

# **A Discussion of the Interface between Legal and Technological Issues in the Provision of Digital Reference Services**

*Prepared by Michael McClennen, Ph.D. for the  
Project on Digital Reference Legal Issues*

*American Library Association Annual Conference*

*Chicago, Illinois - June 23, 2005*

## 1 Introduction

The goal of this paper is to present a technical basis for a discussion of legal issues in the field of digital reference. We will undertake to lay the groundwork for this discussion by delineating the interface between the technical and legal domains, setting out those technical factors whose behavior and consequences have legal implications. In order to do this, we must carry out the following steps:

- 1) Define the set of legal issues that we will be considering
- 2) Define the set of technical factors that are relevant
- 3) Enumerate the legal implications of the various kinds of technology used in the provision of digital reference

We will not attempt to discuss specific legal issues, as that is the purview of the other presenters. Nor will we be evaluating individual products, since these change frequently and are quickly supplanted by newer ones. Instead, we will be focusing on the development of a taxonomy by which any product or system can be analyzed and compared with others.

We will start this process with a summary framework for classifying the legal issues under consideration, in section 2. Section 3 will then discuss the nature of information, and its implications for technology and law. Section 4 will return to our framework, discussing each point in greater depth. Finally, section 5 will discuss the implications of various technologies with regard to these issues.

## 2 A taxonomy of legal issues

The areas of concern that have been chosen for discussion in this session fall into two categories:

- privacy and confidentiality
- copyright and licensing

Interestingly enough, these are both essentially issues of *information integrity*. By this, I mean the following: in the course of using and providing information-based services, whether commercial or public, we rely on a shared understanding about the ways in which information will be allowed to flow, and who will get access to it. This shared understanding underlies the economy which supports the production and dissemination of information in the first place.

So long as this understanding is followed about a given piece of information, we can say that “its integrity is preserved”. This might mean, depending on the situation, that nobody has seen that particular piece of information except for those who are supposed to have access to it, or that copies have been made only under certain circumstances and that all copies are in some way accounted for. If, on the other hand, this shared understanding has not been followed, we say that “its integrity has been violated”. As we will discuss below, the inherent nature of information is such that:

- the integrity of information is very difficult to maintain

once violated it is difficult or impossible to restore completely

These unfortunate points underly not only the reason why this conference session was called, but in general the current turmoil in our society over information and the rules that surround it.

The essential function of the law in this regard is to codify this and other social understandings, setting out rules which must be followed and penalties for violating them. The issues we will ultimately consider in this session may derive from many different legal realms, including: criminal law, government regulation, codes of conduct, and private contracts. However, they are all manifestations of the same basic set of principles.

In general, we can classify the legal issues with which we will be concerned according to which set of headings from the following list each one falls under:

- 1) Type of integrity violation
  - a. Disclosure
  - b. Copying
  - c. Modification
  - d. Masquerading
- 2) Type of information
  - a. Personal information
  - b. Contents of questions and answers, and associated attachments
  - c. Other materials shown to the client as part of the answer to their question
- 3) Location of concern
  - a. Client computer
  - b. In transit over the Internet
  - c. Library server
  - d. Third party server
- 4) Agent of concern
  - a. Outsider acting illicitly
  - b. Patron acting illicitly
  - c. Staff member acting illicitly
  - d. Government agency acting with legal authority

Each of these different headings has technical implications. In section 3, we will discuss the various technical factors which are relevant to these issues, and in section 4 we will return to this classification and discuss each heading in more detail.

### 3 Technical factors

The technical factors that inform a legal analysis of digital reference tools are ultimately consequences of the basic nature of digital information. Some of these are fundamental to information in any form, and some are unique to the digital realm. As for the first case, we will consider the following universal aspects of information:

- 1) Anyone who can view the contents of a collection of information can make a copy of any part of it, at minimal cost, without leaving any evidence of this having occurred.
- 2) There is always a tradeoff between control of information on one hand and expense and ease of use on the other.

We will also look at the following aspects which are specific to digitally stored and transmitted information:

- 3) Viewing, analyzing or transmitting any kind of digital information unavoidably involves making a copy of it.
- 4) Access to digital information is almost always granted or denied based on mechanical procedures, with no human judgement present.
- 5) A reference to a piece of digital information, such as a URL or bookmark, is itself digital information and has the properties we have already noted.

