Archive - Where it started and the Problems of Perpetuity

Dr. Parmesh Dwivedi
National Ocean Data Center
National Environmental Satellite, Data and Information Service
National Oceanic and Atmospheric Administration

ABSTRACT

In the last mass storage conference, the term "yotta bytes"(10^{24}) surfaced. That is not an unimaginable amount of data given the yield rates of advanced technologies, and yet for comparison, it's probably a couple of orders of magnitude more than all the stars in the universe. In 1934 before the age of computing, T.S. Eliot foresaw a fundamental problem with what seemed to be a data explosion, when he wrote a dramatic poem, "The Rock", where he lamented, "*Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"* Then more recently in 1985 just as the computer revolution was beginning to take hold, David Byrne wrote a song which goes, *"In the future there will be so much going on that no one will be able to keep track of it."* And more recently, Joel Achenback or the Washington Post stated; *"The moment in the future when so many technologies have converged - computers, miniaturization, bio-medicine - that they have become auto-catalytic, driving one another to yet greater sophistication, hyper-accelerating."* It is these factors which contribute to the woes for today's data manager in managing and parsing data for access from the ever growing volumes and complexity of data.

Recently, in developing a baseline architecture plan to justify the cost of doing business, the primary question was; "what do we do for a living". When we figured that out, we were instructed to describe "how" we do it. The question of "why" we do "what" we do was never asked. Then, why do we save data in the first place? Is it because storage technologies make it possible? or is it because its disposition is indeterminable and since technology provides the means, why not? or does it have historical or collectors value? or, as in the case of NOAA, does past data and information has continue to have value in the study of change?

An archive is not a process, but merely a collection of data and information. Preservation of these data and information is a process. The almost unimaginable acceleration of mass storage and computing technologies coupled with data collection and sensing technologies has lead to a significant shortfall in the processes employed for managing data. In NOAA, the processes for managing data are compounded by the fact that data and information are expected to be preserved over perpetuity. The manufacturers of mass storage media have produced products which have life expectancies far exceeding those of operating system and form-factor technologies. In fact, media can be expected to last beyond the working life of those who implemented the storage systems, and the industry continues to be driven by advancement to maintain market presence. As a result, the "hyper-acceleration" of technology only compounds the problem of perpetual data management.

The paper addresses the issues associated with preserving data over perpetuity in the face of an ever changing world of technology. It is not merely the data volumes involved, but the need to understand the issues of obsolescence, of information overload, of the needs of data discovery, of the credibility and reliability of data preserved over time, and of the necessity to expand data descriptions over time to ensure its continued value as it ages.

INTRODUCTION

The crux of the archive problem today is that the core to everything that technology requires and, yet provides as a product, is data and information, which, in turn, is used to foster future development. This paper deals with how archives started, their progression in society, and the demands and issues archivists face in today's technological explosion as unimaginable data generation powers which have now been cast among us with the ubiquitous high performance computers. It took civilization, once it became advanced, hundreds of years to record enough information to warrant the first libraries established in ancient Greece. Now, a forever expanding and enlightened global population are exploiting massive observational and computer powers to generate many orders of magnitude more data or information in just a few seconds than was generated in the previous two thousand years. The new era has instigated a major change for archivists from a world of "human-readable" data to one of "computer- ciphered" data, introducing a completely new set of issues and processes.

THINKING BIG

The magnificence and magnitude of the universe is incomprehensible. Just as amazing, is the power of our gift from heaven, our brain. Dr. Ralph Merkle of Xerox Palo Alto Research Center estimates the human brain stores about 10^{18} bits of information and processes these bits at about 10^{16} bits per second (of course, these numbers are not exact, similar to our understanding of the universe, but are close to within several orders of magnitude). Your memory is just one small database that you can access at the speed of light, and your intelligence is derived from the ability to access and retrieve data from this prodigious database. Another big thought; astronomers estimate that there are possibly as many as 100 billion (10^{11}) galaxies, and, based on our own Milkyway galaxy, each of these galaxies could contain more than 100 billion (10^{11}) stars. That would result in a Universe that is populated by as many as 10^{22} stars, or a billion trillion stars. At the same time, if the rate of processing power and the amount of data it generates continues to progress at its current rate of doubling every 18 months (Moore's law), we can anticipate annual data collection volumes to grow to a yottabyte (10^{24} or 2^{80}) in volume within the decade. This is hundred times more bytes of data than all the stars in the Universe. Where the Greeks were "thinking big" to conceive of libraries containing treasures of knowledge, we must begin "thinking big" to conceive of how we can exploit all the knowledge that we have accumulated over the millennium and now multiplying at exponential rates with today's technology explosion.

