A User's Perspective on Electronic Data Archival: The Importance of Standards

by Annette Jones Watters [1]
and Carl E. Ferguson, Jr.
The University of Alabama

Revolution probably wins the prize for the most overused characterization of rapidly changing non-violent events. However, few words better characterize the rapid rise of the microcomputer as the dominant technology of the 1980s. The device *has* revolutionized the workplace, bringing the power of electronic digital computing to the desktop. And, with speed and storage capacity increasing extremely rapidly, the microcomputer continues to transform every task associated with the acquisition, maintenance, and use of information. This paper offers a brief look at the impact of the microcomputer revolution on data distribution and archiving standards, then attempts to chart current trends and conditions in the rapidly changing technological landscape.

Historical Perspective

Digital document archival has traditionally been directed by considerations of space and convenience.[2] Space was a consideration because traditional library or reference facilities simply could not accommodate copies of all the historical information. Although magnetic tape offered relative high storage densities, even tape storage quickly became problematic. The space savings achieved by going from tape canisters stored in wire racks to hanging tape seals was quite significant. However, every unit eventually ran out of room—no one could ever buy enough tape cabinets.

Convenience. Seldom used historical data could not compete successfully with current information for shelf-space. As a result, data progenitors and librarians soon developed usage rules to help establish shelf-life and retention standards for data sets.

The Limits of Space and Accessibility

Limited space dictated that out-of-date items be compressed and/or relegated to less expensive (albeit less accessible) mediums. Numeric data was frequently transcribed from a character format (typically ASCII or EBCDIC) to a much more dense binary format. The resulting files were then written to the highest density magnetic tapes available. Standard tabulations based on these data were candidates for microfiche, 35mm microfilm, or paper microform products. Binary tapes and microform products do offer significant storage densities. However, retrieval has always been a tiresome process. In all cases, accessibility and space savings were the primary considerations and the *end-user* was frequently the loser. The end-user usually played little or no direct role in the determining the method of compression or archival.[3] Indeed, the end-user generally worked through an intermediary who selected the archival strategy. That strategy frequently was not based on the needs or retrieval skills of the end-user. A critical aspect of these archival strategies was the skills and tools available to the person archiving the data, not the skills and tools of the researcher. Archival responsibilities frequently fell to computer programmers who had no sense of the practical value of the data involved.

Machine Time

Prior to the advent of the microcomputer, most data analysts worked with paper products developed and maintained by a group of modern-day alchemists called *the programmers*. Working patiently, the analyst communicated the nature of the application to the wizard, who with cards in hand communicated with *the machine*. This was a most serious relationship, for usually there was only *one* machine in the organization — one computer to be used by all. Competition for its time and attention could be intense.

Trial and error was expensive. Research strategies requiring alternative methods of analysis were expensive. Researchers were allocated a limited amount of machine time and they learned patience. They conceptualized the table or statistical procedure to be run, gave it to the programmer, and waited. Two or three turn-arounds in the morning — maybe the same in the afternoon — meant that they *did not* spend too much time trying alternative methods or procedures.

With the development of the statistical packages, SPSS, SAS and others, the role of programmer as the analyst's interpreter began to fade — though not yet disappear. The canned packages greatly facilitated the analysis of these data and they confronted the analyst with a new challenge. The intellectual cost to the analyst in time and commitment to develop programming skills in FORTRAN or some other higher level language was almost always viewed as excessive. However, the statistical packages were different. Using surprisingly few procedural commands the analyst could now read the data and actually do the analysis. Most researchers immediately realized how much time could be saved by skipping the intermediate step of using a computer programmer. That time might now be used to explore alternative methods or forms of analysis.

Consequence

It was the limits imposed by space, time, and accessibility, that profoundly directed the data distribution and archival standards of the 60s and 70s. Programmers working for data progenitors prepared distribution tapes for programmers working for data analysts. And, programmers chose the archival standards and formats for the day that someone would want to look at *old* data sets. End-users rarely read tapes. End-users worked through programmers and it was the programmers who decided how they would communicate with one another.

The Evolution of Desktop Computing

While Apple and others were offering microcomputers in the late 1970s, the introduction of the IBM Personal Computer (PC) must be regarded as the beginning of the workplace revolution. IBM's entry into the market gave the microcomputer credibility. It was no longer a toy or experimental device for hobbyists — it was made for work and from the first day it began to recreate the office. One could say, "And the rest is history!", but there is too much to be learned from this transformation of the workplace to move on too quickly.