In general, these functional aspects distinguish information both from physical goods and from human-provided services, and such distinction is recognized in law. We discuss this distinction more fully, later in this paper. The following subsections will examine each of the five aspects listed above in detail, in order of their importance to our discussion of legal issues.

#### 3.1 Information can be copied

By far the most important aspect of information from a legal point of view is that it can be easily copied. One of the hallmarks of digital information in particular is that the basic copying cost is almost zero. This fact has had a greatly beneficial effect on our society and our civilization, including making possible the profession of “digital reference librarian”. However, it also makes the maintenance of information integrity (including the enforcement of privacy and intellectual property rights) much more difficult. Large-scale violations, such as multi-person identity theft or the republication of an entire database, can be detected relatively easily<sup>1</sup>. Smaller-scale violations, such as the exposure of a single reference transaction, can be difficult to detect and can potentially have extensive repercussions on the lives of the people affected. Integrity violations that are individually of little consequence, if not checked, can also add up to a much larger problem. There are many significant present-day examples of this, including the copying and sharing of music and the copying of articles from reference databases.

Many different technical mechanisms are used to inhibit the unauthorized copying of digital information. Those most relevant to digital librarianship are *access control*, *logging*, and *encryption*. Each of these can be quite useful, but they are all limited in that they are working against the essential nature of information itself. In general, technical means can be used to

make copying more expensive and more risky, but they can never prevent it completely. This is a point which we will emphasize in the following sections.

### 3.1.1 Access control

In general, an *access control* scheme is one that attempts to restrict access to a certain set of information to a certain set of people. These people are presumably subject to legal requirements, such as contracts of employment or subscription, which constrain their ability to copy the information. An example in our own domain would be a database which stores transcripts of digital reference questions and associated personal information, but will not let anyone access this information unless they first enter a valid username and password. Presumably, these are only issued to members of the staff, and the staff are bound by rules and procedures that mandate the preservation of patron confidentiality. Another example would be a reference database licensed by some organization such as a corporation or university, which only allows access via the local campus network or corporate intranet. Presumably, the terms of the database license allow access to just those persons who have a relationship with the organization in question and thus the right and ability to connect to its network.

Access control mechanisms are almost universally used, but their effectiveness is subject to three major limitations. First of all, they rely on non-technical strictures and sanctions in order to be effective. If someone who has been given access to a collection of information is not afraid to violate the terms of the contracts they have signed (and is also not afraid to violate copyright and/or privacy laws) then there is nothing to prevent them from copying the information if they choose. Secondly, even a single “leak” has the potential to do great harm. If one person chooses to ignore legal requirements and copy some information, that person has the technical ability to make an unlimited number of copies and distribute them widely. Thirdly, it is thought to be impossible to create an access control scheme that cannot be suborned. There are technical reasons for this, but it is also true that the vast majority of unauthorized access happens because someone chose a password that was easy to guess, or wrote it down where it could be found, or told it to the wrong person. Human factors are far more significant than technical ones in causing information system vulnerabilities.

While access control mechanisms are not completely able to prevent unauthorized copying, they are still very helpful in the enforcement of privacy and intellectual property rights in the same way that locks and keys are helpful in the enforcement of physical property rights. To see this, suppose that a violation has occurred. In a system without access control, anyone in the world might have accessed the information and copied it. If, on the other hand, access control is in place, the possibilities are much more limited. Either someone who was authorized to access the material did the copying, or someone broke into the system. This makes it easier to investigate the violation, and in either case the violator is subject to harsher penalties.<sup>ii</sup>

### 3.1.2 Logging

Some of the limitations of access control can be remedied by the use of *logging*. This involves arranging that a record be automatically kept of every successful or attempted access to a collection of information. Typically, the log records the following facts about every access: which piece of information was requested, by who, from where, and at what time. This does not

prevent unauthorized access or unauthorized copying, or make them any more difficult, but it does make it easier to figure out who is responsible. If it is known that the integrity of a particular piece of information has been violated, and a sufficiently complete log exists, it is possible to determine which users have accessed that piece of information. Of course, the person whose identity is recorded in the log may not be the person actually doing the accessing. The perpetrator may have stolen the login name and password of one of the authorized users. But at least this gives a clue that might be of use in an investigation. Finally, logging can also detect the repeated accesses that sometimes accompany attempts to break into the system.