WHAT IS AN ARCHIVE?

Ancient civilizations maintained records for commerce (probably taxes or tolls), but it was the Greek civilization that conceived of a repository of information as an archive. Until recently, the word "archive" was a noun for describing public records. There is no transitive verb to "archive". The latin root word "archiva" came from the Greek word, "arkhein" which meant to rule. The Greeks regarded knowledge from which they could better understand their subjects and trade partners as a prerequisite for ruling. The Romans adopted many of the Hellenistic traits when Julius Caesar overthrew the Ptolemics in 48 B.C. in the battle of the delta and evolved a word for archive, "archiva" as they were the first real bureaucrats who retained records necessary for public services. Incidentally, it was in 48 B.C. that Caesar supposedly burned the great library of Alexandria, though the act of burning the 400,000 scrolls was probably unintentional as the scrolls were inappropriately stored in grain depots at the docks, which were the intended target of Caesar's conflagration. In fact, the Romans subsequently adapted their libraries based on the Greek libraries that originated with Aristotle. In answering "what is an archive", in today's world, as archive is a collection of records and documents, or just about anything else.

THE POWER OF DATA

As mentioned above, the Greeks were the first people who felt the preservation of knowledge had value for the betterment of their society. Aristotle established a collection of documents for his Lyceum in Athens for the enlightenment of his students. The first library of Alexandria was established by Demetrius, a follower of Aristotle, for Ptolemy I and handed down to Ptolmey II in 283 B.C., beginning the practice of establishing a library for attracting scholars. Incidentally, the first "metadata" event was established when Callimachus the poet created the first known card catalog, called the *Pinakes,* for the Royal Library in 240 B.C. which followed Aristotle's divisions of knowledge into subdivisions of observational and deductive sciences. So, as the Greeks developed the legendary library of Alexandria for the sake of collecting all the known world's knowledge, the Romans, who followed after their conquest of the Ptolemics in 48 B.C., continued the concept of a "Royal" or government library for the purpose of keeping public records necessary to govern. After the fall of the Roman Empire, manuscripts were primarily maintained in the monasteries scattered throughout the kingdoms of Europe. The monasteries, which flourished in a non-secular environment, coveted the knowledge they held in their libraries for the sake of control through religious tenets. During the ensuing centuries, Europe entered and endured the "Dark Ages" when religious tenets and public ignorance stymied expansion of scholastic holdings. In fact, many civilization records were destroyed during these times and in subsequent conquests in the name of religious heresy. During these same periods, the far eastern civilizations continued to record data and collect documents as they were a secular society. After the 15th century, record keeping

reemerged with astronomy, as astronomical knowledge evolved despite the dangers of knowledge considered to counter the monastic beliefs of the period (more than one lost their life due to theories which countered the perceived perfection of a preconceived order of the universe guarded by religious tenets), learning institutions began to flourish, and independent archive collections began to accumulate. Even as early as 1472, some 20 years afer the Gutenberg's invention of a printing press, the library at the Queen's College in Cambridge, England, had a grand total of 199 books, reflecting almost the entire record of western civilization to date.

WHY ARCHIVE?