At first office workers were given a machine, and little else. Many quickly learned two new words — hardware and software. They learned that without software the hardware did not do very much!

Software

These microcomputers were fast and could remember things! And they did like numbers. However, they were business machines — they liked documents and numbers. The numbers they liked best were of the financial variety (spreadsheets) and the documents were correspondence. Lotus, Microsoft Word, and others quickly found their way into the market — and the world would never be the same.

Hardware

As more software and data applications became available, the 10MB hard disk quickly filled up.[4] Although the earliest microcomputer chip — the 8086 — was fast, more complex applications quickly called for more speed. Today, the fastest machine uses an 80486 INTEL chip running at 50MHz, 8 MB of RAM, and is typically packaged with a 350MB hard disk.[5] Such a machine will operate hundreds of times faster than the original 8086 and offers more total computing power than large mainframe computers of less than a decade ago.

Socialization

By the end of the decade the VLSI (vary large scale integration) silicon chip, that thumb nail sized computer, could be found in every office and on almost every desk. It was no longer a curio down the hall but rather an extension of the worker. In the decade of the 1980s it was OK for men to type and for senior executives to get their hands dirty with data.

Scientists captured data via analog ports while specialized software, running in the background, conducted the analysis in *real-time*.[6] Survey research introduced CAI. SPSS, SAS, and BMD for the PC were not far behind.[7] Never before could the analyst get so close to so much data — manipulate it, manage it, analyze, and interpret it. Microcomputer based analysis and text processing (eventually to be called *desktop publishing*) skills were fast becoming an integral part of every data user's personal skill set. Whether the analyst was a social scientist, music historian, paleontologist, or Greek mythologist, the power of the micro was sweeter than the songs of the Sirens.

User groups, first formed to provide aid and comfort to practitioners of the infant technology, disappeared as help became available from the officemate next door. *Power users* began talking to software and hardware developers, offering (frequently demanding) new features, more power, more speed. And, as the size of the market continued to grow, the software developers listened; their craft was now a multi-billion dollar business.

And so, what has become of the programmer? Who now sets the distribution standards? What has become of the limits of space and time?

Microcomputer hardware, application software, and enhanced user skills have dramatically altered the traditional role of the programmer data analyst. The installed base of MS DOS microcomputers is now measured in the hundreds of millions and the market potential for a good applications software package can quickly exceed a million dollars.[8] Microcomputer applications software developers have attracted exceptionally bright and creative systems designers and programmers with training and interests in many functional fields. As a result, researchers now have computer based tools unimaginable less than a decade ago.

Standards

The Uses of Standards

Electronic data distribution and archival standards serve the user community in several way. Standards promote

> ease of communication among users and between users and data providers;

equitable global access to data opportunities through improved documentation and communication environments; and

the convergence of distribution and archival media.

Ease of communications between data users and data provides is critical to both analyst and provider. Frequent providers include governments (national, state and local), universities and other research organizations, and businesses. While each has a unique mission in our society, as data providers, they and their user community can benefit from improved communications — improvements through mutually agreed upon standards.

The user community, public and private social and physical science researchers and analysts, share in this responsibility.

All too frequently, a *me versus them* mentality sets in. If there were a common understanding of the technical standard for providing data and common understanding of what is reasonable for the end-user to bring to the table, the level of antagonism would be reduced. These *standards of expectation* do not now exist.

Improved, jointly developed standards, are a major step toward equitable global access. Global communications today is no more exotic then a hard-wire link to your local mainframe in an adjacent building. However, to be most useful, data providers and user worldwide must work closely together to insure not just interagency or national standards but rather international (universal) agreements on media and form.

Such standards must transcend multiple platforms and operating systems. Microsoft DOS machines must be able to easily communicate with UNIX (XENIX), MacIntosh, and others. Communication standards are desperately needed to allow word processing (desktop publishing) software to easily share a document and its complete formatting. Microsoft, with its rich text file (RTF) concept is offering the market one such standard for consideration. And, of course, the need for standards can be found for spreadsheet, database, CAD, and other systems.