### 3.1.3 Encryption

The practice of *encryption* is the scrambling of information in order to protect its integrity. Typically, this operation involves a second small piece of information, called the *key*. Anyone who possesses the key can unscramble the information and return it to usable form. Someone who does not possess the key can only unscramble it only if they are able to “break” the encryption scheme, i.e. figure out the inner workings well enough to bypass the need for the key. Some methods of encryption are easy to break, while others are extremely hard. It is generally agreed that all possible encryption methods can be broken, if enough computing power is available. Given that available computing power is increasing every year, encryption methods that were considered secure a few years ago are now considered vulnerable.

Information in encrypted form can be copied just as easily as any other information, but the copies can only be unscrambled by someone who either acquires the key or breaks the encryption. People may in fact be motivated to copy encrypted information in the hope or expectation that they or someone else will later do one of these two things. However, as long as the integrity of both the key and the encryption scheme is maintained, the integrity of the original information is, too. Any copies that may exist are (for the time being) nothing but gibberish. Assuming that the encryption scheme remains unbroken, one can encrypt a large amount of information using a single key. In this case, the problem of preserving the integrity of the original information reduces to the (presumably simpler) problem of preserving the integrity of the key.

Encryption is extremely useful in a few important domains, but it is largely useless for the general protection of information meant to be accessed by a wide variety of people. Many people, including many so-called experts, misunderstand this. The big limitation on the usefulness of encryption is known as the “key distribution problem”, and can be summarized as follows. Anyone who is supposed to have access to encrypted information needs to obtain the key. Thus, if many people are allowed to have access to the information, then the key must be widely distributed. The key cannot itself be encrypted (if it is then a further key is needed, which leads to the same problem all over again) and is thus subject to being copied and used without authorization. In other words, as a key is distributed to more people, its integrity is more at risk. Consequently, the integrity of all of the information encrypted with that key is more at risk. Many supposedly secure information systems in fact turn out to be quite vulnerable in actual use because the keys are handled carelessly, and are disclosed either accidentally or on purpose to people who should not have them.

That said, we must note that encryption is still a big help in maintaining information integrity if used properly. Most importantly, it is crucial for preventing a third party from “listening in” on communication between two parties (such as a librarian and a patron) over the public Internet. In order to establish a secure communication channel, one must simply make sure that both parties have the same encryption key (or, in some cases, matching keys) so that one can decrypt what the other encrypts. Fortunately, there exists a method (called the “Diffie-Hellman algorithm”, if you care to know) that allows two people to exchange secret keys such that nobody else who may be listening in can find them out. No key distribution problem here.

Most web servers and web browsers have the ability to use encrypted communication. Of course, this only happens if the server is configured to do so. The facility is known as SSL. When it is in use, the URLs generally (but not always) start with “https://”, and by convention the browser displays a “locked padlock” icon. Chat sessions carried out via web pages thus have the possibility of being encrypted. By comparison, instant messaging software typically does not use encryption, nor does e-mail.

### **3.2 Tradeoffs involved in control of information**

A second universal aspect of information in general is the tradeoff between control on one hand and factors such as expense and ease of use on the other. This is easy to see in the case of physical media such as books. For example, a book could be written in code, which would give the author a lot of control over who is able to read it. A book could be kept in a locked case, or in a building accessible only to certain people. A book could be printed using ink which will not photocopy. All of these techniques are in fact employed, and all of them indeed have the effect of providing an increased measure of control over who can read the book and what they can do with it. At the same time, each of these involves additional expense, in some cases ongoing. Finally, each of these measures makes the book more difficult to actually read and use in legitimate ways. Similar measures are routinely taken in the digital realm, and just as with physical media, it is necessary to decide what measures are appropriate to the nature of the information in question and the situation in which it is to be used.

In the domain of digital librarianship, these kinds of tradeoffs crop up quite often. Here are some examples:

- A digital reference service has among its goals maximizing patron confidentiality. To ensure that digital reference transactions cannot be listened in on, one could require the use of encrypted channels to contact the library. All newer web browsers are capable of this, and it would certainly provide an added measure of security. On the other hand, this policy would prevent people who use older web browsers or chat software that does not have encryption capability from making use of the service.
- A database of questions and answers is made available on a website, and one of the goals is to prevent wholesale copying of the contents. It is possible to set up the pages so that users are prohibited from saving or printing the pages or copying the text to the clipboard. (As we will see below, clever users will always be able to get around this, but many would be stymied). On one hand, this would make it much more difficult to copy large sections of the database. But on the other hand, it would seriously limit patrons’ ability to make use of

small portions of the information in legitimate ways, requiring them to re-type it laboriously instead of simply copying and pasting.

