Following the Trail of the Disappearing Data

By Victoria McCargar

Since the mid-1990s, it has become increasingly clear that information stored digitally — unlike physical photos, for example — is unnervingly fragile. Lacking the appropriate systems, workflows and metadata to ensure longevity, news archives are setting the stage for future data loss.

Sitting on my desk is a black-and-white aerial photograph looking up Pasadena's Arroyo Seco at the Rose Bowl on a sparkling winter day. The picture is in very good condition, the emulsion intact, with a couple of minor wrinkles and a mark or two from an orange grease pencil. I can see on the back the carefully applied caption from the day it ran in the *Los Angeles Times*, Jan. 1, 1935 (Alabama beat Stanford, 29-13). This picture looks like it's good to go for at least another 70 years.

On the monitor of my Macintosh G4, I have a JPEG from Mullaittivu, Sri Lanka, from Jan. 1, 2005. It's an arresting image of the forearm and hand of a dead woman, visual evidence of the human tragedy of the Dec. 26, 2004, tsunami. Someone in 2075 (amid global warming-induced flooding, perhaps) might want to see exactly what Mullaittivu looked like on this day. Will he or she be able to pull up this 200-dpi, 584KB nugget of disaster history 70 years from now?

Don't bet on it just yet.

Since the mid-1990s, it has become increasingly clear that information stored digitally is terribly fragile. Newspapers periodically run stories about this phenomenon and give good coverage to heroic data rescue efforts, such as the British project to salvage the Digital Domesday Book, or conundrums, like the difficulties museums are having curating digital works of art. But there appears to be a mysterious disconnect when it comes to another group with an important cultural stake in long-term preservation: newspaper archives.

Research on a global scale is under way to find solutions to preserving born-digital content, but it's a field limited almost exclusively to academic and research libraries, national archives and bureaucratic record keepers — professionals invested with a defined responsibility to keep digital files alive and accessible for a long time.

So it is ironic that even as they're publishing stories about data fragility, newspapers haven't quite made the connection with what is going on in their own electronic morgues. (I refer throughout to newspaper archives, but in fact the same issues affect other news media collections as well — for that matter, any data collection that is supposed to last indefinitely.)

The fact is, photo and multimedia databases, and even text databases are potentially shorter-lived than yellowing newsprint, and some formats in use today will ultimately prove more unstable than chemical color photography. Indeed, the very technologies that have enabled the rapid dissemination of news are conspiring to create a generation-size gap in the historic record.

Only 1s and 0s

Digital data is basically a collection of on-off switches, strings of 1s and 0s (bits) ordered in manageable chunks called bytes. In simplest terms, what differentiates the million bytes of a 1MB JPEG from the million bytes of a 1MB spreadsheet is how the bytes are interpreted by which application. But other factors besides software determine the future accessibility and readability of the 1s and 0s: platform and operating system, storage structure, technical metadata, content description, copyright and even (maybe especially) institutional discipline. Over time, sometimes catastrophically quickly but more likely gradually, a byte stream will tend to become unreadable, essentially reverting to the magnetic on-off switches of storage media, the 1s and 0s.

The task of identifying all the risk factors and putting preservation solutions in place has barely begun. In the meantime, lacking the appropriate systems, workflows and metadata to ensure longevity, news archives are setting the stage for future data loss. It's not too much of a stretch to say that byte streams that have been stored for the past 10 years — and those that will be captured and stored tonight or next week — might already be lost.

It's not hard to see how this happened.

Lured by speed, unprecedented accessibility and flexibility, not to mention gains in staff productivity, publishers and their newsrooms have embraced technologies that enable a wealth of functions: easily captured, edited and transmitted photography, full-page pagination, Web publishing, content sharing and repurposing, and PDF workflows, to name the big ones.

Over in the news library, meanwhile, huge gains in storage density and processing power meant that big, increasingly sophisticated image databases or burgeoning collections of images on CD-ROMs have relegated black-and-white prints in envelopes to the back of the stacks. "Archives" have morphed into "assets," and assets have come to refer to a variety of formats beyond photography and text. Information graphics, analytical databases, HTML pages and digital video all have all become part of the potential multimedia archival mix.

