Preserving Digital Documents: Choices, Approaches, and Standards*

Clifford Lynch**

Mr. Lynch surveys some of the technology approaches related to digital archiving and some of the developments in the area of standards, seeking to help readers identify areas that might be worth tracking and factoring into their thinking.

¶1 I want to begin this examination of issues concerning technology and standards as they relate to preservation by scoping my topic a bit. When we talk about technology and digital preservation, there are really two separate kinds of conversations that can go on. One is about using digital technology—particularly imaging technology—to capture physical documents or physical objects and convert them to digital form, or more precisely, to create digital surrogates for them. You can argue that one of the reasons for doing this is a preservation reason,[1] although there are many other reasons to do it, most notably access. The other conversation is about how we deal with the preservation of things that we might characterize as *born-digital* objects. We are seeing more and more of those. A great deal of material now enters existence in digital form through some kind of word processor or Web-page authoring tool, spreadsheet, or other sort of software system. It may later on in its life go into printed form, but it starts life in digital form; and increasingly there is interest in how to capture things in digital form. And keep in mind that as we author more for the digital environment, more and more born-digital materials can't be reduced to print—at least not without substantial loss of content or function. We are moving beyond just using word processors to author what are ultimately printed pages on paper.[2]

** Executive Director, Coalition for Networked Information, Washington, D.C. This article is based on a paper presented at the conference "Preserving Legal Information for the 21st Century: Toward a National Agenda," Georgetown University Law Center, March 6–8, 2003, and has been prepared from a transcript of that talk, retaining the informal character of the presentation. Given that there has been rapid progress in some areas during the roughly eighteen months between the conference and the publication of this article, the author has felt it essential to add a few notes in proof indicating such developments as of summer 2004, but has not attempted a comprehensive revision.

1. It is worth noting that the legitimacy of digital surrogates for preservation purposes has been a hotly debated issue over the past decade, but with the steady improvements in quality and cost-performance, a consensus that digitization is at least as reasonable a path to preservation as more traditional microform technologies seems now to be rapidly developing. For example, in mid-2004 the Association of Research Libraries explicitly recognized digitization as a preservation reformatting method.
2. At least in the sciences, we are now seeing journal editorial boards declare the digital version of journals published in both print and digital form to be the definitive version of record, directly recognizing the greater range of expressive power available to authors in the digital medium. Much more on this shortly.

¶2 I'm going to confine my focus pretty much to born-digital material. I would point out that if you use technology to preserve the physical by making digital versions, you immediately create the same kind of problems that you would have with born-digital material, it's just that it's usually simpler. Why is it simpler? This is actually worth exploring for just a moment, as it helps us to see the difficulties we face with born-digital materials more clearly. One reason it's simpler is that the capture process is typically a lot less anarchic and variable than the authoring process. When you think about systematically digitizing materials that exist in print or other physical form, it usually involves some kind of library or archiving organization. It's done on a planned and professional basis, and there's some kind of systematic attention to appropriate standards of metadata and things of that nature. Also, there is by and large some rough intellectual consensus on what you're trying to capture when you image a piece of paper or a photograph. There are, in fact, well-developed best practices about resolution, about how to do color calibration in color material, and other similar kinds of issues. In a sense, physical documents represent an implied consensus about what you're trying to capture in a given context.[3] The central issue—what you are typically trying to do when you are capturing a document by imaging it—is fairly well understood.

Setting Preservation Goals for Born-digital Objects

¶3 With that in mind, let's return to talk about things that are born digital. Let me suggest that the essence of the complexity of dealing with this born-digital material is that we really don't have good consensus yet on what we're trying to accomplish through preservation, and the farther away our digital object drifts from behavior that is rooted in physical artifacts the more uncertain we get. To make this concrete, let's think about the progression of a journal to digital form. As you probably know, the vast majority of scientific journals now are available in digital form, although very few of the traditional publishers have yet abandoned print. Now when this first started—journal issues available in electronic as well as printed form—nobody got excited about it from a preservation perspective. It was great from an access perspective to have the digital material, but from a preservation perspective you could say there's really not a problem here. The publisher is doing dual print and digital publication. The two are identical—and, in fact, if you look at some of those journals when they first went online, they are *nauseatingly* identical. What you get online is an image rendering of the physical page complete with artifacts like multiple columns, which are not very helpful on a display