- Digital reference transcripts and personal information about patrons are stored in a database, and one of the goals is to ensure that only staff have access to this information. The simplest way of approaching this goal would be to require a username and password in order to access the database. More sophisticated approaches might include disallowing access from outside of the local area network, making use of hand-held devices that generate ever-changing passwords, or even more exotic technologies such as fingerprint recognition. Each of these would provide more security than a simple username/password scheme. On the other hand, they all entail additional expense and make it more difficult for the staff to log in when and where they need to.

### 3.3 Viewing information involves copying

Perhaps the most important aspect of specifically *digital* information from a legal point of view is that every time that information is accessed, a copy is implicitly made. This has been recognized and legitimized by courts in the U.S., Canada, and Europe, and is true regardless of the type of information and the manner of access. For example, when you view a web page in a browser, the server sends the contents of that page across the network to your computer. This information is then stored in your computer's memory, constituting a copy of the information stored on the server. The same is true when accessing any other type of file, such as audio or video. In the case of "streaming media", your computer may display the information to you bit by bit as it comes in, and may discard each segment after it is displayed. However, aggregated over time, copies of each part of the file are retained in your computer's memory.

This is important for the following reason. Many schemes for maintaining control over how information is used involve restrictions built into the software used to view it. For example, someone who views streaming video over the web may be disallowed from saving it to disk. Someone viewing a web page or a PDF file may be disallowed from printing the information or copying the text to the clipboard. Someone who downloads a music album may be allowed to burn only a limited number of CDs. These kinds of restrictions form an integral part of the modern regime of intellectual property protection. They are often specified in licenses, and are thus subject to legal enforcement, but are also enforced through technical means.

Unfortunately, these technical means are not as infallible as their developers would like. The essential problem is that the client computer has to receive an actual copy of the information in order to display it. Once that information is on the client computer, the possibility exists to either fool the viewing software into not applying its built-in limitations, or as a last resort use (or write) some other piece of software that will display the same information and also provide more functionality. Once the information leaves the server, there is *nothing* that can be done to guarantee that it will not be used and copied arbitrarily. Even encryption is useless<sup>iii</sup>. Recent experience has shown that resourceful people are capable of figuring out how to circumvent even the most complicated schemes developed by content providers. Here are some examples:

- Many web page limitations (on saving, printing, and so forth) can be circumvented by using purely text-based browsers such as Lynx.

- File-sharing aficionados publicly post tips on how to trick one's music software into writing more than the permitted number of CD copies of downloaded music.
- A program called "deCSS" was written in order to play DVDs on Linux systems, and has the side effect of allowing its users to make copies of those DVDs.

From a technical point of view, the only way that an information system can keep control over what can be done with the information it provides is to decide whether or not to send it out in the first place (i.e. by means of access control). Once that decision is made, all bets are off. It is because of this fundamental technical limitation that the DMCA<sup>iv</sup> was enacted. This legislation makes illegal just such circumvention measures as I have described, and attempts to reestablish by means of legal sanctions the ability to control how information is to be used once it gets to a client computer. Unfortunately, as noted above, there is a large segment of society which disregards this law and actively supports the development of technical means for circumventing such controls. Until the legal and/or social climate changes, we should always assume that such means exist.

### **3.4 Access to information is almost always granted or denied based on a mechanical procedure**

Whenever information is accessed from a remote system over a network, some kind of request message must be sent to that system. This is done automatically for you when you click on a link in a browser window, hit the "enter" key after typing a database query, or enter somebody's "handle" into an instant messaging application. Once it gets the message, the system (i.e. the server) must decide, based on the message's contents and context, whether or not to send out the requested information.

Since the system lacks human judgement, it must rely on a strict and unvarying set of rules to make this decision. In a completely open system without access control, the message need specify nothing more than which piece of information is desired. On the other hand, if access control is being used, the request message will typically include *authentication* information such as a username and password, on which basis the system can differentiate between valid and invalid requests.

