11-1-2014

# The Missing Memory of Jehu: An Essay on the Preservation of Data through Time

Ted A. Campbell
*Perkins School of Theology*, tedc@smu.edu

# The Missing Memory of Jehu:
# An Essay on the Preservation of Data through Time

by Ted A. Campbell

ויתר דברי יהוא וכל־אשר עשה וכל־גבורתו
הלוא־הם כתובים על־ספר דברי הימים למלכי ישראל׃

*Now the rest of the acts of Jehu,*
*and all that he did, and all his might,*
*are they not written in the book of the chronicles*
*of the kings of Israel?* (II Kings 10:34, KJV)

If you want to know more about King Jehu, says the book of Kings, you can look it up in "the book of the chronicles of the kings of Israel." Unfortunately, though, we can't. There is a book in the Bible called Chronicles (divided into two in Christian Bibles), but it isn't "the book of the chronicles of the kings of Israel" mentioned in this passage. Some information about King Jehu was preserved, like the Hebrew text above. But "the rest of the acts of Jehu" recorded in "the book of the chronicles of the kings of Israel" has perished. Alas, the missing memory of Jehu.

Historians lament the loss of such information as we continue to obsess over the destruction of the great library of Alexandria. But information is perishing all the time. The problem we face, as Isaac Asimov once explained it, is that "you can only go back so far, and the farther you go back, the less reliable the information becomes—no matter what you do." He thought of this as "an uncertainty principle in information" (Asimov 1988, 272).

Asimov put this in a novel, and we don't have to take it as a given (a datum) that all data will atrophy forever. But how do we preserve texts and other data from such a fate as that of Jehu's chronicle? How do we preserve memories, stories, images, texts, or video and audio recordings, across multiple genera-

tions? How do we preserve business records in the long run? How do we preserve the memories of our families?

### Preserving Data

We might begin by considering how data have been preserved historically and prehistorically. Three principal means for transmitting data prior to the advent of digital computers were oral transmission, transmission via manuscript writing, and transmission via printed writing.

Ancient people transmitted stories and more complex texts by oral repetition or recitation. There were some controls over the oral transmission of information. For example:

- orally transmitted texts were recited in the presence of witnesses who could verify the accuracy of the texts;
- orally transmitted texts were often introduced and concluded with a solemn announcement, often in the form of an oath, of the importance of the text and that the reciter vowed that the text he or she gave out was true;
- orally transmitted texts were sometimes accompanied by curses pronounced on those who altered them and blessings pronounced on those who transmitted them accurately.

Each of these controls on the transmission of oral texts can be seen in Jewish and Christian scriptures, embedded within texts (such as the letters of Paul) that were later written down (cf. Campbell 2009, 15-17). But although oral texts might be transmitted through centuries with some degree of accuracy, oral transmission was always hampered by the possibility of inaccurate remembering, speaking, or hearing of texts, leading consistently to the degradation and corruption of older forms of texts.

The development of writing greatly enhanced the ability to transmit information through time and enhanced the accuracy of information that was transmitted. Subsequent developments of writing and printing improved the possibility of accurately transmitting information:

- the development of more reliable media: clay or stone tablets, parchment, papyrus, vellum, and various forms of paper (cloth-based paper, wood-pulp-based paper, which was a step backwards with respect to preservation, and newer forms of acid-free paper);

- the development of tools for writing on these media: styli, pens, pencils, block printing, printing presses with moveable type, and newer forms of industrial printing, for example, rotary presses;

- the development of particular forms of writing that enabled more accurate transmission, for example, the move from majuscule text (all-upper-case writing, typically with no spaces between words and no punctuation marks) to more elaborate miniscule text utilizing distinct upper-case and lower-case letters, spaces demarcating words, and punctuation and accent marks; and

- the development of processes for the accurate written transmission of information such as the Benedictine use of scriptoria in monasteries, or the more elaborate evolution of printing processes involving composition and typesetting, the production and proving (proofing) of galley sheets, and the correction of these sheets before final printing.

