Washington, D.C.
October 15-17, 1997

Preservation of Digital Information

When Technology Leads Policy

Clifford Lynch, Executive Director
Coalition for Networked Information

I'd like to briefly cover two specific classes of technologies: identifier systems and rights management systems. I want to put these kinds of technologies in a context, and to discuss how technology choices can actually preempt policy decisions. One of my key illustrations follows naturally from some of the points Bob Oakley made at lunch today. I didn't know he would be joining us, and I was delighted when he laid a lot of the groundwork for what I want to say. Bob made the point that some of the important practices we have historically enjoyed and counted on, to the point that we now rather glibly talk about them as "rights" — such as the right to fair use — are really not legal rights at all. Legalistically speaking, fair use is a defense that is invoked after an action. For example, if you have copied something and someone charges you with infringement, you can appeal to fair use as a defense. If you can't do something in the first place, it doesn't matter what defense or justification you might have been able to cite in support of that action, had you been able to take it. Legal defenses are not traditionally arguments for why you should be capable of doing something; rather, they are arguments for why you should not suffer legal consequences for having done something. Specifically, the fair use defense does not, as I understand it, in any way imply that people must have the ability to make copies, but only that, if they can manage to do so, then in some cases they may be able to justify or defend that action through an appeal to fair use.

Historically, in our society we have had a great deal of freedom to act; we are held accountable for our actions. This accounting, when required, takes place within a broad framework of law and human judgment and considers a very broad context of purpose, intent, and consequence. As a society, we have also developed a wide range of commercial and social practices and expectations which go far beyond any legal guarantees. And, until today, there have been many activities that have simply been impractical to monitor, regulate, or prevent. However, we are now being presented with technology options that don't necessarily allow us to take actions and be accountable for them. They also create the possibility of restructuring our commercial and social practices in radical new ways. They

make it possible to control and monitor activities that have historically been impractical to address.

We have developed certain expectations, certain practices that we have so internalized that many of us may believe they are rights. But I'm not at all confident that they really are rights, in the legal sense, within the current legal framework. To take another example besides the area of fair use, I think many of us expect that we should be able to browse or read materials, without necessarily identifying ourselves or the details of our reading to the author or the publisher of those materials. (Note that we may have to go to a library or a bookstore to do this, to borrow or purchase a copy of the work in question. I'm not suggesting there's a widespread view that authors or publishers are obligated to give copies of works to any anonymous reader that expresses an interest in them.) I will leave it to wiser people to discuss the legal specifics of reading anonymously, noting only that my colleague on this panel, Julie Cohen, has written a very interesting paper on this topic. But I think that these two examples — fair use and the anonymous/unmonitored reading — illustrate areas where it's not clear that legal defenses and social and commercial traditions of practice are going to guarantee that we can carry our expectations forward into the digital world. We need to be conscious of the dichotomy between after-the-fact justifications and protections (such as protections of privacy) that one might take within a legal system and actual affirmative rights. This dichotomy underlies all of this discussion.

Let me talk first a little bit about identifier systems. Certainly, identifier systems are the foundation for many activities, ranging from rights management to citation, and I, at least, view them as an essential construct for building our whole record of social and scholarly discourse. Without the ability to cite, we would be in very bad straits. Citations and identifiers to facilitate citations are not new concepts to the library community. We have been using them for centuries in various forms. Identifier schemes have always been very powerful applications in a fundamental way. But now there's something new in the network environment. Identifiers and citations are actionable. Somehow they are not merely passive textual data — they are something that can actually, at least potentially, transport you directly to the material they describe or cite.

This actionable system may be the only path into the network environment. If any of your workspaces are like mine, they are full of stacks of paper. You may have forgotten or lost the exact citations. But searching through those stacks of paper often uncovers that very item that you now remember you need. But did you ever lose a URL to some material? You may remember what site it's on, but often there's no way to get to the specific document without knowing the URL. Unfortunately, there's no analogous concept to serial browsing yet in many of our electronic resources, and we rarely keep stacks of personal copies scattered around our desktops. So naming and citation become just enormously powerful activities. We will have to ask a lot of contextual questions about identifiers when we start evaluating when they make sense and what their purpose will be.