When the new Republic of the United States of America moved its capitol from Philadelphia to Washington in 1800, President John Adams appropriated a hefty sum in those times of $5,000 to purchase "such books as may be necessary for the use of Congress". The books purchased from London consisted of 740 volumes of primarily law books and three maps. For then, almost all literature and art was in private hands. When Thomas Jefferson succeeded Adams as President in 1802, he expanded the role of the Adams library as the first library of the American government, now known as the Library of Congress. When the British sacked and burned Washington in 1814, the 3,000 volume library was destroyed. In 1815, the government bought Jefferson's private collection of 6,487 volumes for almost $24,000 to restore the library. In 1850, the Library evolved into a National Library as Joseph Henry, the secretary of the Smithsonian Institution, argued that the Library of Congress should be viewed as an appropriate foundation for "a collection of books worthy of a Government whose perpetuity principally depends on the intelligence of the people". This is a reflection of the Greek feelings almost two millenium before. The copyright law of 1870 which stipulated that two copies of every book, pamphlet, map, print, photograph, and piece of music registered for copyright be deposited in the Library, gave cause for expanding the Library. Finally the Library was funded to be housed in its own building. In the argument for support of funding to build a Library building, Senator Voorhees of Indiana eloquently expressed his Jeffersonian belief in the essential moral value of books and intellectual property by stating: "Knowledge is power, the power to maintain free government and preserve constitutional liberty. Without it, the world grows dark and the human race takes up its backward race to the regions of barbarism." So from that, the Library of Congress became the National Library to preserve not only the intellectual property of this country, but also to preserve it for the intellectual enlightenment for the betterment of society.

For the first 150 years of our country's history, most government records were stored haphazardly, falling victim to fire, water, and neglect. Historians and federal officials realized that the records needed a permanent home that would be accessible to the government and the public. Government operations at all levels were collecting information regarding population and demography (census), property deed records, marriage records, and birth and death certificates. Finally in 1934, the National Archives and records Administration became a reality. Where the Library of Congress preserved an archive of documents originally intended for reference for the Congress and as well as documents with cultural value, the National Archives and Records Administration was chartered to

preserve the official records produced by the operations of the government. In this role, the National Archives ended up with many historical documents and artifacts, to such an extent that it's function as an official government records repository extended to co-function as a museum and historical reference library.

A CHANGE IS IN THE AIR

Star catalogs developed by Hipparchus around 150 B.C. and miraculously preserved as an early scientific record, became a valuable record to detect change when novae or "stellar guests" appeared. Only recently have we begun to realize the value of data and information as the world is changing at an alarming pace as technologies rapidly evolve. Until now, little attention was given to preserving information outside of public records and historical documents. Much of the actual data and information used in developing and publishing scientific study results were withheld, as data were guarded by industry as proprietary for economic advantage and coveted by researchers for personal claim. Other data records were simply regarded as working copy data without consideration of disposition for archive, and access to data was generally restricted to those few who had the capability or access privileges to parse through data repositories. Since data and documents were generally shelved, the processing of parsing information from the "stacks" was time consuming requiring prior knowledge of where and how to look. The value of an archive wasn't generally appreciated and, as a result, received little economic support. As the stacks grew to overflowing and treasures of information became buried under the pile (today, many documents in the Library of Congress are stacked on the floor for the lack of shelf space), the problem of finding data was exacerbated by the rate of influx of new volumes. It was as if the forest of data grew so much as to block out the light with which to see under the canopy. Then, in less than a decade ago, the Internet and its prodigious outreach capabilities became a reality. Suddenly, the value of data increased as more and more people suddenly had a method for readily exchanging information. The Universities first, followed by the government, seized on the concept of placing data and information into accessible realms to exploit the original intent of the Internet. As an outgrowth, search engines evolved to enable more efficient and broader community utilization of data and information, and now anyone with Internet capability has access, with paradigms to assist in the "hunt, search, and browse" processes to prodigious volumes of data and information. As the people became more in tune with the world and environment around them, a greater appreciation of understanding the changes that are now appearing so rapidly aroused interest in the study of change. It's as if we suddenly realized we were running out of space, as the new age of computers and data have shrunk the world. Now, data in repositories have increased value as an asset for studying change. The driving force behind the efforts to collect almost unimaginable amounts of data is to understand and appreciate change. What we place in archives is longer just for enlightenment, but now, it's a question of "what's going to happen to us?" Perhaps historical data can offer some clues to these questions, and as we look at more information at a quicker pace, answers and many more new questions will begin to surface at an even faster rate. And, of course, the pile of data will become increasingly big.

