"TITANIC 2020 - A CALL TO ACTION"

by Rich Lysakowski and Zahava Leibowitz

CENSA

Research Publications

"TITANIC 2020 - A CALL TO ACTION"

by Rich Lysakowski and Zahava Leibowitz

1 Executive Summary

The potential electronic crisis predicted to hit the Western world upon entering the Year 2000 was the subject of global concern for many months, if not years. The airwaves were filled with the speculations of everyone from diskjockeys to systems analysts to politicians on whether public utilities, banks, or nuclear weapons would fail. Y2K Plus, a company specialized in solving Year 2000 computer problems, estimated that the cost of prevention and repair of Y2K technical problems will total approximately $750 billion dollars. With legal preparations and litigation the total cost will likely be over 1 trillion dollars.[1] Intense preparations for years, replacing old computers and rewriting millions of software programs spared the world from major problems.

While it is no wonder that Y2K inspired thoughts of a coming apocalypse, a much larger and very real crisis is on the horizon now -- one we call "Titanic 2020." It has been brewing for a long time and grows larger everyday. Titanic 2020 results from software business practices that make new products quickly obsolete, cause major hassles for users, and put critical assets at risk. Comparing Titanic 2020 to Y2K, Y2K was like the tip of an iceberg. Like the passengers on the Titanic on her maiden voyage, many of our most valuable records stored by the ship of computers will perish.

Icebergs are dangerous because they do not reveal their size until we get very close to them.Unlike the iceberg that sank the original Titanic, the iceberg looming in our future is man-made. Titanic 2020 is not just one iceberg or one definite deadline like Y2K -- it will be many icebergs scattered across many future horizons. By then it is too late because the real danger is below the horizon.

Simply stated, Information Technology (IT) of the early 21st Century -- and the organizations that use IT -- are generally ill prepared to prevent damage or loss of valuable electronic records or data. Every time we migrate electronic records or data stored in computer systems to the next generation of technology, we potentially face perilous collisions. Some collisions will lose priceless personal data and records. Until some important changes occur in Information Technology itself, large losses will mount. Every organization that uses computers for storing critical data or legal records will have Titanic collisions and full-blown disasters from irrecoverable data and violated records. Countless personal and professional lives will be lost in failed rescue attempts. Costly lawsuits to recover damages for lost properties will tie up courtrooms for years.

This report identifies the primary causes of the impending disasters we will all face as we rush into our digital future. It provides an economic analysis of a recent, extremely costly software migration that resulted from incompatible data formats. The example is at once practical, symbolic, and symptomatic of the problems for technology users caused by the software industry at large choosing a strategy of rapid planned obsolescence. The report also admonishes users to take back ownership of access to their data because today, guaranteed long-term access to records requires a continual subscription with an unknown cost-of-ownership price tag. Lastly, this is a Call to Action to reduce the risks and costs of ownership of Information Technology to finally support permanent electronic records by changing business strategies and buying behaviors. The Titanic 2020 problems are well understood, and will persist by design, until this Call to Action is heeded.

[1] See Ted Smalley Bowen. "Y2K Legal Wrangling Escalates." InfoWorld, November 29, 1999, Cover Story.

2 The Origin of Titanic 2020

The Titanic 2020 icebergs started to nucleate at the start of the Information Age, roughly in the mid-1960s, when we started to rely on electronic computers for business records. The icebergs have been growing rapidly in size and number since then. IT systems of the early 21st Century create and store "fragile" electronic records and data, meaning their integrity is easily broken when they are removed from the technology that was used to create them. Once broken, an electronic record's integrity cannot easily be restored. At first, electronic records were mostly historical in nature, but increasingly we create them for critical business and scientific matters too. For society to move all the way into The Information Age, we need to replace paper and microfilm as the default permanent storage media for data created electronically. We must still take some important transitional steps.

The Y2K problem was the result of using a shorthand, two-digit year code instead of a full four-digit date code. However, for modern electronic records, this date code is only one item in a long list of parameters that must be kept to ensure integrity and access in the future. Like the Y2K problem, which is the result of a lack of concern over the past 30 years with a known parameter problem, the problems leading to Titanic 2020 are also known, but little to no action is being taken to correct them. We continue to produce digital records in formats with short life spans that are increasingly difficult and sometimes impossible to access years later.

"Titanic 2020" is a moniker for the serious problem of widespread software design flaws in today's IT systems. These design flaws will affect every organization that must keep today's electronic records well into the 21st Century. Every organization is building its own fleet of Titanic ships in the form of electronic records technology systems and data systems. Many of the records these vessels carry will be shipwrecked on the dozens of icebergs growing in the future seas of IT. Each migration event is a point in time when records must change carriers and preserve their full integrity and accessibility. In front of each migration event on the horizon floats a potential man-made iceberg with a size and shape unknown at this time.

The designs of today's IT systems simply do not permit facile or reliable migration of their contents, at a predictable or reasonable cost. Unwittingly, by our own design, we keep creating the ships without lifeboats to carry us safely into the future. Not only are we as a society failing to take note of the problem looming over us, like the builders of the Titanic, we are continually taking steps to seal our fate.

3 Focus of This Report

This report focuses primarily on legal records; however, in most cases it applies equally well to less formal or secure compilations of "data." Legal experts widely agree that electronic records and data are admissible in court as evidence. They also agree that electronic data can have a legal weight equivalent to formal, signed legal records -- as long as the data integrity has been preserved beyond a shadow of a doubt since the time of its creation to the time of its surrender in legal proceedings.

