itself. However, it also adds to the number of documents that must be reviewed, especially when two documents are the same, but were duplicated in the filing process either through different names or different filing locations. Corporate intranets and content management systems provide a great window into the past, but it is not always known to what extent the picture seen is the full history. Critical documents might have been labeled confidential and filed in an inaccessible location or not filed at all other than on a personal computer. Here oftentimes too much was kept adding to the difficulty of not knowing which document was the final copy and what the sequence was of the numerous intermediate documents.

But this research situation is not unlike basing research on face-to-face interviews and not interviewing the right people with sufficient knowledge on the topic being researched. In both cases the researcher has to follow leads to other information (or people) and eventually draw a conclusion that sufficient information has been uncovered. In the case studied here, the lack of management of stored content provided great depth of material for study. But, it also revealed a few holes, most notably the lack of a post audit of the transformation program.

One last comment on cyber-archaeology concerns the impact such research may have on the content repository itself. While this research proceeded unimpeded by organizational constraints, the mere revelation of the case study results might limit similar studies in the future. Content management systems may be cleaned up by deleting documents not currently needed by the organization. Additional security controls may be put in place limiting access to material now viewable. In both cases, the result would be reduced material for performing a similar study of the current culture that is being captured on the intranet and within the content management system.

It may also initiate actions to retain electronic artifacts for study by future generations though that is highly unlikely. Where such digital archiving programs are in place, they are being driven by academic researchers and not the organizations involved in creating the documents. But instigating such policy changes was not the intent of this study, which focused on documenting past change initiatives as they are currently recorded within the electronic archives. Yet, once knowledge about the breadth of the archive is known among information security personnel and content creators there is the real possibility of them destroying the parts they deem inconsistent with current information needs. In this manner, a study of this type may not only document the past culture hidden within the intranet and content management system, but also bring about its destruction through its identification.

REFERENCE LIST

Hodge, G. M. (2000). Best practices for digital archiving: An information life cycle approach. *D-Lib Magazine [Online], 6*, 1. URL: http://ww.dlib.org/dlib/- january00/01hodge.html [2003, August, 22].

Jones, Q. (1997). Virtual-communities, virtual settlements & Cyber-archaeology: A theoretical outline. *Journal of Computer-Mediated Communication [Online], 3,* 3. URL: http://www.ascusc.org/jcmc/vol3/issue3/jones.html [2003, September 9].

Lyman, P., & Kahle, B. (1998). Archiving digital cultural artifacts: Organizing an agenda for action. *D-Lib Magazine [Online]*. (38 paragraphs). URL: http://www.dlib.org/dlib/-july98/07lyman.html [2003, August, 22].