3. I will admit the consensus is not universal. There are people who are concerned about capturing watermarks on the paper in support of certain kinds of research. There are people who go around sniffing materials in archives in support of certain kinds of research. But these are fringe concerns that interest only a tiny part of the audience; the digital surrogates may be of no help to these people, who may need to still go to the original physical artifacts and may be very unhappy (and vocal) if the physical artifacts cease to be available for some reason.

screen. But you could say from a preservation point of view: "No problem, it's identical." (Indeed, the more identical the better!) We've got the physical stuff, we know how to take care of physical paper, particularly when it's printed on good acid-free stock.

¶4 No problem so far. Okay, then what happens? Well, the publishers start saying that maybe they'll stop producing the print versions because they are expensive and very few libraries are still buying them. As long as the content is just a digital version of a print page you can solve the preservation problem by saying, "Oh well, we'll just print it out in the library on good paper and put it back on the shelves as a preservation copy." (And this, of course, is stupid. Better to keep the digital versions of the pages. But the mental picture of printing out the digital material onto paper is important, because it says we all know, we all agree upon what successful preservation means—at any point in the future we can convert the digital content that we have been preserving back to print and get the same printed pages!)

¶5 The trouble starts when some creative scholars start saying: "You know, authoring in the digital environment affords me opportunities I don't have in print. I could embed multimedia clips. I could build interactive simulation models. I could attach digital data sets that the reader can manipulate to revisit my reasoning as parts of my article." All of a sudden you have a piece of authoring that doesn't reduce neatly to print. You may be able to print out some aspects of it, but what we're seeing now in science publishing is more and more publishers making this very critical intellectual transition in which they say that the version of record, the *authoritative* version, is the digital version. These publishers will continue for the time being to produce print versions of the journals, but these are impoverished in some sense. They don't fully capture the content available in the authoritative digital version. Now when you start talking about this kind of content, it becomes difficult to understand and agree about exactly what you're trying to preserve.

¶6 Let me give you some examples that drive this point home. In some of its more recent versions, Microsoft Word, a popular authoring tool, has a whole series of functions that let you track changes made while editing a document; there are options you can set (and change) to tell it whether to actually display a history of these changes on screen or when a document is printed. In fact, many people don't realize it, but when they send a document someplace it may have embedded in it, just waiting for someone to ask Word to display it, the whole edit history of the document, unless they take specific measures to get rid of that. There have been any number of embarrassing incidents when people have fired off press releases that way, for example. Then recipients—in the press, for example—have had great fun decoding publicly the revision history and editing of the press release. Do you want to keep that? Is that part of the document? Or is that not part of the document? All of a sudden documents aren't fixed any more. They interact with tools you use to view them in complicated ways. We, as scholarly communities, really have to make choices about what is the right thing to do in trying to preserve this, and the choices are going to vary from community to community and context to

context. The more detail and functionality—particularly in random kinds of things you want to capture—the more complicated it gets.

¶7 Let me tell you another story just to show you how really complicated this can get. There is a lot of software out there that is not very well written, including some well-known and widely used commercial software. Among the annoying properties of some of this software is that of not clearing unused memory before it writes things out. So, for example, an image you create or document you author may contain several kilobytes of random garbage that just happened to be sitting around in the memory of your computer or out on your disk, depending on how this happened. Web pages, old e-mail, other documents—who knows what it was—it could be anything. And of course if you view the document or image through the software that created it and that is designed to be used to read it, you won't see any of this garbage. The software knows how big the object is, it keeps track of the size, and it won't show this other garbage. But if you view it with a different kind of a software tool, you can actually look at all of this random junk that was captured with the documents and is being carried around with them. Do we need to preserve that? The obvious answer would seem to be no—we want only the document or the image. But I would suggest that if we do keep it, deliberately or by accident, sooner or later some scholar will undoubtedly exploit it to very interesting use in some context.