The phrase "almost always" in the title of this section is there because it is possible in some cases to use a person to make the judgement as to whether to grant or deny access. However, this is not possible in the vast majority of cases, since the capacity of a person to make such judgements would be generally outstripped by the number of requests coming in to even a medium-sized information system. This lack of human judgement is important for the following reason: there are so many different ways to fool an information system that it is generally thought to be impossible to prevent them all. In other words, there is no such thing as an absolutely secure system. Depending upon the situation, it may be possible to try many passwords until the right one is found, to make requests of the system and study its responses to figure out how to attack it, or to use a "back door" that has been accidentally or purposefully left in the system by its designers (these are appallingly common). Even more common, studies have shown that it is terribly easy to obtain authentication information from authorized users by means including stealth, misdirection and subterfuge.

The most unfortunate thing about this kind of vulnerability is that depending upon how thoroughly a system is compromised, the unauthorized user may have the ability to not only access information but to modify it. This in our own domain this might include, deleting information such as reference transcripts or staff records, altering the saved questions and answers, or even pretending to be a member of the staff and interacting with patrons.

### **3.5 References are themselves information**

Now that we live and work in a world shaped by the Web, references to information are a crucial part of what we do. Whereas in the pre-Web era, a reference was simply a set of instructions to a human being as to how to find a particular piece of information, today a URL is almost as good as the information itself (almost, because the information itself may be protected by access control, or may have been moved or deleted). These URLs are themselves information, subject to the same rules as any other piece of information.

This has a number of implications, but the most important one is this: once you publish a URL or other digital reference, you can't control who will come to have a copy. You can subsequently disable that reference, either by moving or deleting the information or by adding access control, but you cannot arrange that some subset of the world has access to the reference and the rest do not. If this is unacceptable, the only solution is to arrange that none of your information is reachable via persistent URLs. In other words, you must arrange that every time someone searches or browses your site, a new URL of limited lifetime is generated. This can be done (at the cost of additional expense in setting up your web server) but as with the other measures discussed in this paper comes with a tradeoff in increased cost and decreased usability.

### **3.6 Summary**

We can summarize the contents of this section as follows:

- It is impossible to guarantee by purely technical means that digital information will not be copied, although there are various measures that can make this less likely.
- It is impossible to control by purely technical means the ways in which information is handled once it has been transferred to a client system (i.e. printing, saving, copying).
- It is impossible to guarantee that a system cannot be broken into.
- It is impossible to keep control over who has a reference and who doesn't, once those references are made known.
- Control over information flow, and increased certainty about information integrity, always comes with a cost. This is both in the expense of setting up and maintaining the system, and in usability.

## 4 Classification of legal issues

Based on the technical factors covered in the previous section, we can now return to the classification of legal issues presented in section 2. For each aspect, we will consider the various risks to information integrity, and the measures which can be taken to guard against (though not prevent) them. For the purposes of this discussion, we use the term *library* to indicate an organization that operates a digital reference service (although it may not be a library in the canonical sense of the word) and *patron* to indicate someone who makes use of that service.

### 4.1 Types of integrity violation

Digital reference information is subject to several different kinds of integrity violations. These have different properties, and require different kinds of measures to guard against.

#### 4.1.1 Disclosure

Perhaps the most serious type of integrity violation is *disclosure*, which refers to information being seen by someone who is not authorized to do so. Violations of patron and staff confidentiality fall under this heading, as do unauthorized viewing of other information such as non-public question/answer databases. The greatest difficulty in dealing with disclosure violations is that they are difficult or impossible to repair. Once information has “leaked out” of the system, there is often no way to determine who has seen it, let alone prove that fact in any kind of legal sense. Even if this is possible, any resulting damage will in all likelihood already have been done. For this reason, preventing disclosure violations is perhaps the highest priority.

Disclosure violations most often arise from failure of access control; this may be a deliberate attempt to break in to the system, or it may be due to an error on the part of the system administrators who mistakenly leave certain information unprotected. These vulnerabilities are under the control of the library staff, so that the organization might be exposed to liability if violations occur. However, even if these are adequately protected against, there are many other places at which violations might occur, in ways that the staff have no control over. These are discussed more fully in section 4.3 below.

A less serious kind of violation which would also fall under this heading occurs if a patron is given access to some information from a licensed database during a reference transaction (i.e. through co-browsing) and uses that access to gain access to other information to which he is not authorized. This is also a matter of access control, and can be mitigated mainly by paying careful attention to the design of co-browsing software.