As technology has come to play a larger role in the news archives, responsibility for maintaining content has in many cases been transferred from traditional archivists and librarians to systems analysts. At the same time, the automatic capture of bibliographic and descriptive metadata from the publishing system has resulted, not surprisingly, in heavily downsized archives and library staffs. This is a major shift in information management philosophy, because IT departments arguably have a different approach than libraries to long-term preservation.

Budgeting for Preservation

Archives consisting of envelopes of old clippings and black-and-white photographs didn't require large capital outlays every few years to sustain them; as long as they were protected from dangers such as fire and water, and kept in a reasonably controlled environment, they could survive almost indefinitely.

Digital data is very different, primarily because it doesn't respond well to that kind of benign neglect. To forget about a few envelopes of CD-ROMs in a file drawer for 10 or 15 years is asking to lose them; to skip a couple of upgrades is to put an entire format at risk.

The problem with funding archives, moreover, is that it's difficult for budgeters to see a return on investment. While digital preservation costs are still mostly a matter of speculation, most researchers agree that it will be expensive. True, some news archives generate a modest revenue stream from reselling old images and articles in new digital forms, but beyond that, publishers and chief financial officers aren't necessarily willing to spend money to meet some vaguely perceived obligation to maintain a record of history in the making.

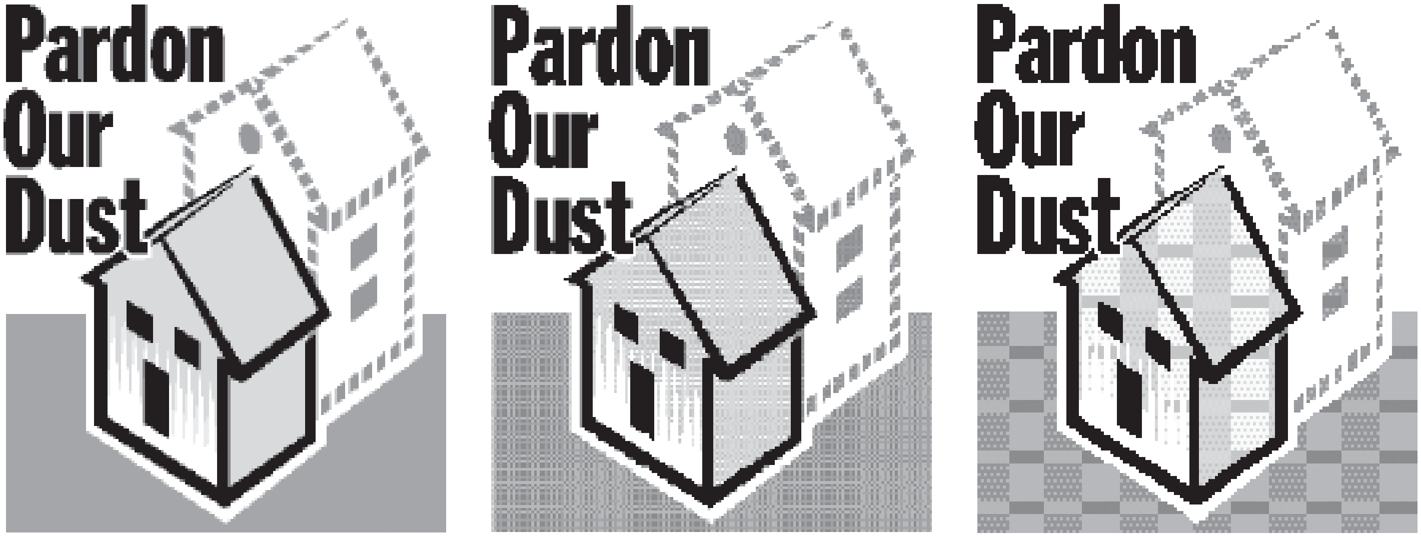
Surviving Space and Time

Digital archives exist in a physical world and are subject to equipment failures, such as burst pipes and the like. Properly backed up, the data will survive physical dangers and be restored. But digital preservation does not equate with disaster recovery — a misconception that IT professionals often have. The threats I'm concerned with here are much more subtle, amounting to the gradual loss of information through a variety of changes over time.