Two Preservation Strategies for Born-digital Objects

Migration

¶8 The two major strategies currently on the table for getting content usably into the future are *migration* and *emulation.* Migration is basically a strategy that says: "I'm going to copy the document from an old format to a new format periodically as document standards evolve." And if you think about it, there usually is sort of a window of opportunity when, if you are dealing with mainstream document formats, you could do that copy and reformat fairly gracefully. Whenever a new format comes along, if there is a substantial installed base of documents based on the old format, there's a strong economic motivation for vendors to build software that handles the conversions**.** The current version of Microsoft Word, for instance, reads documents saved in formats used by older versions of Word and also by some of its competitors. It read even more versions when it had more—and more serious—competitors in the word-processing marketplace. When there was a big need for migration from some of the old CPM-based word processors to the then-newer DOS word processors, there were tools that did that. These tools don't stay around forever. Today, of course, if you are unfortunate enough to have an old eight-inch CPM floppy disk, you will have a lot of trouble finding not just hardware readers for it but software to migrate it.

¶9 But the strategy with migration is that you treat these documents (and I am using the term document here in a very generic way to include images, sound

recordings, videos, and other kinds of digital objects) as a kind of living thing, you pay attention to them on an ongoing basis, and you copy them during the windows of opportunity. This is a strategy that is full of obligations for constant attention. It entails many curatorial choices, and it requires some careful quality control and thought. If you have taken a complicated document from an old format to a new format, you've probably noticed some garbling, errors introduced by the conversion or migration process. Tables, footnotes, diagrams, equations, and other things you might want are frequent victims of this, so you've got to be careful about degrading the document or other objects as you're moving forward. Sometimes pagination isn't faithfully preserved under migration, leading to all kinds of issues about permanence of reference to specific parts or passages within a document in the absence of more robust but perhaps more intrusive apparatus (such as paragraph numbering). And of course, as I've already indicated, migration is a subtle business because we aren't clear about the behavior of the documents we are migrating. Do we want to simply replicate the appearance to the reader—or do we need to preserve more structural things, like the revision history or outline that might be embedded in the document and viewable through one generation of word processor, but perhaps not another from another vendor?

¶10 Migration is one strategy. I think it's a very effective strategy, at least for a large class of materials that have "document-like" characteristics. But it's a strategy that is annoying in the following ways. Can you prove it works in perpetuity? Well, you can't. You just say it worked up until now. How do you cost it out? Well, you don't, because it's going to be another hundred years before we really have sufficient data points on how often it's necessary to navigate one of the format conversions and how much it really costs when we do. Every migration has a nasty ad hoc element of judgment and curation to it. So we don't have a general theory at this point, or not much of one. And we don't have good economic models, but as a sort of best guess it looks quite practical, particularly for things we think about as static or relatively static documents, as opposed to say complex interactive programs.

Emulation

¶11 The other major approach, which has been championed most notably by Jeff Rothenberg of RAND, is called *emulation.* Here you try to move the computer into the future; but instead of trying to move the physical hardware of today's computer into the future, what you do is write software on new computers that emulates old computers, then you keep building layers and layers of these emulators. For those who are not terribly technical this probably sounds like a completely insane idea, but in fact, there's a long tradition of this. Back in the 1950s and early 1960s, every time IBM brought out a new computer model it was completely different from the former model of computer. It didn't run the software that the old computer ran—at least directly on the new hardware—until the days of System 360 in the 1960s. Customers didn't like this: they basically had to throw out a lot of the software they'd written and start over every time they went to a new series of computers.

So IBM actually wrote emulators back in the late 1950s and early 1960s so that when you bought the latest and greatest new IBM computer you could still run your old program—albeit very slowly using this clunky emulator software. More recently, in the 1990s, Apple Computer wrote an emulator when it went from the old Motorola 68000 chips to the newer PowerPC chips in the Mac. Apple did this on the consumer product well enough that it hardly broke anything, and most people didn't even know they were dealing with emulators.