This report focuses on matters of critical concern to individuals, businesses, and governments.

Individuals may think a report about electronic records does not apply to them, since they may use personal digital devices to create data for their eyes only. However, this report applies to any collection of any information with integrity that you may want to preserve and access years later.

At the individual level, personal digital devices make our lives easier on a daily basis. We increasingly rely on them in our daily lives. However, digital devices do not yet provide reliable and easy ways to move our digital results across generations of technology. Digital records sometimes change irreversibly as we move them across different devices and software packages over time -- sometimes altering critical characteristics of authenticity, reliability, or accuracy that allow them to be used as historical records later. We increasingly entrust priceless sentimental works and events to digital-only devices. If a record's content is not entirely lost in migration, sometimes it can be corrupted or damaged irreversibly, a fact which may only be discovered after great personal loss of time trying to recover it. Such time and effort spent to reliably migrate digital data through software versions over time needlessly encumbers and wastes our personal lives.

At the levels of business and government, records are the lifeblood of accountability for past actions, and data are the basis for current actions. Data can be useful for very long periods of time even though it changes from one data processing software to the next, and from one computer system to the next. At some point processing stops, conclusions are reached, and a permanent record is created. In the rush to become ever more efficient and effective, manage knowledge and leverage intellectual capital, we have overlooked the vessels to carry electronic records of action far into the future for both legal and historical reasons. In any endeavor, seeds for tomorrow's success come from today's intellectual property and other recordings of knowledge. We are killing those seeds by using vessels that can easily sink and destroy the integrity of the seeds. At the same time, we are creating unknown legal and economic problems and losses for ourselves that will be nearly impossible to estimate until they are upon us.

4 The US National Archives: Experienced Sea Captains in Need of Better Ships

As the keeper of the government's records, the United States National Archives and Records Administration (NARA) was among the first groups to experience, on a relatively small scale, the danger that lies in front of us all. Its buildings in and around Washington, D.C. are packed with as many records as they can physically hold, including many original audio and video recordings of historic moments in time, all on various forms of outdated media. In order to access these media, the original audio players and video projectors must be reconstructed and then each recording must be played -- sometimes in its entirety -- in order to catalogue its contents and separate out accessible from inaccessible information and key historical information from the unimportant. Due to a lack of funding and resources, many of these records will molder, crumble, or rust before they can be reviewed and catalogued by archivists. It is part of NARA's responsibility to decide which records submitted to them they will have the resources to preserve in an accessible form and which are beyond their resources (and may be lost forever.)[2]

Digital media such as digital audio, digital video, and other more complex digital data types (scientific, engineering, and business data) are rapidly proliferating. Digital media is quickly becoming the preferred recording method -- strongly favored over traditional paper and analog methods. Digital formats permit innovative computer processing, searching, mining, and reuse of all types of knowledge. However, accompanying this ease of use is "The Digital Time Bomb" - an indefinite period of time after which data is inaccessible because no one paid money or attention to ensure its accessibility. Complex digital records are increasingly difficult to deal with because there are few universal standards for long-term preservation and access. Each software supplier creates its formats and changes them as needed for a variety of reasons. Many people hope that the World Wide Web and technologies such as the eXtensible Markup Language (XML) will offer some stability and solve universal access problems. However at the current frenetic pace of innovation and market competition, chaos will remain the norm until enough end users and suppliers are organized and work together effectively to produce stability in the marketplace. Currently, nearly everyone using computers is forced into a defensive position of preserving records in the lowest common formats -- in the worst cases -- by printing to paper.

5 Modern Records: They Don’t Build Them Like They Used To

The problem faced by the US National Archives is in many ways much simpler than Titanic 2020. The records stored at NARA represent only the history of the US Federal Government. NARA's repositories hold historical records of operations of all branches of government and public service since the United States was formed. NARA's records are the primary instruments of accountability of past public administrations. NARA typically does not receive records from a public administration until after it leaves office. While the records hold tremendous value with respect to our understanding of who the American People are and how they got to the present, loss of these records is unlikely to dramatically affect the day to day lives of most citizens. Such is not the case with most electronic records being created today in business or by individuals.

Most electronic records today are produced and relied upon in the normal course of everyday business. Individuals become most aware of how much they are affected by the integrity of these records when their bank statement or credit card bill has an error. Businesses are even more

[2] See Stille, Alexander. “Overload.” New Yorker. March 8, 1999, p38ff.

dependent on the integrity of their records, since record integrity is critical to defending patent rights, demonstrating compliance with government regulations, defending tax returns, etc. As the world population grows rapidly, and society becomes increasingly litigious, the number of records that individuals and businesses must create and depend upon is increasing exponentially.

6 The Growth of Electronic Records

Accurate models of the growth of electronic records are difficult to construct, but we know that several factors are important. The most important ones are population growth and the accelerating adoption of information technologies to achieve greater efficiencies of scale and sharing of knowledge than paper-based systems can provide.

Models of growth for the Earth's population provide the most compelling drivers of growth. Population experts predict that the global population will expand to around 11 billion by 2050 and potentially to 24.8 billion by 2150.[3] It is obvious that for each person an increasing number of vital records must be kept: certificates of birth and death; marriage and drivers licenses; mortgages, stocks and other property transactions, plus many other types of routine transactions that we need or want to record. With so many transactions of lasting legal value, it is easy to see that the number of records will continue to grow exponentially. In fact, most institutions today -- whether corporate, government, or even small businesses -- harbor not-so-secret fears that the volume of important records is growing so fast that they will soon become, if they are not already, impossible to manage.