Software obsolescence. This is such a seemingly ordinary problem that it's tempting to think that it really isn't one at all. If systems administrators are careful enough to make every upgrade on schedule, the objects will migrate naturally to the next version, or so the thinking goes. But batch migration of thousands or millions of individual objects from one version to the next is not common practice. The typical workflow is to leave an object in its original version until a user needs it for some new purpose.

But what if a user retrieves the object created in version N, and the only available software in-house is version N+5? Backward compatibility will never be unlimited, and the nature of forward migration is to introduce errors with every upgrade, however minute or undetectable. Even with well-executed batch migration, over time those errors are cumulative and the data gradually becomes unreadable (see illustration).

That assumes the software continues to exist and function. WordStar, a nearly ubiquitous word processor in the 1970s and 1980s, is often held up as the poster child for digital obsolescence. No current word processing programs will open a WordStar file, and the



Even with well-executed batch migration, over time errors are cumulative and data gradually becomes unreadable.

company stopped manufacturing the software in 1991. Cracking old WordStar files now amounts to a hobby for computer enthusiasts.

Hardware obsolescence. Every new data storage format signals the end of its predecessors, be they Zip disks putting an end to 3.5-inch floppies or EVDs (enhanced versatile disks) putting users on notice that there's a format beyond DVD. While it's true that few people do any serious archiving on Zips (or their successor, memory sticks), many news archives have consigned their photography to CD-ROMs, and they're now looking at having to shift to DVDs. Inasmuch as CDs are turning out to be subject to more physical deterioration sooner than thought, having to reformat on the more stable DVD platform is probably a good thing. But it's still a moving target.

If photographs are stored in large databases with industrial-strength hard disks and tape-drive backups, the material is easier to move forward than collections of disks.

Inadequate metadata. In a January 1995 *Scientific American* article, RAND Corp. researcher Jeffrey Rothenberg pointed out that if modern civilization is going to hang onto digital information into the future, its denizens are going to have to create a lot of other information about the information to go with it, to enable future seekers to write new software to "bootstrap" their way into rendering the obsolete data into some form that humans can read. That information about the information, or metadata, is critical to the preservation process — probably a great deal more important than software or hardware, in fact. Much of the research agenda in data preservation focuses on what that metadata should comprise.

In his article, Rothenberg proposed that the information include, minimally, specifications about hardware, operating system and software requirements; byte-stream interpretation, and enough information about the software code itself to allow a future user crack it — essentially a digital Rosetta Stone.

That so-called technical metadata is in addition to the more familiar content and context metadata: the journalistic who, what, where, when, and why of good caption-writing; bibliographic data such as date of publication, section, edition, part and page; and enough information about the copyright status of the object to ensure that future users know what their access rights are.

Some of this metadata can be captured or generated automatically, but a lot of it cannot, and producing it will not be inexpensive. Assigning index terms according to a controlled vocabulary, sometimes known as keywording or taxonomy, is a good example of this. As much art as science, good indexing provides ways to limit searches and zero in on the subject of an article or image, saving the user from looking at a lot of irrelevant material.

As multimedia databases grow and become more complex, smart metadata will make the difference between a useable database and one that merely contains objects. If an object can't be searched for, found, retrieved and used, it is as good as lost. As brilliant as Google is, simple free-text searching isn't up to the kind of sophisticated searching that news users need. No one will want to slog through a Google-scaled 10,000 or 20,000 hits in his or her own multimedia database.

And just because an object is never retrieved doesn't mean it doesn't still reside in the database. Over time, systems analysts and budget writers will find themselves supporting — and financing — a larger and larger chunk of this "dark" data.