¶12 There is a thriving hobby activity in emulation right now, writing software for modern personal computers that emulates the old video games you used to feed quarters to in bars and video arcades. This works well enough that people still read memories out of these very old—now collectible—arcade machines in order to get the game programs themselves. They read the memory out into a file (we will ignore the copyright considerations here), feed the file to the emulator on the PC, and up pops the arcade manager's configuration screen saying: "How many lives do you want to give the customer for a quarter?"

¶13 Actually, emulation really does work to an amazing extent—it is a very well-established and respectable practice in computer science and engineering. Among the things that get troublesome about emulation is, first, timing. Timing is often completely off, which means that anything interactive will get screwed up. Interestingly enough, while I have just described how video games are being run on these emulators, they are often quite annoying to play because the timing is all loused up, the rhythm of the game is off, so they really aren't as satisfying as you might think. They are a way of getting an idea of how the game works, but are a far cry from a faithful reproduction of the experience of playing the game.

¶14 There are other problems with emulation. As soon as the computer wants to communicate with peripherals like storage and display devices, there are often problems. This is very complex to emulate, if for no other reason than the huge variety of devices that are around, and the fact that they often come from many different vendors and represent a very delicate set of negotiations about how specifications are interpreted and who is the final arbiter of these interpretations. In addition, we are increasingly facing the problem that computers don't stand alone; they exist in a very complicated network context, and they rely on services and content that are scattered all over the network. It's not at all clear how to emulate the network environment that a computer exists within and depends upon today. Emulation by itself won't emulate the whole network, just one machine. You've got to question how much of the environment around the document, or around a digital object more generally, we need to reproduce for it to be meaningful. My view is that, again, there is a set of curatorial and archiving choices that are best made in a focused way around certain genres and classes of documents. You don't just say, "Oh, I can emulate the hardware that the reader ran on, so the problem is over." You've got to consider the content you are trying to preserve and how you are viewing its essential characteristics and attributes for emulation as well as for migration.

The Importance of Context

¶15 It's easy to get very focused on documents in the digital world when we talk about preservation and to forget the lessons that archivists have learned about the need to save not just documents, but also the *contexts* for the documents in order for them to be meaningful and interpretable. And when we move from saving published digital works (which, after all, are expected to stand alone to some extent) to more complicated things, and particularly materials that might exist within complex organizational settings, context becomes very important. Here's a quick example to give you a feel for this. Suppose I'm a company and I decide I'm going to save all my corporate e-mail as a record because I'm supposed to. That e-mail is apt to be nasty to interpret over time unless something also is done to save the corporate directory that identifies the e-mail addresses. That's not part of any specific piece of e-mail necessarily, but a sort of contextual thing off to the side. Same thing with instant messaging.

¶16 Nowadays we are starting to see technology deployed to do digital signatures to deal with authenticity and origin of certain kinds of digital objects (for example, so-called Public Key Infrastructure, or PKI). Preserving that means coming up with a whole preservation strategy for a pretty complicated piece of infrastructure that is not part of any individual document but rather is in classes of documents—or the context. When we start thinking about actionable citations, or *hyperlinks* if you prefer, this again implies a whole contextual infrastructure of identifiers, things for which strategies need to be developed. So focusing too narrowly on documents or digital objects without at least some consideration of their broader digital ecosystems raises the likelihood that you are going to miss some things that you really want to hold on to.

Standards for Digital Preservation

¶17 As a final topic, let's consider the state of standards that are relevant to digital preservation. When we talk about capturing surrogates of physical objects digitally—in other words, using digitization for preservation of physical objects and enhanced access—there's a wealth of experience to provide best practices and good guidance. You can look at things like the *NINCH Guide to Good Practice*[4] and the work that has come out of the Digital Library Federation.[5] And the Northeast Document Conservation Center's School for Scanning has done a

4. Nat'l Initiative for a Networked Cultural Heritage, The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials (2002), *available at* http://www.nyu.edu/its/humanities/ninchguide.
5. *See* Digital Library Fed'n, *at* http://www.diglib.org (last visited July 19, 2004) (listing one of the Digital Library Federation's goals as "identifying standards and 'best practices' for digital collections and network access").

lot of valuable educational work in this area.[6] There is a lot of prior work you can build on, although most of it isn't formal standards because the right answers are somewhat situational. Nonetheless, you can find guides to good practice, which are informed by the experience of others, that take you through a decision-making process suitable for your material.

