Archival Issues Related to Digital Creations

Mark Kieler
Carnegie Mellon University, USA

Michael J. West
Carnegie Mellon University, USA

INTRODUCTION

The authors define "intellectual creations" as human expressions embodied in text, music, or other forms of art. Increasingly, we encode these creations in digital formats that have extremely short life cycles. Eventually, backward compatibility is lost. Thus, after very little time, a digital encoding format becomes obsolete, and intellectual works encoded in the format may become irretrievable. In contrast, the cultural worth of an intellectual creation may not be realized for generations. Additionally, future generations must access artifacts, including intellectual creations, to understand a culture in historical context. We contend that technology – intensive storage and manipulation of data may result in an inability to gain this access. Technology creators have some responsibility to facilitate future retrieval through careful documentation, and by selective maintenance of hardware that may be required to access archival media.

BACKGROUND

Cultural artifacts nearly always outlive the technologies that made them possible, which is particularly obvious with digital technology. Imagine the discovery hundreds of years from now of a floppy diskette containing a document written in WordStar®. Once a code has been lost, have all texts written in that code been lost as well?

At some point, supporting a technology is no longer economically viable if it is abandoned by enough people. It ceases to exist except in archival form. Currently, home videocassettes are being replaced by the DVD (Digital Video Disc), which subsequently will also be superceded.

Linguists stress the organic nature of language, which it is in constant evolution. Grammars are codifications constructed of "snapshot images" of spoken and written language usage. Despite the fact that both are considered examples of English, speakers of the vernacular of Shakespeare and the vernacular of rap music would find each other incomprehensible. With technology, as with language and culture, timing is everything, and context plays an important role in shaping how we interact with each other and with technology.

HOW DO WE KNOW ABOUT PAST CULTURES?

In engineering, Shannon (1948) described elements of a communications system: An information source generate data. A transmitter encodes information to travel over a channel through distance and time. At the other end, a receiver decodes the signal for the destination, which is the person or thing for which the message is intended. Since we are worried about humans, we consider them to be the source and destination. While parts of the transmission system have changed greatly over time, one could certainly argue that technology has caused the other portions of the system to change more rapidly than the human elements. Linguists, beginning with Ferdinand de Saussure, established semiotics as a science of signs, which likewise focuses on the sender and the receiver of a message. Semiotics also posited the arbitrary nature of the sign and looked at the ways human languages encode and transmit messages. (Saussure, 1974).

Physical artifacts that survive from the distant past reveal much about a culture, depending on their purpose and the quality of materials from which they are made. Our record from ancient civilizations is far from perfect, but archaeologists can construct some details about them from these clues.

Significant events have often been recorded in the living embodiment of a storyteller, and oral traditions still form an important part of many cultures. The historical record has often been related by language, as well as by performance (e.g., a ritual dance). Some of these oral histories survived long enough to be recorded in other more permanent media. However, not only have many oral traditions died with the last generation of storyteller, others have assumed inaccuracies and exaggerations as a result of being passed serially through generations.

As languages evolved and became standardized, it became possible to encode events in written form. Because writing has traditionally been the province of the learned few, written documents were recorded on long-lived media, and special care was accorded to their storage. Fortunately, many ancient documents have survived, albeit, with significant degradation. Given the constant change in a living language, when a culture dies, the language often dies with it. Language experts

attempt to reconstruct meaning by looking for patterns that may establish types of words and contexts and similarities to more modern languages.

In addition to printed text, human expression is also accomplished through artistic means such as music, painting, and dance. While we only have been able to preserve musical and dance performances for a relatively recent portion of human history, for centuries we have had a written procedure for recreating music, using notes, measures, and time definitions. The methods for recording dance instructions are much less standardized and rely on interpretations and tradition.

Art works degrade over time, depending on the types of inks, paints, or dyes used, the media on which they are deposited, and overall interaction with the environment. Even sculpture is subject to environmental degradation, such as damage from acid rain.

THE INTRODUCTION OF TECHNOLOGY

The invention of the printing press made wide distribution of printed information possible, and wood pulp-based paper made it affordable for the general public. However, unlike expensive fiber-based paper, pulp-based paper has usually been manufactured through an acid-based process, and residual acid in the paper eventually destroys it. Thus, paradoxically, fiber-based books from the 19th century are often more legible than their 20th century wood pulp counterparts.

Text

Text was the first means of expression to be converted into electrical form. In fact, text went "direct to digital." Morse code is a duration-encoded digital signal, unrecognizable to anyone who does not understand it. Initially, storage was primitive, used mainly to "buffer" the information until a human could reconvert it to text. Thus, long-term storage and archiving of Morse code traffic was not an issue.

The first modern bit encoding of text occurred in 1874. Emile Baudot, a French telegraph engineer, devised a 5-bit code for each letter of the alphabet. Unlike Morse code, each symbol had a fixed length representation, dependent only on the presence or absence of electrical current. The Baudot code was durable, used by news service teletypewriters throughout the 1970s (Freed, 1995). Crude paper tape punches were often utilized for storage. The digital code could be read, albeit slowly, merely by holding the tape up to a light.

ASCII (American Standard Code for Information Interchange) uses 7 bits. The added bits allowed upper and lower case letters, as well as numbers, punctuation, and other special characters. It endures as the "plain text" standard.

The rise of WYSIWYG (what you see is what you get) computer interfaces, and the availability of sophisticated word processing programs, made it possible to digitally encode additional expressions to text. Different art styles of text (fonts) could be used, and these could embody visual variations such as italics, bold, superscripts and subscripts, and underlines. Word processing evolved to encode these variations. This was accomplished by adding bits to the original text data, or by software commands that controlled a section of ASCII text. These techniques represent a deviation from international standards into conventions of the word processing software. As the level of sophistication increases, we become increasingly unable to understand an encoded section of text without the software used to create it. In fact, the actual text becomes only a tiny portion of the code. If the page is graphically encoded, as occurs in programs such as Adobe Acrobat®, then plain text representations are lost all together (Kieler et al., 2004).

Audio and Visual Technology

Storage of audio and visual data began in the 19th century and progressed in sophistication throughout most of the 20th century. While not strictly correct in each instance, the general paths of progress could be viewed as follows:

- Visual Still and Audio-Only, to Visual Motion, to Visual Motion with Audio
- Mechanical or Chemical-Based Recording, and Storage to Electronic Recording and Storage
- Mechanical Reproduction to Electronic Reproduction
- Analog Encoding to Digital Encoding

The most significant impediments to easy retrieval of an intellectual creation involve electronic encoding and storage, and digital conversion.

Electronic encoding of visual information marks the historical point where the original signal cannot be recovered merely by converting stored data directly into light or motion. The visual image is stored as a series of lines, and voltage references signal new lines or new collections of lines (*frames*). Thus, one must be able to properly decode the various electronic levels and pulses unique to the method, and understand how the lines are ordered and encoded in order to correctly reconstruct the scanned image. Incompatible formats include PAL (Phase Alternating Line), NTSC (National Television Standards Committee) and SECAM (Systéme Electronique Couleur Avec Mémorie) (Abbott, 1994).

With the advent of magnetic storage, one could no longer directly "see" the information on the storage medium without a great deal of technological aid. As an example, by the late 1920s, light modulation of audio onto motion picture film was possible (Hochheiser, 1992), and both audio and video information are clearly visible on the film. In contrast, the