The transition from oral to written means of transmission took centuries. Older written texts such as the Jewish and Christian scriptures transmitted originally by manuscript contain oral texts embedded within them, and the process of copying in monastic scriptoria involved the oral reading of a text while monastic scribes wrote down what they heard, thus opening the possibility of degradation of texts over time even when they were written. But reading written texts aloud as they were copied allowed for more integrity than simply reciting texts from memory.

The nineteenth century saw the beginnings of new forms of information storage that could transmit images, audio and video recordings. The advent of digital computers from the period following the Second World War has brought the possibility of new ways to transmit information, though many of the older problems persist and in fact transitions from written and other media formats to computerized data are really still in their infancy.

Computers do three things: they manipulate data, they store data, and they link data. The first item, the manipulation of data, was the initial impetus for developing digital computers, emerging from the work of code-breaking during the Second World War.

Means of data storage developed through the 1960s and beyond, some derived from earlier forms of recorded media: punched cards, magnetic tape, magnetic disk drives, optical disc drives, and various forms of persistent RAM (random-access memory) including solid-state drives (SSDs). "Cloud storage" may sound like a new way to store data, but it's really just a new way to get at data stored on someone else's server on hard-disc drives or SSDs. A variety of data formats have also been created which can store texts in various languages, and can store images, audio and video recordings, and other forms of data (e.g., databases and spreadsheets) as well.

Despite early hype about computerized storage and especially optical media – Microsoft famously branded CD-ROMs as "The New Papyrus" in the mid-1980s – they and other computer media have proven to be impermanent and have raised the critical issue of competing and mutually incomprehensible data formats in which information is coded and stored. On the other hand, one of the very positive outcomes of digital computer technologies for the transmission of information has been the advent of computer networks and eventually the internet in the 1980s that have brought the critical matter of data linkage (embedded links between data) to the forefront and have enabled new and creative means of linking data together.

### The Law and the Commandments for Preserving Data

Although particular media and techniques for transmitting information have changed vastly, there are some basic principles that govern the transmission of data through time, no matter the media in which data are encoded. The primordial law governing the transmission and preservation of data through time, something like a "Zeroth Rule" for data preservation, can be stated thusly:

**Remember, O mortals, that thy data are preserved**
**in the short term by storage on media**
**and in the long term by copying.**

The most important part of this statement is the last part: no media have proven to be permanent; all media will eventually atrophy or decompose and the information stored on them will be lost. Only the copying and recopying of information will preserve it through long stretches of time. This applies to written texts and computerized data as it did to orally transmitted stories. Information left only on media and not consistently recopied or retold is bound to atrophy. Copying data such as texts frequently and in volume counteracts the tendency towards data atrophy. It also helps protect against censorship. The sheer volume of books printed by Wycliffe and Tyndale, to take two examples of censored works, ensured the survival of their texts against attempts to eradicate them.

The moral here is: copy and recopy. Accurately. I will say more below on principles governing the copying of data, but the point here is that apart from copying, data are bound to disappear. I have all of my own computer files including all the books and articles I have written, all of the genealogical data I have compiled, and I also have some of my family's computer files including my deceased parents' files including their own digital images, stored email, and PDF files that my father made of a lengthy book on family history that his brother compiled. I have them all now, triply backed up, but the media on which they are stored will not persist. Only copying on the part of new generations will ensure their transmission into the future. Someone failed to do that for "the book of the chronicles of the kings of Israel," which just might have been more important than my family's files.

This most general principle warrants more specific rules for preserving data across generations. The first specific rule applies to the process of developing information and it recognizes one of the most consistent problems in this enterprise:

## I.
## Thou shalt identify and isolate only one editable copy of data.

More than one editable copy of a document is a road to disaster. Critical editions of ancient and medieval texts often have a "stemma" showing branches of the alterations or corruptions of a document; the branches often occurred when more than one scribe or copyist was working on the same work.

This principle is doubly important when working with computer media. Producers of database software realized early on the problems inherent when two or more people were working on a single database concurrently. Two people would begin editing a database at the same time; the first editor saved the database and over-wrote the original version; then the second editor saved the database and over-wrote the version saved by the first editor. Only the edits of the second editor were left and the edits from the first editor were lost. The same scenario can occur when two people are editing a word-processor document simultaneously and only the edits of one person (the last editor) are saved.