What has changed is the role of the user. The size and sophistication of the user community both commands the attention of the developers and shares with them the responsibility to develop and adopt global standards in each of these functional areas. Poorly written documentation and a lack of common understanding on what documentation is supposed to cover disrupts international data distribution, even without intervening language barriers.

The Future of Standards

While it may seem contradictory, standards are dynamic. Distribution and archival standards will continue to be affected by technological change. For all the progress to date, we have yet to achieve fully error-free exchange of information, easy retrieval of archived data sets, or interoperability of hardware and software. More technological changes are inevitable. The size and sophistication of the user community, high-density storage media, and an unparalleled applications software development effort have rewritten rules for data standards. And, in the judgment of these authors, it is the combination of the three that has had the greatest impact.

The globalization of data uses necessitates communication on the issue of standards. International business, international academic research, and United Nations programs are examples of sophisticated uses of data sets requiring new distribution and archival standards. The integrated European market; the political changes in Germany, Eastern Europe, and the former U.S.S.R.; potential tariff agreement in the Western Hemisphere; and strengthened copyright laws in the Pacific Rim countries will accelerate the push for easy communication among data users and promote the development of international standards.

Hands-on Users

Without a common understanding of distribution and archival standards and principles, forthcoming changes may not all be improvements. As the largest producers of research data, government agencies worldwide face the unprecedented challenge of serving a rapidly growing end-user market now numbering in the millions. Today, data end-users number in the tens of millions and all of them expect to interact directly with the data product.

During this period of flux, the development of distribution and archival standards can benefit from input by a knowledgeable user community. And, IASSIST members are on the leading edge of understanding the need for governments and analysts alike to practice "safe data." As the traditional technical role of the programmer has faded, end-users have acquired a new responsibility for safe-guarding the welfare of their own data resources and distribution channels. Yet, it is unclear how much of this responsibility users will co-opt for themselves and how much they will hand back to the professional data processing community.

What will be role of the *data archivist* as we move into the 1990s? How many of the distribution, retrieval, and archiving functions will be done by business professionals, research analysts, and computer professionals? IASSIST members will be facing these questions head-on in the coming years.

Selected Bibliography

Chartrand, Robert Lee, ed. *Critical Issues in the Information Age.* Metuchen, N.J.: The Scarecrow Press, Inc., 1991.

Snowhill, Lucia, and Meszaros, Rosemary. "New Directions in Federal Information Policy and Dissemination," *Microform Review* 19:4 (Fall 1990): 181-185.

U.S. Congress. Office of Technology Assessment. *Critical Connections: Communication for the Future.* OTA-CIT-407, 1990.

U.S. Congress. Office of Technology Assessment. *Informing the Nation: Federal Information Dissemination in an Electronic Age.* OTA-CIT-396, 1988.

Wells, Norman E.; Chow, Ivan, and Johnson, Linn D. *Document Format Considerations for a Document Tracking and Storage System.* NTIS, 1991.

[1] Presented at the IASSIST 92 Conference held in Madison, Wisconsin, U.S.A. May 26 - 29, 1992.

[2] Digital documents are defined to be those that reside, are distributed, or primarily archived in digital form — traditionally on magnetic tape.

[3] The term archival is used throughout this paper to mean the transition or transformation of digital-data from that form normally associated with daily or regular use to an alternate compressed form intended for less frequent use and/or long-term historical storage.

[4] MB is an abbreviation for mega-byte or million bytes (characters) of storage. To put this in context, an average single-spaced page of text contains approximately 3,200 characters. Thus, a 10MB hard disk could store the equivalent of approximately 3,100 pages of text.

[5] The original 8086 operated at an internal clock speed of approximately 4.7 MHz (4.7 million cycles/second). A 350MB hard disk can hold the data traditionally stored on approximately 10 magnetic tapes recorded at 6250 BPI, the current recording density.

[6] Scientific experiments now frequently incorporate instrumentation that measure such information as temperature which is automatically captured by PCs as the experiment occurs. Other programs running on PC analyze these data continuously as the experiment occurs.

[7] CAI is an acronym for computer assisted interviewing. SPSS, SAS, BMD are all statistical packages that began on the mainframe and which are now available for use on the PC.

[8] MS DOS is an acronym for Microsoft Disk Operating System. It is the dominant operating system (control program) used by microcomputers. The chief rival to MS DOS is the Apple MacIntosh.