JUST COLLECTING DATA

In 1934, T.S. Eliot foresaw a fundamental problem with what seemed to be a data explosion even then when he published the dramatic poem, "The Rock", where he lamented, "*Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?*" Archivists are reluctant to throw anything away. Keeping everything was reasonable in the past, as there were few copies available as compared to today. After all, there was only a few stone tablets produced (only 90 or so have been recovered) and if it had not been for the Rosetta Stone discovered in 1799, we might not been able to interpret most of them. Just a few more hand-scribed manuscripts survived, and then, an explosive growth of circulated copies after the Gutenberg printing press was developed in 1455. At about that time, the library at the Queen's College in Cambridge, England, had a collection of all the world's knowledge in all of a 199 books. Now, almost a half million books and journals are published annually. The Library of Congress has accumulated 113 million items in the 200 years since its inception. The National Archives as accrued over a billion pages of texts and documents in its short life of 65 years.

Now, the computer revolution that evolved in the last four decades have changed our entire concept of collecting and managing data and information. Earlier, the statement was made that the core to everything that technology requires and, yet provides as a product, is data and information. As computer technologies advance, the capability to observe, assimilate, and output incredible amounts of data accelerate at even a faster pace. The result is, today data are being provided at rates beyond which we can cope. If you haven't noticed, the information explosion is happening, or perhaps, the explosion is just beginning to happen, and, in an analogy to the creation of the universe, we are in the early phase of the explosion and have a universe void yet to fill.

The amounts of data collected from natural and science-induced events initially were limited to the capability to observe events and to transcribe the events into a recordable form. After all, there were only so many stone tablets. There was only so much time in an adverse environment for Lewis and Clark to record the events and discoveries of their amazing journey across the North American continent. Even early computers were restricted to assimilating just a few inputs and producing even fewer results. But as recording capabilities expanded and the ability to detect and produce data from events and the ability to capture the sensed information expanded, the amounts of information started growing faster than the exponential rate of technology advancement. Weather observations are a good example. Just a couple of decades ago, weather observations were transcribed to digital form on an hourly basis, though analog recorders provided a continuous data record. There were a limited number of stations, 154 stations in 1888 reporting once a day, growing to 250 or so reporting hourly in the U.S. In 1997, the Automatic Surface Observation System (ASOS) was installed, and when fully deployed, can provide as many 1700 sites observing weather events in digital form on a minute by minute basis. The data gathering potential thus increases by a factor of 400:1. At the same time, the U.S. national Doppler radar system was installed, providing a digital data record at prodigious rates. This is nothing compared to the amounts of data being generated by particle super-collider systems used in atomic research. They can accumulate data at petabyte (million billion)

volumes in the matter of months. The reason space borne sensors do not collect more data than they do (the civil earth observing environmental satellites are expected to collect about less than five petabytes of data from their inception in the 1960s through the year 2015), is because of cost and political constraints which limit power resources and broadcast options.

SUDDENLY EVERYTHING IS DIGITAL

Digital television is around the corner. The entire collection of commercial audio-recorded data are now available in digital form. Also, telephony is becoming all-digital, publications and entertainment media are digital, and as more people come on-line, messaging is evolving to all digital. This has a compounding effect on the amounts of data be collected and saved in digital form. Technology enables anyone to place data on-line so people can have access to it. The Internet is a perfect example. In 1999, it is estimated that there are over 50 million web sites and 300 billion pages of information available over the web. Since everything associated with printing these days is now in digital form from inception to the press, it is easy to place material in a data base and provide it over the Internet. This digital information explosion has resulted in entirely new paradigms for managing information, evolving from dealing with "human readable" records to that of dealing with information that requires a translator tool – the computer. The concept is no longer keeping track of documents held on a shelf, but controlling or losing control of indecipherable information. Suddenly, the methods of managing archives developed over the last two millennium are forced to undergo a radical change.