Some contemporary software (Google Docs, Microsoft Office) has built-in facilities for collaborative work. But if you're not using software like that and observing its own protocols for collaboration, the only solution is to develop a fail-safe system by which only one person at a time can edit a document and save its contents. Other copies (backups) can be saved, but they should not be used for editing, and should be understood as, and perhaps marked as, "read-only" copies. More elaborate data structures such as large-scale databases might need to employ a more complex system where only one person can edit an isolated or unlocked portion of the data and write it to the database at one time. In this case, the editable portion is "checked out" and then "returned" to the overall project, using the analogy of checking out and returning library materials. But in either case, the editable data has to be identified and isolated (restricted to one editor) to avoid lost edits.

This raises a related issue that I'll list as a separate principle:

**II.**
**Thou shalt distribute information that is not to be edited in read-only formats.**

This may seem dead obvious, but it's quite customary at this time to distribute documents as editable text, especially in the form of Microsoft Word documents. Sooner or later this is going to backfire badly. I fear that students will produce their own creatively edited version of a course syllabus ("The instructor is required to provide drinks and refreshments for each class session") or members of a corporate board are going to offer their own version of minutes from a critical meeting according to which (surprise!) they won a crucial vote. Or again, an editable document will be edited simultaneously by two or more persons, invoking the scenario envisioned in principle I above.

I am now in the habit of distributing almost all materials in a read-only format such as Adobe Acrobat's PDF format. The only exception I make is for a document that I'm offering to be edited by others, e.g., lecture outlines that individual students can use as a basis for their own lecture notes. One of the issues, though, is that Adobe Acrobat and other read-only formats are in reality only partially read-only; readers can still cut and paste from PDF documents. They may lose the format and it makes it more laborious to copy large amounts of data, but determined readers could copy most of a document.

The jury is still out on the question of which data formats should be used for long-term digital storage. The best advice I can offer at this point is this:

**III.**
**Thou shalt store data for long-term storage in multiple formats, in the most widely accessible data formats possible, and thou shalt recopy existing data to newer formats.**

Several years ago the use of something called SGML (Standard Generalized Markup Language) was advocated as a universal format for texts (as in the *Chicago Guide to Preparing Electronic Manuscripts*, 1987), and although that still makes sense to me, SGML simply hasn't caught on. HTML (HyperText Markup Language, an extension or application of SGML used in web sites) and such formats

as Microsoft Word and PDF are widely used. A great deal of textual material is available in simple text (ANSI/ASCII text format). A similar situation applies to formats for images, and for audio and video recordings. I think it's safest to keep data in multiple formats that are (at least at present) widely accessible, keep copying, and copy to new formats (in addition to older ones) when new formats become popular.

## IV.
## Thou shalt keep thy data close unto thee.

Keep your friends close, keep your enemies closer, and keep your data, in the immortal words sung by Karen Carpenter, very, very "close to you." If I might state this in slightly more expanded King-James language: "Keep thy data on thine own storage media, and do not thou entrust all of thy data unto The Cloud." It's not that I'm completely against The Cloud. Cloud-based storage like Dropbox is wonderfully helpful, for example, when you're traveling and need to have a certain body of data constantly available. Of course this only works as long as you have an internet connection, which is not always.

One advantage of cloud-based storage is that it can automate the process of storing backups of data, freeing users from some of the laborious work of synchronizing and backing up their files. But I don't like keeping everything in The Cloud because the data are too exposed. If it's not the NSA, it's going to be some corporate entity that's interested in my personal habits and inclinations and proclivities as revealed in my data. And yes, dear, our data reveal those things. For some reason I don't feel like sharing that. And it's not just that my data may be exposed when they're sitting on someone else's server, they're also vulnerable when they travel up and down the tubes to and from The Cloud. Very vulnerable, I'd say. I'm keeping most of my data close at hand. But that means that the responsibility falls on me to keep it sync'd and backed up. Which leadeth unto…

## V.
## Thou shalt synchronize and backup thy data methodically.

Data backups and synchronization are the methods by which we most frequently engage in the crucial enterprise of copying by which data can be preserved for longer terms. But there are some rather strict rules that govern the synchronization and backing-up of data.