Lack of standards and best practices. Preservation researchers agree that tight standards are key to solving the data longevity problem. The academic and research library and archives worlds, which have been grappling with the digital preservation problem for most of a decade, are coming at it from a foundation of fairly rigid standards for digital data structures and description, beginning with MARC (machine-aided cataloging) in the 1960s, and proceeding through today's emerging standards like MIX (technical metadata for still images in XML) and METS (metadata encoding and transmission standard). They are, consequently, well prepared to begin adding preservation metadata to their institutional workflows as standards begin to take final shape in the next few years.

News archives practice has developed in response to the deadline demands of news research and, more recently, the requirements of repurposing material for the Web and other products, including sharing content with sibling properties. One-off systems and local customization are gradually giving way to discussions of ways to interoperate, developing best-practices workflows not just within a single news organization, but within a corporate chain. The venerable IPTC (International Press Telecommunications) "header" is a logical place to start talking about standards for preservation, but eventual solutions will come at the expense of flexibility and the latitude to customize.

Lack of institutional discipline. Customization has usually been born of necessity. Meeting production deadlines and the "get the paper out at any cost" mentality that is the hallmark of working in a newsroom tend to produce some really creative workflows. However, in the automated capture and processing of metadata, spot innovations and one-off workarounds can play havoc with the digital record.

Best practices for digital archiving suggest that the process actually begins with the photographer or reporter and continues through the entire editing process. But the burdens and requirements of well-formed metadata are way beyond what can reasonably be expected of shooters, wordsmiths and artists. On the archives end, the only way to guarantee the compliance of the record is a set of quality controls, which are usually humans drawing a salary and benefits. Without them, the resulting record is basically an anomaly and, over time, subject to becoming invisible to a future search engine.

Moreover, any current and future efforts to develop digital preservation solutions will be aimed at solving a standardized problem — developing a uniform migration path for JPEGs to a future format like JPEG2000, for example. If an individual news archive isn't IPTC-compliant, is using a slightly different version of JPEG or has incomplete technical metadata because of one of a dozen possible user workarounds, the standard "rescue" solution might pass it by.

XML is frequently mentioned as a preservation solution because of its platform independence and highly intuitive, self-describing tag-sets. XML in theory and XML in practical application are quite different, however, and the rigid workflows required for well-formed XML are hard to come by in most newsrooms, especially at the design desks, where a lot of last-minute changes take place. When deadline performance is at stake, the creative workaround will trump the compliant workflow every time.

Copyright. It's not a technological problem, but it's almost as big a threat as obsolescence and could turn out to be even harder to solve. In the fallout from the Supreme Court's 2001 *Tasini v. New York Times* decision over the rights of freelancers, large parts of news archives disappeared from their host databases, either moved offline or deleted outright. As digital copyright continues to evolve, archive managers are struggling with how to handle freelance material, for which in many cases archiving is *verboten*.

How the Government Saves Its Assets

How do you build a system for long-term preservation? Even though many of the potential tools and processes are still theoretical, there is an immediate need for systems that will retain data for long periods. Currently, the public sector is leading the charge.

TranTech Inc. is a U.S. government contractor that works with federal agencies and technology companies to build systems that meet federal standards for data preservation, among other requirements. The Seybold Report interviewed Mark Wells, TranTech's technology director, by phone from the company's headquarters in Alexandria, Va.

The Seybold Report: *Talk about TranTech and its role in digital preservation.*
Mark Wells: Our concentration is mainly in software development, database development and digital media for government agencies. When it comes to preservation, the government has its own set of requirements that our clients have to follow, especially Department of Defense (DoD) directive 5015.2 ("Design Criteria Standard for Electronic Records Management Software Applications," 2002), which says how any record management application must be built. What DoD establishes, everyone follows. Also, the U.S. National Archives has Title 36 from the Code of Federal Regulations, which tells you how to maintain historically important records and documents. There's a whole series of other regulations we deal with.