IS THERE A DIGITAL CRISIS?

First of all, no technology is finite, in fact, as the technology curve bends at a more exponential rate, the half-life of systems, form factors, and interface requirements will become increasing shorter, placing data at risk. Even when dealing with "human readable" documents, archivists have problems with preserving data because new lower cost, mass produced paper products are acidic resulting in deterioration problems. The same is true for film products as emulsions tend to fade (emulsions never cease to be active). In the past, the major concern of the digital archivist for the loss of data, aside from fire or storm damage to the storage facility, was because of media deterioration. However now, at least for the digital archivist, the problem has an added dimension of system and form-factor obsolescence brought on by the fact that media technologies have succeeded in producing highly stable materials havin significant shelf life. For example, magnetic tape used to develop physical deterioration problems in ten years or so. New encapsulated magnetic tapes now are expected to have a shelf life of 25 or more years. Optical media is touted to have a shelf life of up to 100 years. With the technology curve being what it is, even the short lived 25 year magnetic tapes will face obsolescence problems long before the media experiences physical changes, putting data at risk. This is no more apparent than to

personal computer users who find that documents created only a few years ago cannot be read by new, supposedly improved applications. Suddenly, everyone is at the mercy of manufacturers for preserving their data. The manufacturers look at their products as a short half-life, throw-a-way item, as they continually advance technological capabilities to maintain market position. In the market place, the majority rules. If the data archivist is not in the majority, as they obviously aren't, they lose, and the data are at risk.

CAN WE MITIGATE OBSOLESCENCE?

The answer is probably not entirely. Unfortunately, no one is willing to make decisions on what data should be preserved versus what is not essential. Most say everything should be saved, but when faced with the sometimes enormous cost of migrating data from one form to another, a choice has to be made on what is saved. This problem is compounded by the volumes of data being captured.

Three major issues are brought to bare when considering the prospect of preserving data indefinitely. One is simply keeping track of what you have. Another, is being able to read the information in a current systems environment. And the last is, the cost of moving data from a prior technology to a new one. For example, if one collects data for ten years onto a media sized to a particular data rate, then presumably it would require ten years to re-record it onto a new media even if the new media system is orders of magnitude faster than the prior system (as the truck bumper sticker stated, "I may be slow, but I'm in front of you"). In this case, one would have to go to the expense of replicating the initial recording capability by whatever speed up factor they wished to migrate the data. For example, if the effort is intended to migrate a ten year data set in one year, ten of the original units would have to be acquired to complete the task in one year. Then there is the problem of acquiring the older technologies often no longer produced. The result is inevitable, there will be data that will not be recovered. Even the most perfectly controlled and documented data will face this problem, and the problem severity is directly proportional to the volume of data in question.

For example, a recent Washington Post article on March 12, 1999, pointed out that the Library of Congress intended to digitize 10,000 out of the 113 million human readable texts they expect to have by the year 2001. At the same time they are accumulating new texts at the rate of 10,000 every two weeks. This is equivalent of trying to use a two gallon bucket to empty a reservoir fed by a fast moving river (remember the movie *Fantasia?*).

WHAT NOW COACH?

Technology is following a curve that promises to make the media more indelible, in some instances mitigating, and others complicating, the data storage problem. This places a greater emphasis on adhering to standards for recording data, and for extensive indexing of the data to ensure future access. Hopefully, access is a key element in the data archive paradigm, as that should be the criteria for saving data in the first place. The question of whether data and information is a national asset

or a national heritage has to be addressed. Using a historical model, data always starts out havine an asset value,having economic, military, or political advantage, and ends up as a heritage trove having historic value for science and education. Generally, there are economic reasons to fund the management of data while it is an asset mode and little appreciation or understanding of the value of maintaining data once it enters a heritage mode. Preservation of heritage data suffers from lack of support because it's beyond a "budget cycle", or "it's not on my watch", or the historical value is not fully appreciated or understood. The result, data continues to accumulate, now exponentially, because no one will decide on its fate. More often than not, the decision of whether to preserve data or not will be made for us as a result of readability obsolescence.