A data *backup* means that data are copied literally from one storage location to another, e.g., you copy the whole contents of your hard-disc drive to another (backup) hard-disc drive. *Synchronization* means that data from different storage locations (two or more) are synchronized so that all the media contain all the appropriate data. Synchronization is far more complicated than backup, because a host of issues have to be resolved: e.g., should deleted files and/or directories be deleted or restored? For conflicting versions of the same files, should multiple versions be saved (and under what names?), or should the newest version be saved, over-writing older versions? Synchronization is more forgiving than backup, where files can perish eternally, so synchronizations need to take place before a backup of the data takes place.

But synchronization can be very tricky. If synchronization software allows only a two-way sync for data stored in more than two storage locations, then a complicated round-dance of sync has to occur. Suppose, for example, you have storage locations (perhaps hard-disc drives) A, B, and C. You sync A with B, and then you sync A with C. But if C had data that was not on A or B earlier, it's still not on B at this point, so the full dance for three storage locations has to be:

A with B,

A with C, and then

A with B again.

That gets everything on A, B, and C synchronized, as long as none of them changed in the interim. Fat chance. But of course this gets even more complicated

with four or more media storage locations, in which case you should Seek Professional Help.

I'm not going there. As mentioned above, Cloud-based storage (storage on external servers accessible through the internet) can automate the processes of synchronization and backup, but it can also involve the exposure describe above. But one way or another, you need to synchronize and back up your data methodically.

## VI.
## Thou shalt cultivate accurate metadata.

"Metadata," in the case of printed books at least, means all that fine-print stuff on the back of the title page that indicates which author or authors claim responsibility for the work, who claims copyright for the work, what press assumes responsibility for the work as it is made public (published), when the work was first made public (published), and when the work was last edited or revised. "Metadata" in other cases would indicate similar information appropriate to the media: who is responsible for and claims copyright for photographs or other images or for audio or video recordings, when the media were made available, etc.

Metadata can denote more informal information accompanying data. My mother regularly marked on the back of photographs and on the margins of 35mm slides indicating the subjects of the pictures, often identifying who appeared in the photographs. Now writing on the back of photos is not uniformly recommended: her writing bled over onto the front of the next photo in many cases. But her notes are an excellent source of meta-information on the photos. Similarly, I once videotaped some of our family's 8mm silent home movies from the 1950s and 1960s. Although the recording itself was dreadful, the recorded conversation with family members as the movies were playing provided a rich source of information on the content of the movies. This too should count as metadata.

Metadata is utterly crucial information apart from which data can scarcely be searched or understood, but the internet is rife with poorly documented data, that is, data that lack the basic metadata or meta-information that appears in almost every published book or every publicly distributed audio or video recording. So students in my Christian history courses regularly show up with references in their papers to *The Catholic Encyclopedia*. It's a wonderful source for learning what Catholics thought in the 1930s about a variety of topics, most interestingly, about Protestants. But the free version of this work on the internet fails to indicate that the information is 70 or 80 years old and that there is in fact a *New Catholic Encyclopedia* that has significantly updated the information from the old one. But I'm not just kicking *The* old *Catholic Encyclopedia*. The truth is that very few sites on the web ever give the kind of metadata we always expect in published works, even in web sites for academic institutions including libraries that ought to be deeply committed to the principle of providing metadata for published works. Web sites are indeed "published" in the sense that they have been made public and they need rich and accurate metadata like any other published work.

Finally, and here acknowledging one of the great the strengths of the web, I offer this principle or admonition:

**VII.**
**Thou shalt cultivate rich, accurate, and persistent data linkages.**

If you think about it, "… are they not written in the book of the chronicles of the kings of Israel?" involved a kind of data linkage: the text refers readers to another text for further information. It wasn't as easy as clicking on a hyperlink, and now you can't just go check out "the book of the chronicles of the kings of Israel." But it did involve data linkage. The same thing could be said for footnotes or other forms of reference in traditionally printed articles and books; they allowed readers, at the cost of some time spent digging around in a library, the opportunity to check out claims or quotations or to verify an author's sources. At another level, library catalogs provide critical linkages between information; even

the ways in which books are arranged in a library imply linkages between the data they contain.

