Chris Krause

Digital Curation Journal

Background:

I have a habit of posting all essays I have written worth reading (i.e. ones I actually enjoyed writing) on my website/blog krauselabs.net. Yet I have had issues with my webhost in the past, including one event which even included a possible loss of all data on the server. A fire resulted in the original data being destroyed, and there was a period in which the server administrators were unsure if the backups, located offsite, were intact. The latter site, by some freak occurrence of outrageous fortune, had also simultaneously suffered a systems failure. Luckily, my data, the corpus of my life's work, was not lost. It was a shocking wakeup call: storing the data directly on my website was the sole means by which this data was stored, and all of my writings and creative endeavors were nearly erased without recourse. When my website's data was finally restored I immediately blitzed the backend and saved offline copies of the SQL databases which contained my work.

While it is a simple matter to backup SQL databases, it is not a simple matter to migrate text and other content from SQL to readable documents. Furthermore, SQL database files tend to be massive in both the number of lines and in file size. A few corrupt sectors on a disk could result in a corrupt file. This is to say nothing of the software dependencies that come along with having a Wordpress blog. Accordingly my digital curation project was aimed at migrating these “born digital” files to a more easily accessible format, creating another redundant location for backup and for ensuring archival permanence.

A note on the data:

Many of the essays which are included in the digital curation project were originally created in Microsoft Word and then copy and pasted into Wordpress, via my web browser Firefox. For purposes of this project I consider the provenance of the records to begin as hypertext entries on my website. My custom in the past has been to create a record locally using Word and then to migrate it, more or less immediately, to hypertext via my blog. The pre-history of the assignment involved a normalization of these content to the respective file formats (txt or pdf) included in the preserved and bitstream files. In other words these files did not exist until I created them; they were not laying around on my hard disk. I will discuss why I decided to normalize the website hypertext to what I consider to be archival-grade digital file formats in an upcoming section.

Intent:

Permanently preserve personal essays of an academic nature.

Record selection criteria:

Records were more or less selected in a semi-random fashion. The first ten essays that I encountered upon reviewing my website were selected, although I did make an effort to vary the content and scope of each record. I also made a conscious effort to preserve my undergraduate thesis, which I consider to be a work of great personal importance.

Media:

I selected TXT and PDF formats for purposes of archival preservation. If possible within the parameters of the assignment then TXT would have been exclusively selected, whereas:

1. TXT documents are a fundamental format in personal computing, accessible by both Macs and PC, using basic operating system programs. There is little threat of the file becoming “obsolete” or otherwise becoming inaccessible due to changes in software and hardware.
2. TXT documents rarely degrade, and in the event of corruption, are more easily restored, both partially and completely, when compared to the alternatives; a few corrupt bytes in the header will not compromise the entire file. This is in contrast to Word documents and JPEGs, which are notorious for becoming corrupt over time, and rarely can be restored in completion.
3. Universal presentation, even in regard to newer text encoding methods. While the text documents I used for the project were ANSI encoded, non-ANSI text readers (from antiquated operating systems) can still interpret and open the files. It is wise to choose a format which is backward and forward compatible with software and hardware platforms.
4. Small file size; Word documents with comparable word counts are often tens of times larger in size. Small file sizes ensure the viability of a continually expanding archive.
5. Extremely flexible for migratory and normalization purposes, non-formatted raw text can be easily copied into a myriad of programs and then be converted into alternative

formats. The file can be appended, modified and viewed with ease and without special software.

This being said, text documents do have some weaknesses: elaborate text formatting is not possible, and integrated content (such as pictures and graphs) is impossible. For all intents and purposes I would urge the use of text documents for archival preservation, but when the record simply will not react well to the rules of the format, PDF is an acceptable alternative. PDF is capable of storing the advanced formatting of Word or Excel documents but frees the record of the dependencies normally associated with such latter formats. Word documents using a non-standard font must be accessed on a computer which has that font installed, otherwise style is compromised and the record presentation lacks similitude.

PDF files are not prone to this flaw, and present the record as it was originally, more or less creating a “photocopy” of it. PDF is also universal, and while although a relatively new format in the chronological history of computing, can still be accessed by Mac and IBM-compatible PC alike. Unfortunately the client must have access to PDF reader software (Adobe’s readers are offered freely available online), and the file is useless without it. For this reason I was wary to adopt PDF as an archival format, as my schema involved permanent storage and access. Ultimately I chose the format because at least two formats were required as per the assignment instructions and PDF was the soundest medium aside from raw text. One must look hundreds of years ahead.