WILL WE BECOME LOST IN THE FOREST OF DATA AND INFORMATION?

In 1985, David Byrne wrote a song called "In the Future", in which he lampoons those making predictions, where the song goes, "*In the future there will be so much going on that no one will be able to keep track of it."* As the market place expands, the progression of advances in computing capabilities continues to double about every 18 months with the same physical and cost constraints (a fact attributed to Gordon Moore, Chairman Emeritus of Intel). Part of the reason for this rapid evolution is the fact that technology tends to feed on itself. As far as the future is concerned, there is no reason that Moore's law will not continue to hold. The pace of development and capabilities will compound the rate at which data are being generated. The major obstacle to advancement will be managing data for access. So as we compute at teraflop rates, communicate at terabit rates, and store data in petabyte, exabyte (billion billion) or yotta bytes (trillion trillion) quantities, will we be able to parse information out of all the data that will be generated as a result? I wonder. Also, how can we afford the cost and effort to save it forever? My guess is that we can't and won't, and data will be lost. This brings up another point, with so much data streaming from so many sources, how is the credibility of all these data assessed? Will the proportion of misinformation grow at the same rate as credible information, and how will we ever know? As a result, future emphasis must focus on data management processes extending the data warehousing, and data mining just started to enable future generations to even begin to extract value from what will be the heritage data sets in their time. Or will the new generation's "throw away mentality" be content to abandon data and information out of a perpetual fear of obsolescence? Or, with increased use of "snippets" of information in today's increasing fast pace of life, will the new generation be content with only "snippets" of heritage data? I suspect there may be no other choice.

Then, there is a more disturbing factor. Large data producers are becoming more obsessed with fast access to data than the preservation of data. The new "hot button" in data management is Storage Area Networks (SANs) where data are physically distributed but managed as a logical whole in order to accelerate access to the data. Data management paradigms are concentrating on instant access, where data are striped across multiple storage devices to relieve access contention while at the same time providing a high degree of access fault tolerance to mitigate instantaneous data loss. This places all the control of managing data in the realm of access taking away any form of control for

the data in a repository environment. The process of parsing data into segments of files or objects for the purpose of access creates an environment where data could lose its identity as a whole. In this fast paced world of providing data services, there appears to be less and less consideration given to archive management. This is happening now because access drives the market place.

PRESERVATION

The is a likely hood that some data will be lost, either because of inattention due to the lack of funding for perpetual management, or, data parcels become buried in the pile of data accumulating at prodigious rates. As a result, we must begin to characterize or summarize data and information to that extent necessary to provide some lasting utility in the event of loss. But, even before this problem is addressed, there should be some filtering mechanism to determine which data should qualify for archive and once selected, its degree of credibility. Data, like many things, has a purity or pollution problem. With the exponential growth of data and information, misinformation will similarly grow, creating a domino effect on the future integrity on an entire population of data. So, along with information management, there is a mis-information management requirement. At the same time, the sources of data and information are becoming distributed to such an extent where a focused characterization of data and information may become an formidable task. All of these factors point to the reality that no data set can ever be assumed to be perfect. As a result, the act of preservation not only includes the requirement to protect data and information from technology change, but must also be concerned with attending to its credibility factor. This leads to a requirement for far reaching descriptive information to accompany the data as it ages to service future access paradigms, and to serve as a last gasp representation of data in the event of partial or complete loss. With all things considered, the processes of preserving data to retain its utility can result in an accumulative cost that exceeds the cost of collecting the data in the first place. The real decision then, is it worth it, and if so, where is the funding?

AND FINALLY

In April 19, 1998, there was an article in the Washington Post newspaper by Joel Achenback, addressing the approaching singularity of technology. It states: "*The moment in the future when so many technologies have converged - computers, miniaturization, bio-medicine - that they become "auto-catalytic", driving one another to yet greater sophistication, "hyper-accelerating." Predictions will be worthless because everything is changing so fast - an event horizon beyond which we can detect nothing!*" And, as a subsequent Washington Post article pointed out, with the information glut, the world is not ending, it is just becoming incomprehensible.