In the most recent ARL newsletter1 I've authored a discussion about some of the

digital identifier schemes that are currently under consideration. I don't want to reiterate their specifics, but I do want to highlight a few characteristics that are common to all identifier schemes and suggest that we think about them in the context of what sort of world we want and what their implications are for the management of content and the ability to use and reuse that content. There are many questions to answer: Who gets to assign identifiers? Who can make decisions about naming and about when two things are the same or different (whether they get the same identifier)? Who can create objects in a world of identifiers so they can be referenced? Do identifiers last for ever? Are they resolvable forever, or do they go away after a while, like so many URLs? How are they resolved, and how is the resolution process managed and controlled? How do you find out what the identifiers are for digital objects?

All identifier schemes need to struggle with the difficult philosophical question of exactly what they are identifying. Sameness and difference are very subtle and contextual kinds of issues. One community will say two things are the same (or close enough): all instances get the same name. Other communities will draw fine distinctions between variant forms of a similar object, and develop complex ways of identifying the differences among these variations (naming the variations). We face those kind of problems in areas as simple as differentiating hardbacked and paperbound versions of the same work. Despite their identical intellectual content, some identifier schemes distinguish their binding because it's important for sales channels.

When we move into digital environments, we have some profound questions which interrelate in rich and subtle ways to some of the questions we were talking about in the preservation and reformatting context earlier. If I have a document in two different formats, are they assigned different identifiers? Or should they all be named the same, and should choice among formats be an artifact of the transaction that I perform in order to access the document (in other words, the system just asks me if I'd like it in ASCII or PostScript)? Jeff Rothenberg showed us one way we can answer the question: When is meaning altered in a fundamental way? These are imprecise, contextual, and ultimately philosophical issues, but they are fundamental to understanding and evaluating identifier systems.

There are two final considerations about identifier systems that are not talked about much. First of all, you want some method of resolving identifiers to the objects that they identify. In the print world we have various printed directories, for example, of ISSNs. These resources have evolved into electronic databases. Most of these databases are used by libraries, suppliers to libraries, and booksellers — today more as part of the business apparatus, than as an end-user access apparatus. As identifiers are used to create electronic interdocument links and references, however, they will become an ever-larger part of the basic access apparatus that every reader will need to navigate.

We need to recognize that, as soon as a print resource turns into a database, use of that database potentially leaves electronic trails. Therefore, it's always worth asking: Who gathers those trails and what do they do with the information? What

are the privacy issues? What are the statistical aggregation issues? What are the market issues? We've seen, for example, the use of citation indexes in tenure and promotion in some institutions. Suppose you actually had counts of how often materials were being read. Is that really information you want public? And, if so, under what constraints and framing? Are we going to have scholars building software robots that do nothing but reference all their articles, as many times a second as their computer can do it, to run up those counts? We need to think about this collection of issues. As we start talking about identifiers that are user-oriented, we need to think very carefully about these privacy and control issues.

Finally, I want to discuss the issue of citation. Actionable, navigable citations are one of the most attractive aspects of the networked information environment. In talking with readers, one of the things that they are most excited about is the ability to move from citation to cited work at the click of a link. Yet, I believe we need to be very careful about the use of identifiers in citation. One of the most appalling prospects that I can imagine is the introduction of a citation system whereby one needs some form of permission to cite — for example, where the identifier value for the work to be cited can only be obtained from the rightsholder, or from a third party that has a monopoly control over identifiers and their resolution. I think that strikes at the heart of everything we've come to assume about the way we do scholarship, the way we do reviews and criticism and analysis of events, and the way we conduct our discourse. Any kind of identifier system that introduces that requirement is tremendously dangerous as an underpinning for citation.

Let's go on to rights management. First, I want to observe that if you can't identify material, you can't really talk about rights management. So there's a very direct connection there.

There are two basic classes of rights management systems — or, more precisely, things that are being generally referred to as rights management systems — with which I'm familiar. (Actually, for completeness, there is a third class of systems dealing with watermarks and unauthorized copy detection, which I won't discuss further today. I would also argue these are not appropriately termed rights management systems.)