NARA provides a good example of the explosive growth of electronic records. Between 1972 and 1989, NARA archived 5000 electronic record files. Between 1989 and 1999, it archived 100,000 electronic records. In 2000, it expects to archive at least 1,000,000 records. In 2002, NARA expects to archive at least 40,000,000 e-mail records from the Clinton Administration White House alone. Simultaneously, it is preparing to archive approximately 90,000,000 personnel records from other branches of government.[4] NARA does not archive state and local records, that is the duty of state and local archives. Nor does NARA archive the millions of records of experimental and medical research, tax records, or countless other types of confidential records that could be important to future citizens. Generally each institution must archive their own records, unless they are of particular historical importance to the public.

Jean-Pierre Wallot, the former National Archivist of Canada, stated in 1994 that "more records have been generated during the past ten years than in all of prior human history."[5] Because of the rapid adoption of computers for business in the Information Age, the time required to double the number of critical records is decreasing rapidly. At the current rate, the number of records will double in 5 years, and double again less than 3 years later. If the number of records continues to grow at the same rate of growth as that of the human population, the numbers are staggering; simple math tells us that ***within 10 years, the number of records produced on the planet could be doubling every 60 minute*s.**

[3] Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat (1999). *Long-range World Population Projections: Based on the 1998 Revision*. (ESA/P/WP.153).
[4] Ken Thibodeau, Director of the Center for Electronic Records Archiving at NARA supplied these statistics.
[5] See Wallot, Jean-Pierre, The National Archivist of Canada. "Archival Oneness in the Midst of Diversity" American Archivist, Vol. 57, Number 2, Spring 1994, pg. 380.

Luckily, several mitigating factors will constrain the problem somewhat. Good records management practices dictate that business are only required to retain records for the number of years mandated by law. Some types of records can be deleted soon after their creation. For example, electronic commerce transaction records are short-lived, requiring retention for only two to three years. The IRS requires that personal tax returns only be kept for 7 years.

The Titanic 2020 problem is compounded by the fact that longer retention periods are mandated by law for some types of business records. The Food and Drug Administration (FDA) requires records for demonstrating regulatory compliance to be retained for 25 years or more. Patent records must often be maintained for as many as 35 years (including filings for claim extensions). Trade secret records (e.g., formula for Coca-Cola, for example) must be maintained until it is made public, if ever. The FDA requires records created during development and manufacturing of implanted or replacement body parts (pacemakers, organ transplants, and, in the future, genetic therapy records) to be retained for at least as long as the affected patients are alive. The FDA also recently introduced a new law that requires electronic records to be kept in addition to paper records.[6] In the future, other federal agencies will make this a requirement as well. It is not enough for these records merely to be retained, they must maintain their integrity and be accessible. Preserving record integrity and accessibility can be particularly difficult since this involves the ability to view and print an accurate and complete visual rendering of the record at any time it is needed in the future.

7 Rapid Planned Obsolescence: The Heart of the Titanic 2020 Problem

With regard to records created over the past 10 years, The National Archives faces a relatively small problem that is the result of old technologies becoming obsolete and unavailable. Individuals and businesses that create and drive our economy, our government, and our educational and research institutions face a similar problem of Titanic proportions. Whether or not data integrity is compromised during a migration (a potential financial burden in itself), a tremendous financial burden accompanies almost all data migration. Format incompatibility results in reduced productivity, which becomes a constant and costly problem. Even though modern technology allows electronic records to be created, accessed, and searched more quickly than records in any other form, because of planned obsolescence and the rapid decrease in time to obsolescence, this same technology ensures that these records must be regularly migrated to the newest available technologies (which are frequently unable to read the older formats). Miss some migrations and you may lose access to critical records entirely or regularly migrate data and absorb the associated financial loss. Neither choice is one we can afford.

[6] See www.censa.org/21CFRpart11.html or www.21CFRpart11.org for extensive information on this regulation.

8 A "Day-In-The-Life" Hassles of Migration and Incompatibility

File incompatibility is a frequent problem between old and new versions of the same software. The two examples given here, Microsoft Office and Lotus SmartSuite, were chosen only because they are familiar to many people. The problems described here are so commonplace that almost any other software for creating or managing data or records could have been used to illustrate the problem.

In the case of Microsoft Office, a file incompatibility exists between the Office95 and Office97 versions of MS-Word, which for simplicity we will call Word'95 and Word'97.[7] Although the price of the software upgrade was only $80, the incompatibility creates many hidden costs in the version upgrade. The hidden costs arise from the loss of productivity as users unsuccessfully attempt to migrate documents from Word95 to Word97. More hidden costs arise from the need to train employees on all the differences between the two product versions, as well as strategies for migrating documents. Many other hidden costs exist depending on the context of the upgrade, e.g., personal versus business. Hidden costs associated with software upgrades are usually many times greater than the actual cost of the software.