TSR: *What are the typical issues you confront when you are required to create a preservation-oriented system?*
MW: The good news is, when you do business with the government, the regulations are laid out for you: "Here they are, you will abide by them." That makes it a lot easier, because you don't have to discuss what the rules and regulations are in regard to accessioning and disposition: what you take in and what you discard. They tell you at any point in the lifecycle of a document what can be done with it. Regulations spell out what an important document is, how long to maintain it, what you do based on document type: Is it operational information, classified data or something that is historically beneficial to the public? Those considerations tend to drive how long you have to keep things — some for one year, some for three, some for seven, some for 25, *ad nauseam*. If you don't abide by the rules, you could go to jail. So having a policy is important.

[Systems developers] don't always talk to the right people. Have you talked to the archivists or record management people at the agencies, the people who are really knowledgeable? Does everyone talk to them when they're implementing a system? No. Do people build systems without talking to the right people? Absolutely.

The DoD model has built-in checks and balances, so if the system isn't right, you'll find out. Yes, you may build a system that doesn't necessarily comply with the requirements. But before you go online with it, it goes through a series of checks, and if it's wrong, you'll hear: "Wait a minute. You didn't comply. You'll have to go back to the beginning."

TSR: *Are more companies starting to develop preservation-compliant systems?*
MW: Vendors are definitely getting up to speed; it's more and more part of the IT world. The federal government has made a strategic move away from what's called GOTS (government off-the-shelf software) — one-off systems, which are hard to support — to COTS (commercial off-the-shelf software).

We're working more and more with commercial vendors. We say to them, "Here's what we have to do, here's the method to get compliance, here's what government needs." If you're a commercial vendor, for your software to be declared a "System of Record," it has to be tested to fulfill DoD 5015 compliance.

Government is good at driving standards. It likes to produce them and it has been doing it for a long time; 5015 got its start back in the 1970s. As the established policy becomes more and more accepted, other parts of government start using it. So even though 5015 started at the DoD, it is now a government-wide standard. Vendors like Documentum, Interwoven and other builders of document management systems have gone to the trouble of making their software compliant, because they want to do business with government. Commercial enterprises will start to get the benefit.

Mark Wells

TSR: *What's your advice to a company seeking to undertake long-term digital preservation?*
MW: Commercial organizations have requirements, too — laws applying to financial management, like Sarbanes-Oxley. There are also state and local requirements for asset management, which drive preservation and records-management rules. First and foremost, you need to know what laws and regulations affect what you do. Once you figure out what those are, look at your document lifecycle and ask yourself what you might start doing differently.

The big thing I say to people is, "You have to decide between preserving everything and preserving some." There are two entrenched sides to this, and the problem is, you end up with a zealous war between two factions. You need to work at finding common ground.

You also need to identify what's best practice for what you do. Here again, there are battles between two extreme groups: those who want to keep the exact original and those who don't think it's necessary. I tell people to concentrate on the "essence" of what they're preserving, to do what I call an essence study. Can I change an object from format to format to format?

Take videotape. Video-to-digital can easily transfer the essence to DVD without any real loss. Paper photos can be digitized without much loss of essence. The digitized picture is the same thing, you get the same reaction to it, it has the same bits of detail.

However, there are other issues. For example, for the National Archives, a screen capture of the Declaration of Independence doesn't necessarily capture its essence. There's much more to it than the "data." There's historical value there, which is lost if you transfer to digital. But the preservation of essence can be an enormously expensive undertaking. How far are you willing to go?

TSR: *Overall, how expensive is this going to be?*
MW: The argument is still out on cost models for preservation. Because of the factions involved, decisions are complicated. One faction can show that there's no cost to keeping digital objects. But for the faction that is so involved in preserving originality, like cultural heritage institutions, the cost can be very high.

It also comes down to the costs associated with the legal ramifications of preservation. But even those are hard to track, because government requirements can change on a whim. The Patriot Act and Sarbanes-Oxley are examples. Information that used to be thrown away now has to be maintained. We're talking about billions of dollars to change systems just because of the way a law is written. Sometimes a single word can have astronomical impact on cost. In the Patriot Act, changing an "a" to "the" cost billions of dollars. **TSR**

— By Victoria McCargar

What can a newspaper or magazine do with freelance stories and photos to archive its own published record? The answer, surprisingly, is to microfilm it with the rest of the paper.