One system is built on registries about rightsholders. At its most basic level, such a registry is a database would be indexed by various kinds of identifiers that name who holds the rights to the material and where to contact them. It may also include other information, such as standard terms and conditions and royalty rights for common uses of the material. This is useful, particularly for people in the multimedia area who want to repurpose and incorporate content into composite multimedia works. These multimedia developers are often spending more money tracking down rightsholders and clearing rights (and not necessarily in actual royalty payments to rightsholders, but in legal and detective work so they can make the payments) than they are for any other component of building multimedia works. We desperately need databases like this.

(I can't resist a small aside here. One of the charming changes that has happened fairly recently is that copyright law now passes a work into the public domain as a function of when the author died, rather than when the work was published. Before this change, you could usually make some assumptions about when a work was likely to pass into the public domain. Now, you need to hire a detective agency. Registries that include such details as whether or not the author is deceased and, if so, when, are going to be very useful. Hopefully, everybody is doing something to preserve the obituaries from their local newspapers, because we're going to want them.)

That's your basic model of one fundamental type of rights management system. And you can extend this model a bit by enabling automated transactions of certain kinds of rights clearances. For example, if you are willing to accept a standard license agreement, you can submit a payment through the system and immediately obtain authorization for 50 copies of an article for a course reader. Systems like this are already well along. They don't prevent anything: rather, they facilitate compliance with copyright and commerce in copyrighted materials. They reduce transaction costs.

The other kind of system that falls under the general rubric of copyright management goes by names such as envelopes, secure containers, lockboxes, or Cryptolopes. Some of these are really product trademarks or perhaps service marks. For such systems, content comes integrated with (often aggressive) rights management software. For example, the content is "wrapped" in a layer of software, so that you can only print it, view it, or duplicate it by going through the wrapper software; you can only utilize the content with the permission and cooperation of the wrapper. And the wrapper can keep track of who is reading what, and how often, and can potentially not only gather data but can also report to third parties. In most variations of this theme the content is encrypted so that you cannot bypass the mediating wrapper software. In some variations the encrypted content doesn't come with its own software, but rather requires that a general-purpose trusted "reader" software environment already be installed on your machine; this reader interprets the content and its accompanying rights management data. This kind of wrapper technology is troublesome for several reasons.

The first reason is technical, and I'll only go into these issues briefly because they are complicated, because the details vary greatly from scheme to scheme, and also because many of the details of specific schemes are highly proprietary. But these comments should offer some grounds for concern. The software that is needed to make use of this wrapped content effectively "colonizes" your machine on some other agency's behalf in the name of creating a safe haven for intellectual property. It creates a "trusted" environment on your machine — protected from you — in which the wrapped intellectual property can exist and be used "safely." The words here are interesting, and revealing: "trusted" means that a third party, like a rightsholder, will trust your machine to ensure that the its constraints on the content take precedence over actions you may attempt on your own machine. "Safely" means that digital objects on your own machine are safe from your own ability to perform (and, presumably, when necessary, to justify or defend) actions

upon them.

In computer science this is sometimes termed a distributed "trusted system" model, the idea being that your machine forms part of a larger system with behaviors that can be depended upon by third parties no matter what the owner of each individual machine may do, and that the owners of individual machines cannot subvert the trusted system's expected behavior. We have many examples of trusted systems (of varying degrees of robustness and incorruptibility) today: for example, taxi meters, which are installed and verified by third parties, supposedly cannot be altered by the taxi driver. They serve as a fair witness to the fare for both driver and passenger. Certified weigh scales are another example. Most of the trusted systems that we are familiar with and accept are relatively passive, and record and redistribute rather than constrained information. But I would suggest that a taxi meter with a video camera that recorded passengers, their behavior and conversations, and their points of embarkation and debarkation and broadcasted this information for access to any third party willing to pay to access it would meet with considerable resistance.