For example, when Word'97 was released, documents created and saved in a Word97 format could not be accessed via Word'95. Thus, when a Word'97 user sends a Word'95 user a document, the latter is unable to open, view, or print the Word'97 file. If the Word'95 user is unable to speak directly to the Word'97 user, all work stops until she may do so. When the Word'95 user and the Word'97 user manage to connect, they may quickly discover the problem is due to the fact that they are using different versions of Word, if they are experienced or sophisticated users. The Word'97 user then resaves the file in the older format and resends it. If the users are inexperienced, it may take them a long time to discover what the problem is and how to solve it. Additionally, even though the Word'97 user saves the document in the older format, if the document contains new (and incompatible) formatting commands, some parts will look jumbled when it is opened in Word'95. The Word'95 user must again contact the Word'97 user to determine whether or not the document's appearance is correct. If one or both of the users are experienced or sophisticated, they may quickly realize that the new formatting will either need to be removed or manually reproduced in the old format. If they are not sophisticated users, they may not be able to recognize the source of the problem, or they may conclude that the document file is corrupt and must be entirely regenerated from a much older version, or even from scratch. In any case, manually reformatting even part of a document can be tedious and time consuming if the document is complex and uses many styles of headings, tables, and figures. When the Word'95 user believes the reformatted document is finally complete and correct, the document is generally faxed to the Word'97 user for a final review. At this point, after spending anywhere from a few minutes to several hours on solving problems due to incompatible format issues, many users will choose to fall back on the reliability of the older technology.

Incompatibility problems were very common for the first two years after the release of Office'97. Many large organizations have a policy of not upgrading to newer versions of software for many months or even years after new versions of software are available in order to avoid bugs and other

[7] The version of Word released with Office95 was actually Word version 7.0. We refer to it as Word95 to avoid confusion between MS-Word version released with Office95 and MS-Word version released with Office97.

costly problems inherent in popular software. Unfortunately, the longer a company waits to upgrade, the longer they and the organizations they work with must suffer because of incompatibility problems. Sometimes vendors will create point solutions to these problems but fail to announce that such solutions are available. For example, about 18 months after the release of Office'97, one of the researchers of this article called Microsoft to find out if a Word'97 to Word'95 translator was available for Word'95. After paying $35 for the technical support call, the author was informed that a file translator was recently posted on the Microsoft Technical Support website. To the authors' knowledge -- confirmed by polls of fellow users of Word'95 -- no notification of the availability of this file translator was ever sent to customers.

Often, some people assume that with enough finagling, it is possible to work around the incompatibility problems between older and newer formats, but this is not always the case. Complex documents, particularly those that include graphics and tables may convert very poorly, some not at all, and some may actually become corrupted. For example, it is not uncommon for pointers (invisible electronic markers that indicate what should be displayed where in a document) to be misinterpreted, damaged, or destroyed in the process of opening a file in a different format. Such corruption often results in the total loss of formatting and/or content.

9 Economic Model of the Users' Cost to Upgrade from Office'95 to Office'97

The following economic model is based on a population of 100,000,000 users at work. One hundred million was chosen because in 1997, Microsoft claimed to have sold at least 80 million copies of MS-Office. After only two years, this would add up to more than 100 million users.[8]

For calculations, it is assumed that each user work 37.5 hours per week, 48 weeks per year.

During the first year after the software upgrade's availability, it is assumed that 10 minutes was lost per week per user. This is an extremely conservative number; many users reported losing much more time before recognizing and learning to solve the problem -- even fairly sophisticated users.

Year 1 - Time Lost		Description
Minutes	10	Minutes Lost on Average per Week
	(10 x 48 weeks) = 480	Minutes Lost Annually per User
	(480 x 100,000,000 users) = 48,000,000,000	Minutes Lost Annually by Total User Base
Years	(48,000,000,000 ÷ 108,000[▽]) = 444,444	Work-Years Lost by Total User Base

[▽] 108,000 is the number of minutes in a work-year (37.5 hrs/wk * 48 wks/yr)

In the second year after the software upgrade's availability, it is assumed that time lost was reduced 50% because people learned to recognize and solve the problem faster, and more people have upgraded to Office'97. Thus, during the second year each user lost an average of 5 minutes per week. This means that 222,222 work-years were lost by the total user base in the second year.

Similarly, in year 3 after the software upgrade's availability, it is assumed that the time lost due to incompatibility problems was only 1 minute per week per user, and 99% of users have upgraded. As a result, 44,444 work-years were lost by the total user base during the third year.

The cumulative time loss derived from adding the conservative numbers above is astounding. For every 100 million users, a conservative estimate of ***711,110 total work-years*** [9] were lost during the first three years the software upgrade's availability. To translate these work-years into United States dollars, the total work-years is multiplied by cost per employee per year. A cost of $110,000 US dollars fully-loaded annual headcount cost was selected.[10] Thus ***the total labor cost due to this particular file incompatibility problem was $42 billion dollars.***

This cost of lost labor does not include the upgrade purchase price, which is also substantial. To calculate the cost of the upgrade itself, the model assumes that for a group of 100 million users, it assumes 25% of these users upgraded during year one, 75% of these users upgraded by the end of

[8] A full spreadsheet model with the variable model parameters is available for download at www.censa.org.

[9] A work-year equals the amount of work one person would do in one year when working. In this example, we use 37.5 hours per week, 48 weeks per year.

[10] Fully-loaded annual headcount cost means the total annual cost of an employee to an organization. This include salary, benefits, other compensations, and other expenses on a per employee basis (utilities, rent/mortgage, office supplies, etc.). Fully-loaded costs range from $70,000 for low-wage earners in small organizations to over $225,000 for high-wage earners in large organizations. A median of $110,000 was chosen as a conservative annual headcount cost.

the year two, and 99% of these users upgraded by the end of the year three. At a cost of $80 per upgrade, $6 billion dollars were spent the first year, $2 billion dollars were spent the second year, and $80 million dollars were spent the third year after availability. This means that the total cost of the software upgrade to the user base was about $8 billion dollars.