The electronic version, on the other hand, may exist in a digital limbo, moved to the archive in an automated workflow, invisible to users, its status uncertain. And creating metadata for copyright that will be meaningful 50 or 100 years from now seems to require a rather large crystal ball.

Coping Techniques

While preservation-oriented standards, practices, users and vendors sort themselves out, there are a few seat-of-the-pants techniques that work fairly well, as long as alert people in the organization stay on top of the content they're trying to keep. None, however, is more than a short-term, stop-gap method. At this point, that's simply all there is.

Migration on demand. Files are upgraded piecemeal as the need for one in the newer version arises. Unneeded files remain in the old version indefinitely. The migration process also necessitates accounting for the transfer of all the metadata, which might exist in a separate format, while retaining all its connections to the original object if the metadata is not contained, or "encapsulated," with the object. A thorough, well-documented testing program is essential before undertaking a larger-scale migration, and careful documentation is necessary for future users to understand the outcomes of successive migrations.

Technology preservation. This involves keeping one or more older computers running and maintaining the software versions that require older machines. Files that can't be migrated are stored here, too. This is actually a fairly good, inexpensive approach, as long as the machines are in working order or can be repaired if damage occurs. It's not a viable solution beyond a few years, though. Similarly, the files might not be formally backed up anywhere, meaning a system crash is potentially the end of the data.

Normalization. This refers to saving the object in a single format that is easier to preserve. In practice, this can mean exporting files to flat ASCII or even printing everything out on paper (popular for e-mail). The development of the so-called "archival" PDF, known as PDF/A, is another example of this approach, one that aims to extend "normalization" to any system in any institution. Loss of functionality of the original document is an obvious drawback, and there are further issues of how to authenticate the "original" if that is a consideration. (For example, a PDF of a freelance contract, which is a legal document, will require a fairly sophisticated method of authenticating the signatures — yet another bit of software that will somehow have to travel with the document for the life of the contract and beyond.)

Bit-level preservation. This is a fancy term for hanging onto problem files but giving up on the ability to render them pending some future technological development. The hope is that if the data can be preserved, someone will eventually figure out a way to render it. Interestingly, systems administrators might already be doing a fair amount of bit-level preservation without knowing it, depending on how many files they're accumulating in their databases that are obsolete, can't be opened, are no longer identifiable, or lack enough metadata to support search and retrieval. Whether that mass of dark data eventually is measured in terabytes or more is a function of how comprehensive the metadata is and how thoroughly the whole asset management process has been documented.

Hard Questions

News archives have a comparatively long track record in what is now termed digital asset management. Nevertheless, it's important to remember that we're still in the early stages of trying to support digital content into the future, and what seems like a workable solution now probably won't be after a number of years. All told, media archives have about 20 years' experience with text databases and half that with large-scale digital image archives. The success or failure of successive migrations after 70 or 80 years won't be known for some time yet, at which point there will be

Standards Are on the Way, But Will They Help?

When every upgrade promises bigger, better, faster features, the word "standards" tends to provoke fear and loathing among some technologists. Standards, almost by definition, suggest the lowest common denominator — hardly an environment to foster innovation and competition. Yet standardization in a number of areas, including workflow (otherwise known as "best practices"), formats and metadata, is held out as the overall solution to the long-term management and preservation of digital information.

There are standards, and then there are standards. We refer to PDFs and Word or Excel files as "standard" formats, and indeed they are, but they are *de facto* standards, meaning they are standards ("in fact") only as long as Adobe and Microsoft choose not to change them. For example, the increasingly full-featured PDF, which allows such bells and whistles as embedded scripts or moving images, is behind the ongoing effort to create a standard, "archival" PDF, called PDF/A. This simple format — relatively speaking, the digital equivalent of paper — is currently in the review and balloting process toward becoming a *de jure* ("by legal right") standard, one that is determined by an international standard-setting body and can't be changed without deliberation and a vote by the group. That sort of rigid standard will help determine the future sustainability of digital objects, because standard preservation solutions will follow. There is much more risk where standards are absent or insufficient.