Now, there are reasons why you might want to go along with the installation of a trusted system that takes over part of your computer as a condition of being able to access digital content, but I don't think anyone has thought adequately about how to audit and vet such systems. Certainly computer scientists, particularly those who work in security, will agree with me when I say that nobody really knows how to guarantee that a trusted system doesn't open up all sorts of security and privacy problems — especially if it can interact with other machines over the Net — other than empirically, that is, the system has been widely studied and tested and used and nobody has noticed a problem yet. (Remember that many of the current trusted systems being proposed for managing intellectual property are highly proprietary.)

We have seen some examples that give us a taste of the issues. There was much concern over "registration wizards" (as one large company termed them, though I think that the idea is far from unique to that company) that were shipped along with various commercial software packages. Nominally, they were intended to "serve you better" by making sure that the software vendor had enough information about your configuration to diagnose problems... until people realized they weren't sure what these wizards were doing — uploading a copy of your configuration and all the software that you had installed to a third party (is all of this software legal? do you really want every body to know what software you have?)... or perhaps a list of all the names of your data files... or perhaps a copy of all your documents? Paranoia, rumors, and urban legends abounded, and it became clear that most users had no idea how to figure out what such a wizard was doing, or even whom to trust to figure this out. That's a good example of what happens when there's no audit or verification mechanism for trusted systems. This is a technical issue that we can not and should not overlook.

Personally, the security and privacy issues involved in trusted systems — particularly those deployed by potentially interested parties to facilitate commerce in a competitive marketplace — make me very nervous. I'm not at all

enthusiastic about having these on my machine, even though I doubt I have anything very interesting to hide. Perhaps I'm old-fashioned and paranoid: I'm not very comfortable with programs that automatically download updates, install system files without telling me in advance, have web browsers that download executable code from remote sites, or lots of other things that seem to be growing more commonplace these days.

There is another problem with these trusted systems which isn't technical. To return to my first point in this talk, these systems constrain your actions rather than permit you to be accountable for them. These systems can prevent you from making fair use of content. Remember, you don't have a right to fair use today. If you can make a copy of something without permission, and if the rightsholder challenges your action, you can mount a defense based upon fair use. That these systems can prevent you from using content at all in some specific ways, or from using it without paying for that use, is troublesome. It makes the issue of fair use moot, or predicates your ability to make uses (and later argue that they fall under fair use) on your ability to circumvent the trusted system. I am not optimistic that we'll see software that puts up a payment menu and then offers a button that says "Click here to bypass payment if you believe you are making a use of the content covered under fair use." Similar questions arise around the ability of readers to read anonymously.

And I don't think these content guardians, if we accept them, are necessarily going to go away gracefully as material passes into the public domain. (It would be very hard for them to determine whether 75 years have elapsed since the author of the material has died — think of the problems involved in engineering software that reliably self-destructs based on such a criteria. Simply engineering it self-destruct or become inactive at some certain date is an enormously difficult problem.) There is no right or guarantee that we can make copies of public domain material, at least as I understand it, only protection from legal action charging copyright infringement for materials that are in the public domain. (An exception is that government agencies are required to provide people with copies of some materials that are in the public domain by virtue of being created by the government.) At the core of much of the work that libraries and other social institutions do to preserve materials is the belief that we are holding these materials in trust for future generations who will be able to use them when they ultimately enter the public domain. Putting protective mechanisms on content, and making content available only when coupled with these protective mechanisms, undermines that premise in a very basic way.

I titled this talk "When Technology Leads Policy," and I hope that I have at least illustrated a few areas in which technical developments — if they find acceptance in the marketplace and if the technologies themselves prove successful — threaten to radically shift well-established social practices and expectations. As I've tried to show, the legal framework that has developed to underpin and guarantee these practices and expectations is not likely to be equal to the challenge raised by these new technical developments. Knowing that actions are justifiable or defensible means little when those actions themselves become impossible in the new technical framework. It's important that we not simply

capitulate to technological determinism in the digital environment: we need to think carefully about what limitations on our actions we will accept, and perhaps even whether some of our cherished practices and expectations need to be re-codified as affirmative rights, balanced with appropriate accountability for their exercise.