Adding the lost labor costs and upgrade purchase costs, for 100,000,000 users, ***the upgrade from Office'95 to Office'97 cost the user community $51 billion dollars.***

Unfortunately, the *actual* cost of incompatibility is much higher. For example, this model assumes each user lost 10 minutes per week the first year, 5 minutes per week the second year, and 1 minute per week the third year. Many users report having lost three or more days in a week on multiple occasions to resolve incompatibility problems. Similarly, although this model is based on the estimation that 99% of users have upgraded by the end of the third year, recent surveys still suggest that many large companies (as many as 7%) still have not upgraded from Office'95 to Office'97.

Additionally, while the model indicates a lost productivity cost of roughly $43 billion dollars, this does not reflect other related costs such as retraining (which is more difficult to estimate). Retraining costs are a significant cost of using software upgrades, sometimes requiring less sophisticated users to take new training courses. To include training costs, we would need to include course costs, training time costs, and lost productivity costs. Since the total cost of training includes course training costs, training time, and lost opportunity costs, the actual costs could total several times $43 billion dollars.

End users lose this time and money in business and at home. To avoid or solve the incompatibility problems, users are often forced to buy the latest release of a product. One would hope that such incompatibility problems are never intentional. Yet upgrades provide substantial revenues for vendors. So, from a vendor's point of view, incompatibility and rapid planned obsolescence are both ways to drive customers to upgrade.

Although in this research report, Microsoft Office is used as the example for modeling costs of planned obsolescence, similar analysis of any other commonly used software would yield similar results. For example, Lotus SmartSuite also produces documents for business communications. A Lotus representative reported that at least 25,000,000 copies of Lotus SmartSuite had been sold by late 1999.[11] Lotus claims that the Lotus SmartSuite market share is 10%, which would mean at least 250 million office automation suites total have been sold in recent years, including Microsoft Office and Corel's PerfectOffice. Another commonly used software, Lotus Notes, has sold at least 40 million copies total, with an annual sales "run rate" of at least 10-15 million more per year.

Taken cumulatively, the numbers we input to test the economic model above are very conservative. The model was developed primarily to show how to establish some minimum costs of rapid planned obsolescence to society. The total number of word processors sold is probably in the many hundreds of millions, so the actual costs are undoubtedly larger than those given here. When the other types of software packages are counted, the costs of rapid planned obsolescence could skyrocket to trillions of dollars by the Year 2020.

[11] Collected from a Lotus Worldwide Public Relations representative on 12/15/1999.

Many software proponents would argue that instead of focusing on lost productivity and lost opportunity costs, users should really focus on the productivity gains possible from having more powerful office automation tools. Productivity gains from these automated tools are unquestionably great. It is possible to develop models to calculate financial benefits gained from improved office automation tools. Indeed, such economic models were developed years ago for word processors and spreadsheets, and have been relegated to the annals of business history No one would never think of going back to manual typewriters, scissors and tape, and correction fluids instead of automated word processing tools. The models and facts given here are meant to inspire critical consideration of business practices that routinely thrive on rapid planned obsolescence. It is important to expose some of the actual costs of planned obsolescence. There should be more and better automation -- not less -- with planned preservation available as a standard option without extreme additional cost.

10 Breaking The Status Quo: Giving Data Control Back to Its Rightful Owners

While moving data between two different versions of the same software can present migration issues, this problem pales when compared to the problems encountered when data must be moved between different products or applications. An outdated mentality persists among software vendors, that of "keeping customers by holding their data hostage." In order to create and maintain captive market share, vendors frequently define their products as the center of the data universe. In a capitalist economy, this is one way to create a successful business. Software vendors whose products are not the center of the data universe strive to compete and re-center the data universe around their own products. One way for a vendor to achieve or maintain this goal is to make it easy to import data into their software, but make it difficult to export complex data to other systems. Data export and external access to important system-level data structures such as file structures and database indexes are usually left as an "afterthought" for the next major software version (if they are added at all). However, more and more businesses increasingly rely upon many different systems to accomplish their business processes and goals. Thus, standards for data sharing and access across system space and time boundaries are becoming urgent priorities.

Obviously marketing reasons exist to prevent sharing data too easily. If data may be easily exported, then customers can more easily migrate to competitors' products. For this reason, vendors rarely dedicate their limited resources to solving export problems. Instead, niche vendors and integrators respond to these customers by providing costly custom data import and export and conversion utilities. This creates another industry that provides a continuing stream of revenue for smaller vendors and integrators as the larger vendors introduce regular upgrades and enhancements. Because most of these data integration and conversion solutions are not simple for end users to deploy or use, users must pay highly-trained, usually expensive, consultants to integrate their applications and ensure that users understand how to use them. The result is a set of endemic software industry practices that ensure that while the data created, manipulated, and analyzed by computer software belongs to the consumers, the long term access to that data is controlled by the vendors who sell the software (thereby locking up the customer base).

These problems are enhanced by the fact that vendors regularly leave the marketplace in one way or another: businesses fail, old product lines need to be retired, or a new product line is created without a practical and easy way to migrate from the old to the new product. Other vendors, in an effort to grow powerful in the marketplace, produce newer, more innovative systems before data is migrated from existing systems. Then they will provide easy import capabilities, with poor export

capabilities. In these cases, time and money must be spent developing custom application and data migration applications, which are quickly out of date, and may only be useful for one migration. The problem is increasingly serious because we as a society are increasingly dependent on computer systems that fail to address major record keeping issues.