Records:

10 personal essays of an academic nature. Records were normalized from hypertext or Word to text and portable document formats via Microsoft Notepad and Bullzip PDF Printer. The end result was 5 text documents and 5 PDF documents.

2 essays were also migrated from their original format to the respective other. April112008Essay.txt was migrated to PDF format, while April272009Essay.pdf was migrated to TXT. In the first instance the continuous text document was converted into a two page PDF document using Bullzip PDF Printer. Notable changes from the original included the addition of a header and footer not present in the original: the document reads “April112008Essay.txt” at the centered header of each of the two pages and includes a centered page number footer. April272009Essay.pdf’s migration process, performed using Adobe Reader’s “save as text” function, concluded with a document which was quite different from the original:

1. The margins and spaces of the original document are truncated: the original contained a dedicated title page with large spaces between title and class information, the new txt document has only a single line break between each string of text.
2. The word Krause followed by a number precedes each “page” of text. A line up from “Krause” is the Unicode character FF.
3. The bibliography’s formatting has collapsed, as was to be expected. Margins do not exist.
4. The text block concludes with a FF Unicode character.
5. Footnotes are not formatted and instead appear as undistinguished numerals.

The aforementioned document is quite readable, although quite different than the original. I believe this would create issue in an archival institution but does not affront me, as I am happy to just have the information, rather than the original essence of the record, preserved.

Preservation Plan:

1. Near-line local storage on my personal computer
    a. Preservation operations as per the assignment instructions, including weekly integrity trials and backup.
    b. Deemed unsatisfactory for permanent storage due to the threat of localized catastrophe.
2. Online storage at an enterprise-grade webhosting facility 1&1 Internet
    a. The files are stored on RAID arrays so as to avert localized catastrophic failure of a single component or hard drive within the server.
    b. I transferred the files from (1) to the server by FTP.
    c. These files are automatically backed up to offsite redundant mass storage facilities (3) on a weekly basis.
    d. Physical plant located at: 701 Lee Road, Suite 300, Chesterbrook, PA 19087, Fax: 610-560-1501
    e. Archive accessible at: http://www.krauselabs.net/archive/
        i. The mechanism exists for this archive to be protected by layers of network security, including encoded passwords and IP verification.
3. Redundant offsite, near-line storage at an enterprise-grade webhosting facility 1&1 AG

a. Physical plant located at: Elgendorfer Straße 57, 56410 Montabaur, Germany - 02602 96-0.

b. Three distant, redundant storage locations ensure the viability of archival permanence.

Process and Reflections:

I did not run into the sort of technical hurdles spoke of in the sample journal or the discussion forum posts. No mismatched bytes, no missing files, no corruptions. I think a lot of this has to do with the fact that I had a clear vision and preservation philosophy already in mind, and am quite familiar with semi-permanent storage and computing. A career data hoarder, I have encountered the problems of faulty hard disks, proprietary, obsolete file formats and precious gems lost in translation. Over time I have come to trust simple formats, robust hardware and the virtue of redundancy. As our late 19th century brethren might say: copying our records, to many redundant locations, ensures survival! I consciously avoided problematic technologies and eagerly adopted technologies which I have effectively ministered in the past, buttressing my understanding of computer science with newfound library and preservation science wisdom. While my previous archival escapades tended to be more chaotic and unstructured, I see now the value of deliberate metadata and the usefulness these documents can serve in ensuring permanence and responsible information storage. While I was initially quite hostile to OAIS and other highly structured models of digital preservation, I can now see the value of standardized metadata sheets. That being said I am still suspicious of standardized data access and retrieval schemas in regard to system design, as I believe that notion clashes fundamentally with the pillars of Web 2.0 internet. That discussion is outside the scope of this assignment however.

As far as the future is concerned I do not foresee myself dedicating the time on a weekly basis to checking bytes. In fact, the system I have engineered does not require human intervention at all. The offline archive on my personal computer is a safety net in the nearly impossible event of another double catastrophe at my web server. This is to say nothing of the effort needed to manage an archive of this sort on a larger scale – a truly time consuming deed. I would say that it is important to have a redundant storage site which is separate from your online storage, but that it is not critical to check bytes on a weekly basis. Purchasing a robust system of RAID hard disks nearly makes moot the task of byte checking. This manual labor only becomes critical for the most important records. I think the system I have in place for my everyday life is extremely easy and reliable: if I want to preserve a file I just drop it on my FTP or post it on my site using Wordpress. I pay a monthly fee for Enterprise-grade industry technicians to oversee the data servers and ensure that my data is redundantly stored. My records are doubly made safe by my local archive. All is well, and little anxiety is on my mind.