One area where standardization is undergoing close scrutiny is the development of metadata for preservation. For the past decade or so, institutions of all kinds, from newspapers to libraries, have rushed to digitize their collections. That has been followed by a similar rush to develop metadata suites designed to enhance search and retrieval, as well as long-term access. Over the past few years, attention has been zeroing in on what is called technical metadata: information about a digital object that would allow its contents to be retrieved and understood even if the original software and operating system are long gone. While researchers concur that capturing technical metadata is critically important, it remains an expensive, largely manual process. Moreover, there are so many domain-specific "standard" approaches to what constitutes technical metadata that it summons to mind the old joke, "The great thing about standards is there are so many of them."

There is one widely watched project to create a core metadata standard for long-term sustainability. Known as PREMIS, for Preservation Metadata Implementation Strategies, the project (under the auspices of the Research Libraries Group and Online Computer and Library Center of Dublin, Ohio, which brought us the

Dublin Core standard) is developing a series of "semantic units" that describe certain characteristics of a digital object and who created it, all of which are deemed necessary to resurrect that object in the future.

The goal is a core of metadata that stands independent of domain, format, hardware, type of enterprise or how the system is implemented. More than 18 months of intensive work has gone into developing this core suite, which has reached the final draft stage and should be released this spring in the form of a data model and data dictionary explaining all the semantic units. The ultimate translation of these into XML tags will enable the PREMIS metadata to be nested into a larger, more domain-specific, de jure standard such as NISO Z39.87, Technical Metadata for Digital Still Images or others that have been developed — or will be soon.

With the proliferation of approaches to data management, accessibility and sustainability, tracking the various solutions and standards has become a research enterprise unto itself. Enter the "registry" concept, which, in theory anyway, assigns to a third party the task of keeping all of this confusing information straight. Two current registry projects are worth mentioning.

InterPARES (short for International Research in Permanent, Authentic Records in Electronic Systems) is an international research team based at the University of British Columbia in Vancouver that is undertaking the cataloging of all the extant and emerging metadata schemas that in some way touch on the life cycle of digital records. The registry is intended initially to begin sorting out all the overlap and discrepancies among various metadata strategies, looking for commonalities that might suggest a more systematic approach to metadata development. Once the registry is fully operational, an institution or enterprise looking for the optimal schema for its asset or archives systems could search the registry for the appropriate solution.

Meanwhile, the national archives of the United Kingdom has mounted an effort to develop a registry of file formats and their technical requirements, including formats both current and obsolete. Online since February 2004, the PRONOM registry has accumulated an initial database of almost 600 types of file formats, 250 software products and about 100 vendors. Companies such as Microsoft and Adobe have contributed information about their formats to the registry, and an online submission form encourages participation from others.

A registry such as PRONOM offers several advantages. Rather than populating dozens of fields of technical metadata in a preservation scheme, one field might simply hyperlink to the same metadata at the registry. Further, PRONOM researchers hope to offer testing and information about data migration paths, and which formats are facing imminent obsolescence. A major challenge for the project is pursuading software companies to divulge enough useful information in their code to support preservation activity, something they're understandably reluctant to do.

Code is proprietary material, even if it is obsolete.

— Victoria McCargar **TSR**

no analog original, such as film negatives or prints, to fall back on.

While solutions evolve, news archivists should be asking themselves a few questions that will go a long way toward putting solutions in place, once they emerge, in an ongoing dialog among IT, news librarians and journalists about the process of archiving.

What are we archiving? In the days of shelves and manila envelopes, limits on archives were a function of space, and it was obvious that periodic decisions had to be made about what to discard. One of the interesting developments of the Digital Age is the gradual abandonment of archival policies, written or otherwise, that spelled out what was going to be kept permanently, what was to be kept temporarily and for how long, and what was to be "de-accessioned" outright. Creators and archivists didn't always see eye to eye on the policies, though, so it's not surprising that as technology improved, creators began asking archivists to take in more material than ever before, whether or not they were equipped to handle it.