The alternative today is if you want records that you create and own to be accessible long-term without continual expense, it is your financial burden to design or buy your own system to make your data system-independent, even though it was your data to begin with. However, by taking the records out of proprietary formats, the integrity of the data is frequently lost, so the risk is high that you could lose the purpose of your investment anyway.

Historically, it is easy to see the trends that led to these practices. Enormous amounts of time, money, and expertise were necessary simply to solve the technical problems standing in the way of meeting basic business needs. Today, many of the original problems have either been solved or new tools exist to help solve existing problems more quickly. A more effective way of building and keeping a large consumer base is to begin to focus on the grander needs of consumers, including long-term data access. Vendors can develop successful marketing strategies around providing the best customer quality and value, rather than proprietary tricks to hold customer data hostage.

With better infrastructure technology standards, data access will get easier and more open. The World Wide Web has the potential to revolutionize attitudes and behaviors about data ownership more then anything else in history. Many infrastructure standards exist such as the eXtensible Markup Language (XML), Java, JavaBeans, Lightweight Directory Access Protocol (LDAP), Common Object Request Broker Architecture (CORBA) services, and the proprietary Microsoft standards Common Object Model (COM) and ActiveX, MAPI, Open Database Connectivity (ODBC), and others. The problem is that these protocols change very quickly relative to the time scales of large organizations.

With the newer protocols, such as XML and Java, the "Browser Wars" are still raging. The Browser Wars are where vendors such as NetScape, Microsoft, Oracle, IBM, and others have created their own proprietary extensions to the HTML and Java languages ostensibly to enhance them. The net results have been improved protocols, with minor incompatibilities that make users' lives difficult, make website developers' lives miserable, and make HTML and Java far too unstable to be suitable for long-term data preservation and access applications.

Vendors' most often-cited reasons for not supporting particular standards are:

1) the cost (It takes scarce time and resources away from competitive features to define and implement standards. Standards are always a cost center, never a profit center for vendors.),
2) poor current acceptance (no one wants to be first to say goodbye to their proprietary technology),
3) short lifetime of most standards (especially if no one really commits to them!),
4) the need to "enhance" standards ("our ideas are better" or "there were always some unacceptable compromises made"),
5) the Not-Invented-Here syndrome (everybody wants to be the standard, nobody wants to follow someone else's standard.)

These reasons are sometimes legitimate; frequently they are simply competitive smokescreens. The persistence of these reasons put the onus on large corporate customers to push vendors to support existing and emerging standards for long periods of time. Buyers must finally make design for long-term preservation and access and interoperability into high-priority, baseline requirements for all new IT products for electronic records or knowledge management. Part of this Call To Action is for large organizations -- government and industrial -- to push for standards by "voting with your feet" to stop purchasing IT products that do not support long-term preservation and access of electronic data and records in application-neutral, open and published formats.

11 Avoiding the Icebergs: The Itinerary Has Been Changed

The Titanic 2020 problem is likely to get much worse before it gets better. Universities do not routinely offer courses or curricula on how to design digital systems for long-term preservation and access to data or records.[12] An entire generation of computer science and engineering professionals needs to be taught basic designs for preservation, not just innovation at the expense of history. Some records are too valuable to entrust to today's computer systems. Commercial products sadly address only a "long-term" horizon of a few years. They most certainly do not include built-in capabilities for their own replacement. The mind shift called "paper minds to electronic records" must still take place to move us all the way into the Information Age.

Practicing information technologists must develop a fuller appreciation of the disciplines and practices of record managers, archival scientists, and library and information scientists. These professionals are the guardians of corporate accountability. They are also important gatekeepers for intellectual capital, and levers for getting the most out of it. Information technologists must design acceptable technology solutions that meet the needs of these communities of business professionals.

There is an exploding need for universal digital archiving systems and formats. Universal formats for each data type (e.g., motion pictures, audio, digital certificates) are badly needed by many communities. The World Wide Web Consortium[13] is sponsoring work to harmonize semantics for web-based information that will apply generally to as many communities as possible. However, competitive wars are expected to stretch out real semantic unification for information for many more years. In the mean time, we badly need universal information container formats that allow high-quality publishing of collections of dynamic information collected from many sources. An electronic legal record is compilation of data that is recorded, signed, and sealed in an immutable form, typically in the form of documentary evidence. Thus, one of the greatest needs for electronic records has been format standards for publishing compound documents that can include text, graphics, numerical data, tables, etc.

The storage format that comes closest to a universal standard for high-quality publishing is Adobe System's Acrobat Portable Document Format (PDF).[14] PDF is a "container format" that permits capture, publishing, sharing, and preservation of complex compound documents with embedded