The internet, and especially the web, created by some of the staff at CERN many years before they nabbed the Higgs boson, has offered unparalleled opportunities for data linkage. Complex linkages between data open new possibilities for ways in which computers contribute to the development of human knowledge. If computers are thought of on the analogy of human brains—and it is strictly an analogy—then data linkages are the "synapses" that connect cybernetic "neurons" (data) together. It is in this particular area that the computer may be the *novum organum* of our age: not in an alchemical ability of computers to "think" as women and men think (Penrose 1989). The establishment of complex linkages between data is more than a means of *presenting* data; I would say that it involves the creation of new forms of knowledge.

If that's an exciting prospect, a few hours' research on the web shows how inadequate many of the linkages are. Many hyperlinks on the web are broken, point to non-existent web pages, and internet specialists don't seem to have developed standards for persistent linkage between data. The linkages are sometimes inaccurate, that is, they point to the wrong data or information, and I'm led to wonder if there should be something like a *Good Housekeeping* seal of approval for web sites that maintain consistently accurate and persistent linkages in addition to helpful metadata (see the previous principle). And the linkages can be richer. I can envision a time when, inspired by the possibility of accurate linkages that persist over time, we will celebrate those unique individuals gifted with the charism of linking data, an Academy Awards to honor their unique genius.

## Conclusion

Meanwhile, information is perishing. Some of it of course is crap and can be left to the scrap heap. Some of it is like the memory of Jehu. Sometimes we really do not know what looks like crap to us but will be valued by our progeny.

Books printed on pulp-based paper in the 1800s are turning brown and flaking, many of them have already perished. Forever. Librarians feel that they are being asked, with very limited resources, to play the role of gods and decide which of these works from the past will be preserved for the future. Family records, photographs, perhaps recordings, are thrown out with the garbage after the elders die. Despite our technologies, libraries continue to burn, or at least, the information that might be preserved in them continues to perish.

So here's my best shot at how to avoid the fate of Jehu today:

**REMEMBER, O MORTALS, THAT THY DATA ARE PRESERVED**
**IN THE SHORT TERM BY STORAGE ON MEDIA**
**AND IN THE LONG TERM BY COPYING.**

**I.**
**THOU SHALT IDENTIFY AND ISOLATE**
**ONLY ONE EDITABLE COPY OF DATA.**

**II.**
**THOU SHALT DISTRIBUTE INFORMATION THAT IS NOT TO BE EDITED**
**IN READ-ONLY FORMATS.**

**III.**
**THOU SHALT STORE DATA FOR LONG-TERM STORAGE**
**IN MULTIPLE FORMATS,**
**IN THE MOST WIDELY ACCESSIBLE DATA FORMATS POSSIBLE,**
**AND THOU SHALT RECOPY EXISTING DATA TO NEWER FORMATS.**

**IV.**
**THOU SHALT KEEP THY DATA CLOSE UNTO THEE.**

**V.**
**THOU SHALT SYNCHRONIZE**
**AND BACKUP THY DATA METHODICALLY.**

**VI.**
**THOU SHALT CULTIVATE ACCURATE METADATA.**

**VII.**
**THOU SHALT CULTIVATE RICH, ACCURATE,**
**AND PERSISTENT DATA LINKAGES.**

Keep copying, and blessed may you be in so doing.

## References


Asimov 1988

Isaac Asimov, *Forward the Foundation*. New York: Bantam Books.

Campbell 2009

Ted A. Campbell. *The Gospel in Christian Traditions*. New York: Oxford University Press.

*Chicago Guide to Preparing Electronic Manuscripts* 1987

*The Chicago Guide to Preparing Electronic Manuscripts: For Authors and Publishers*. Chicago, University of Chicago Press.

Penrose 1989

Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford and New York: Oxford University Press.


## Version, Copyright, and License

Version 1.0, 1 November 2014