From a human standpoint, one of the great things about digital storage is that it's compact, convenient and, unlike bulging shelves, out of sight. But the bottomless accumulation of unpublished pictures from photo assignments, for example, is likely to be every bit as expensive, or more, than shelves of prints, if the intent is to keep the files viable indefinitely. And if users, archivists and IT support personnel haven't arrived at a mutual understanding of what the system requirements are, including appropriate expiration or selection strategies, the result will sooner or later be an unmanageable, minimally described mass of data weighed in terabytes or petabytes. Making policies now will save a lot of grief later.

How much is preserving digital archives going to cost? There are so many variables that preservation costs are difficult to estimate, but some researchers put it conservatively at $1 million per terabyte per decade, assuming that the institution has already developed (and paid for) all the necessary metadata analysis and creation; has seamless, reliable, ironclad workflows; and has established failsafe migration paths for all of its format types — three pretty hefty assumptions. In other words, once the expensive work of development has been accomplished, it is still not going to be as cheap as maintaining paper and emulsion in manila envelopes.

Who is going to be responsible? There is a natural partnership to be fostered among information professionals in the news library and technologists in the IT department. Hardware and software, the centerpieces of the IT approach, are only half of the equation. The rest is metadata development, standards compliance and user workflows — the domain of information professionals from libraries and archives. But the system can't succeed without buy-in from users in the newsroom, who need to be included in the development of realistic policies for long-term preservation, as well as help to promote intelligent, compliant workflows among their creative colleagues.

Responsibility extends to understanding standards and compliance, and keeping a close eye on developments in the field. An emerging body of literature about preservation metadata will eventually influence standards, XML schemas and, in turn, systems developers and integrators. See Preservation Metadata: Implementation Strategies, or PREMIS (www.oclc.org/research/projects/pmwg/), for information about one important effort. But since vendors won't develop preservation-aware solutions until customers start asking for them, it behooves media properties to be well-informed about preservation and their own internal long-term retention strategies.

How do we pay for this? Some of the thorniest questions concern how to pay for sustainable digital collections. There are more questions than answers. What is the value of the collection, and to whom? What is the ROI for text, images and other material, such as Web pages and video that is of little or no commercial value, but has intrinsic historic worth? The contents of news archives are the history of a city, a nation, a culture, a snapshot of an epoch of humankind, but if you can't sell it on your Web site, how can you justify the expense of maintaining access decade after decade?

The short answer is that it might not be feasible. The problem might just be too big, too complex and too expensive over time for individual media properties or even their parent companies to sustain on their own.

In the research and academic world, there is ongoing work to scope out models for "trusted digital repositories," third-party entities that have the mission and expertise to take in the digital contents from outside archives and do the preservation work on behalf of their customers, guaranteeing continued access according to a predetermined set of criteria.

Cooperative efforts — perhaps an industrywide project — would leverage what limited expertise exists while the field grows and attracts more practitioners. Research and development funding, moreover, could be spread among a larger pool. But that will still require a concerted effort at standards development and best practices to be a realistic proposition. This will require partnerships between media companies and vendors, as well as rethinking established newsroom workflows.

What about what we have already archived? Another provocative question is, what has already been lost? News databases are full of complicated multiplatform formats, compound, complex objects and nonstandard, locally customized metadata schemas. A standard for preservation metadata is close, but implementation will take a few years. Without these critical components of a preservation-oriented archive, how will old data move forward or how will it be rescued after the fact if migration fails? Is there already a gap in the historic record? Some archivists believe the 1990s are already gone. Only time will determine whether they're alarmists — or actually right.

Fortunately, I know that my Jan. 1, 2005, picture from the devastation at Mullaittivu will be human-readable in 2075. It'll be on microfilm. **TSR**

About the Author

Victoria McCargar is involved in newsroom and library technology support and strategic planning at the *Los Angeles Times*, where she is a senior editor. A frequent lecturer, she is a member of two international teams researching digital preservation and is investigating standards and preservation strategies for the newspaper industry. She is an adjunct professor at UCLA and holds masters degrees in information science and journalism. She can be reached at mccargar@mac.com.