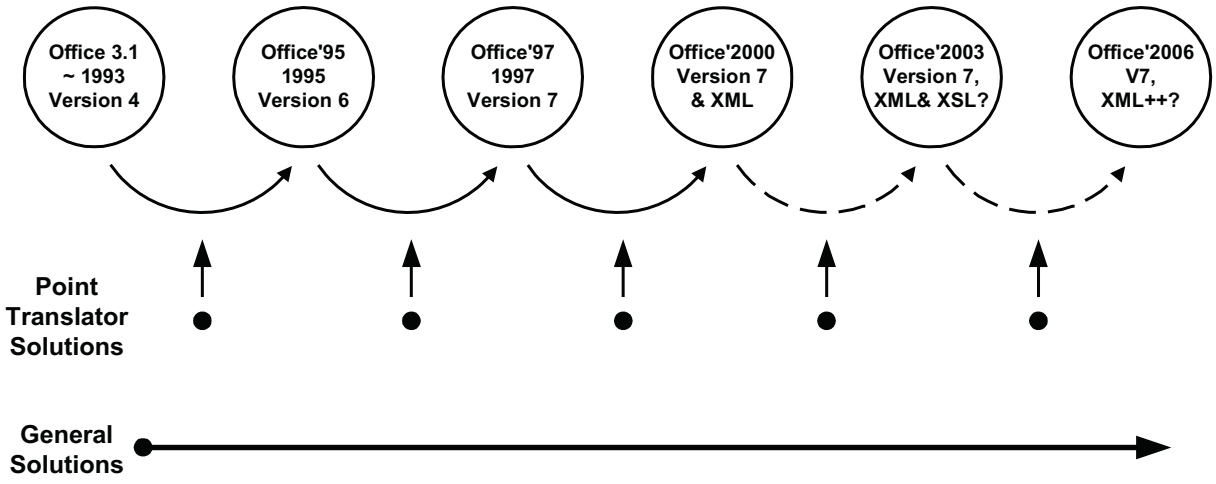
[12] Research and codification of the rules of system design and engineering for long-term preservation and access to electronic records are underway now in an international collaborative research project called the InterPARES Project. Courses on this subject are being developed and taught in a few leading universities and industry now. See www.censa.org/industry-interpares, www.interpares.org, or www.is.gseis.ucla.edu/us-interpares for more information. It will be a few years before these principles are routinely embedded in higher education.

[13] See www.w3c.org for more information on the World Wide Web Consortium.

[14] For more information on Acrobat and PDF, see www.adobe.com/products/acrobat/adobepdf.html.

digital audio, video, and other dynamic data types. PDF permits communication of these published documents across all major types of computer platforms (MS-Windows, Apple Macintosh OS, and many flavors of UNIX, including Linux) and across time. Although it does not work for all types of data, it is good enough for archiving of many types of high-quality digital documents and data. Extensions to PDF or other products still need to be developed to support permanent archiving of large commercial databases, spreadsheets, and other specialized scientific and technical data sets.

An explanation of the terms "forward compatibility" and "backward compatibility" will help to clarify the following discussion. "Forward" or "backward" always means relative to a software version or to a file format version. Backward compatibility of software means a newer software version can read and process files created by an older software version. Forward compatibility of software means an older software version can read and process files created by a newer software version. People most often desire backward compatibility of the software to read and process older files, since the software is continually updated. Forward compatibility of the software is less frequently needed, only by people who fall behind in software versions, e.g., trying to read an MS-Word'97 file with MS-Word Version 7.0 ("Word'95).



The diagram above illustrates the difference between long-term, general solutions and "point solutions" in time. Point solutions rely on translators to convert from an older version to a newer version of a file format. MS-Office includes file converters to translate nearly all of its older file formats into the latest file format, and vice versa, obviously with a loss of features that did not exist in the older formats.

General file format solutions can be read directly with any newer software versions, without using a translator, and if they are high-quality general solutions, without generating errors. General solutions are more reliable than version-to-version file translators, because version-to-version incompatibilities can cause critical information to be lost, especially embedded graphics and links, table formatting, and other software version-specific details. However, general solutions are much harder to create, because IT changes so fast and competition is so fierce. More important, it requires a tremendous amount of foresight and creativity to create products that evolve gracefully in such a fast-moving world, plus a firm business commitment to long-term compatibility. Relative to lower-tech industries, the IT industry per se is very young, immature, unstable, and competitive.

This makes it difficult to create general solutions with high-level of fidelity and interoperability that the leading IT vendors agree to support for long periods of time. With a few notable exceptions, vendors have perceived file and data formats as a competitive weapon to use against competitors and customers.

Adobe Systems (San Jose, CA) is an example of a company with a truly unique philosophy and commitment regarding long-term product compatibility. Adobe has a broad and public commitment to provide 100% backward compatibility of the PDF format *for at least the next 25 years*. No other software vendor has such a forward-looking philosophy and commitment to its customers. An important business reason for this commitment is that Adobe is beholden to the publishing industry, which has standardized on Adobe's PostScript language and many of Adobe's other multimedia publishing products. PDF is a new and improved but compatible version of PostScript, that also supports high quality publishing on the World Wide Web. The publishing industry simply cannot afford to retool every few years. Publishers must be able to access, view, and print their archived documents and books on demand for many decades to come.

The publishing industry's practical need for a general solution that has stability and backward compatibility was recognized early by the United States Internal Revenue Service, the Food and Drug Administration, the Environmental Protection Agency, the Patent and Trademark Office, and many other government agencies. *More than 140 government agencies worldwide have adopted PDF as their standard for document submission and archiving.* Many industries have followed the lead of these government bodies and adopted PDF, primarily because it makes it easier to do business internally and with the government.

The Collaborative Electronic Notebook Systems Association and its end user member companies, which include several large pharmaceutical corporations, adopted PDF in 1997 as the preferred format for document communication and publishing and permanent storage of electronic records for three specific reasons:

1) PDF's is the defacto government standard for document publishing and archiving;
2) PDF's high quality rendering capabilities for multimedia publications; and
3) Adobe's long-term commitment and long track record for successful and complete backward compatibility, going all the way back to the version 1.0 of PostScript released in 1984.

The third point is the most important from a practical point of view. PostScript printers of 2000 can still faithfully render files created in PostScript Version 1.0 format. This track record of 15 years speaks volumes in a world where Web years are measured in weeks or months on the calendar.