¶18 When you deal with file formats, there's nothing overly magical about this. There are well-understood and well-documented image file formats that are used for a multiplicity of reasons and applications—TIFF, JPEG, etc. The key issues are making sure that you don't inadvertently get involved with lossy compression,[7] and that you carefully document contextual information such as color calibration.

¶19 But when we get into the born-digital area, there are really very few standards or even guides to good practice to draw on. Basically, what you get concerned with is popular document format specifications and standards. Some of the key standards are proprietary (think about word processors, for example). One thing that is worth tracking is a working group that is looking at archival issues around PDF, the portable document format developed by Adobe and now in very wide use. This effort is happening under the leadership of AIIM (Association for Information and Image Management), and NISO (National Information Standards Organization) is also tracking the work. PDF is a good example of something that looks simple, yet is anything but. As you really delve into it you discover that things as simple as fonts can create all manner of trouble when you start recognizing that faithful pagination—which is, in fact, tied to fonts—may be an important issue. The result of this work is a specification for a more archivally manageable subset of PDF called PDF-A, which is currently working its way through the standards process.

¶20 Beyond that there has been some useful work done on metadata standards for preservation, although that work is not highly advanced. Part of the problem is that a lot of the work on preservation metadata has given rise to organizational guidance, the kinds of things you should think about as you attach metadata to objects when you want to preserve them, rather than hard specifics that would be

6. *See, e.g.*, Northeast Document Conservation Ctr., School for Scanning: Building Good Digital Collections, *at* http://www.nedcc.org/prelsfs/prgen.htm (last modified Jan. 13, 2004) (describing 2004 conference in Chicago, June 2–4, 2004).

7. Without getting overly technical, there are two kinds of compression: lossless and lossy. Lossless compression simply codes data more efficiently, less redundantly: if you take a set of bits, compress them losslessly, and uncompress them, you get the same set of bits back. Lossy compression techniques exploit knowledge of the human perceptual system to produce a more compact representation of something—a sound recording, or a picture, that "looks like" or "sounds like" the original, but is much more compact. However, it doesn't give you the same bits back when you go through a compress and uncompress cycle; and while often these lossy compression techniques produce results that are very good, they are subject to various kinds of "artifacting" where they fail on specific sorts of data patterns.

more typical in interchange format, because preservation metadata today is seldom shared across organizations.[8]

¶21 The other standard that is relevant but limited is the OAIS, Open Archival Information Systems model. This is work that came out of NASA and the scientific data archive community initially, then was adopted more broadly, and now is in the final stages of becoming a formal international standard. It is a very, very useful standard, but it's one that many people misunderstand. It is a high-level conceptual model for understanding the functions of archiving in the digital world, for talking about them, for establishing common terminology. It's not a blueprint for how to build a digital archive. It's the framework from which you might start to build such a blueprint, but it is not itself the blueprint. So there's a great deal of work that needs to be done, from the general kinds of guidance and models that OAIS gives you to very specific archival and digital preservation strategies. You should be very worried if anyone tells you that they are doing OAIS-compliant digital archiving with the implication that this is comprehensive and definitive.

Conclusion

¶22 This has been a very rapid survey of a very complex area. I hope that it has been useful, not just in pointing out some of the standards developments that may help us with digital preservation, but in framing some of the fundamental policy and curatorial issues that have to be addressed in developing digital preservation strategies, and in underscoring the reality that digital preservation is not accomplished through a single "right" strategy, but is ultimately a set of carefully considered choices in support of specific objectives.

8. Since this talk was given there has been a great deal of progress in relevant areas here. While I do not have the time or space to describe the work in detail, I would point the interested reader at the work on METS (Metadata Encoding and Transmission Standard), the PREMIS group which is dealing with provenance-related metadata, and the NISO Still Image Technical Metadata draft standard and its representation in XML called MIX.