With extensions to support XML, PDF can apply to all types of data published on the web and elsewhere, and for exchange and archiving of more specialized scientific and technical data types. The most recent release of PDF added security features to support signing and witnessing of PDF documents, allowing users to create self-contained, portable electronic records. Finally, an army of third-party software developers is using the open and published PDF format to extend and apply PDF to different applications and niche markets. PDF has a long future ahead of it.

12 The Die Has Been Cast for 21st Century Information Technology

Adobe has proven to business leaders, IT professionals, and governments worldwide that stability by design, coupled with rapid innovation are not only possible, but profitable too. Most important, with Adobe's commitment to long-term stability and support, the die has been cast for end user's expectations of the IT industry in the 21st Century. This report does not mean to single out Adobe for benevolence or philanthropy -- clearly the company has benefited economically from its long-term commitments. In every age of technology there have been companies that have built fortunes by getting into new areas of business early. In the near future, suppliers who commit to stable, flexible *and* open technology designs for electronic records and data systems will have a huge competitive edge. They will profit over those who choose to maintain a perpetual blind eye to the need for stability in at least one part of their technology design -- the data or records owned by their customers. In the future, the stability of digital records, whether corporate, government, or personal, must be measured in decades to centuries rather than web years.

Innovation does not have to come at the expense of the integrity of our data or records over time. Information technologists are among the smartest and most creative people on the planet, clearly capable of creating flexible technologies that are extensible far into the future *and* backward compatible. Stability, flexibility, extensibility, and backward-compatibility are not opposing or incompatible requirements but simply additional business constraints on present and future electronic record and data storage technologies, products, and systems.

13 The Ship Has Left the Dock

The Titanic was designed to be the biggest, most technologically advanced ship of its time, and indeed it was. No one would argue that the ship failed to meet the expectations of the passengers who paid enormous sums of money to stay in her staterooms. But the fact that the ship appeared to set a new standard against which all other ships would be measured inspired too much confidence in it creators. They did not believe it necessary to make sufficient provisions for the safety of their passengers. Of all the questions about the Titanic's tragic demise, a particularly haunting one stands out: would the passengers have booked passage if they had known the boat had no sufficient exit strategy?

Like the passengers of the Titanic, it is easy for users of computer systems to believe in the marketing messages of vendors and become overconfident in the safety of their data and records. However, experience has given most modern users the chance to understand what passengers of the Titanic could not, that within the current, albeit often-impressive solutions lie dangerous pitfalls that cannot be ignored. Perhaps the market dynamic that once allowed Red Star Line to create the Titanic and now allows computer businesses to create systems used by millions of users shouldn't be broken. But the disaster that visibly looms in our future cannot be ignored. If the Titanic had had enough lifeboats for all her passengers, she would perhaps still be remembered as the biggest and most advanced ship of her time. She would not have become symbolic of the inability of man to recognize and plan for his own fallibility. Like the Titanic, current computer systems are in part a result of man's unwillingness to recognize and address the shortcomings of an industry that has mushroomed both technologically and financially. If this problem is not addressed now, in approximately 20 years, the Titanic may be replaced with a new icon of disaster.

14 Background on CENSA

CENSA, the Collaborative Electronic Notebook Systems Association, is a market development association focused on catalyzing markets for advanced automation for research and development organizations. CENSA uses unique market development programs that involve end users, suppliers, and government agencies working together to identify, specify and create solutions to market problems faster than normal methods. With over 800,000 employees in CENSA member organizations, CENSA has a broad impact across many sectors, including software, high-tech, consumer products, chemical, pharmaceutical, biotech, healthcare, government and many others. CENSA emphasizes regulated industries where intellectual property protection is key to success.

CENSA's current market development programs focus on systems for long-term preservation and access of electronic data and records, electronic notebook authoring applications, architectures and tools for component-based application and systems integration, and high-priority scientific, technical, and business components. CENSA's deliverables to its members are specifications for complete systems, component software and hardware technologies and products, and extensive specialized knowledge for setting up and using the systems and tools successfully.

CENSA sponsors the Global Industry Interagency Group on Electronic Recordkeeping and Digital Archiving Systems (GIIG on ERDAS), a non-profit organization of government agencies and industries dedicated to defining and implementing acceptable systems for long-term preservation and access of electronic data and records in industry and government applications.

15 Background on the Authors

Rich Lysakowski, Executive Director, CENSA

Rich Lysakowski is the Executive Director of the Collaborative Electronic Notebook Systems Association (CENSA). He has 20 years experience in various scientific, engineering, marketing, and project management roles in the private and public sectors. He has a Ph.D. in Physical and Analytical Chemistry, with a specialization in R&D and laboratory automation. He has done over one hundred presentations at conferences and workshops, authored over 30 publications, and edited two books. He also teaches regularly for the American Chemical Society and other professional organizations. He holds two patents on software technologies.

Zahava Leibowitz, M.A., ABD, Researcher and Writer, CENSA

Zahava Leibowitz is a researcher and writer for CENSA. She specializes in medical and technical writing, technology research, regulatory compliance and validation issues and documentation, document management systems, and project management. She has numerous publications to her name, and has been the senior editor for two books. She has over ten years experience in teaching writing and editorial strategies.

16 For More Information

For more information on CENSA and solutions to these problems, please contact:

CENSA, 800 West Cummings Park, Suite 5400, Woburn, MA 01801 USA
Phone: 781-935-9600; fax: 781-935-3113; E-mail: info@censa.org; website: www.censa.org.