#### 4.1.2 Copying

The most widespread type of integrity violation is certainly *copying*, by which we mean unauthorized reproduction of information. Such acts are legally, in most cases, violations of copyright. Some violations of this type will also be disclosure violations, while others may not. This depends upon whether the information in question is or is not intended to be public. In the context of digital reference, the most important example of a copying violation would be the unauthorized reproduction of licensed materials made available to users as part of the answer to a

question. Another example would be the unauthorized reproduction of questions and answers from a knowledge base.

As was noted above in section 3.3, it is impossible to prevent patrons from copying any information sent to them. The only to control such activity is to not send out the information in the first place. As a result, the focus of attempts to limit copying violations involving licensed materials should probably focus on ensuring that nobody but the library staff and the patron who asked the question has access to the answer and any accompanying materials. Again, this is essentially a question of access control.

### 4.1.3 Modification and Masquerading

These types of integrity violations are substantially less common than the previous two, although if they do occur they indicate deep-rooted security problems which must be urgently addressed. The violation of *modification* refers to the unauthorized modification or removal of stored information, such as questions, answers, transcripts, personal data and so forth. The violation of *masquerading* refers to someone successfully pretending to be a librarian and interacting with clients on that basis, or substituting their own web pages or other information from those which are supposed to be available through the library's server. These kinds of violations are so very serious because they represent a violation not only of the integrity of individual pieces of information but of the integrity of the digital reference service itself.

Fortunately, these kinds of integrity violations are more difficult to carry out and, because they have detectable effects, difficult to conceal. The best way to guard against modification violations is, once again, careful attention to access control. Masquerading violations are more insidious, because they can be carried out without any access at all to the library's computer system. For example, there are ways in which a patron's computer system can be tricked into contacting a different website, which appears to the patron to be the regular library site but is actually under the control of a malefactor. There is not much that can be done in advance to prevent these kinds of attacks, but at the present time they appear to be mostly directed at financial institutions. When and if libraries ever become subject to them, the best remedy will be careful vigilance.

## 4.2 Types of information

The different kinds of information involved in the provision of digital reference service are subject to different vulnerabilities.

### 4.2.1 Personal information

Personal information about patrons is most vulnerable to disclosure violations. Disclosure of such information violates the principle of confidentiality. Even the mere fact that a patron has asked a question of a digital reference service is considered to be confidential. As noted above, guarding against these kinds of violations is mainly a matter of careful attention to access control. As an extra measure of security, many libraries have adopted policies of discarding such information after a short period of time, even if they keep the contents of questions and answers.

## 4.2.2 Contents of questions and answers

Questions and answers are certainly vulnerable to disclosure violations, which if they include personal information represent a breach of the principle of confidentiality. In addition, when stripped of personally identifying information, they become a valuable resource in their own right. As such, they are subject to copying (copyright) violations. If large collections of questions and answers are made public, it is important to remember that technical means alone cannot prevent them from being copied. Only non-technical means such as the threat of legal sanctions will have any effect.

## 4.2.3 Other materials provided to patrons

Materials provided to patrons in the context of answering a question are typically licensed to the library by information providers. The library therefore has a duty as licensee to uphold the information providers' copyrights. Such information can be provided to the patrons in two ways. Either it is sent directly as an attachment, or the patron is led to it via co-browsing. In either case, as has been noted many times in this paper, there are no technical means which will prevent patrons from copying this information once they receive it. Therefore, if such copying occurs it becomes a purely legal question involving the amount of material in question and the terms of the license under which the library is authorized to provide it.

There is one aspect in which technical measures are important, however. If a record is kept of completed reference transactions (as is almost universally done with digital reference), copies of any materials sent as attachments to the patron are likely to be kept with it. This information is therefore vulnerable to disclosure and copying violations, and measures such as those described above should be taken to avoid this.

## 4.3 Locations of concern

In the process of digital reference, information flows between a number of different locations, and there are thus many places where information integrity violations can occur. Some of these are under the control of the library staff, but many are not. There is an important axiom of computer security that is important to note: physical access trumps everything else. In other words, anyone who has physical access to a computer system or network can generally circumvent any measures of protection that have been put in place.

### 4.3.1 Client computer systems

Any digital reference interaction involves, by definition, at least two different computer systems (and in many cases three or more). Except in the special case where the patron is physically sitting inside a library facility while using that library's digital reference service, some of them will not be under the control of the library staff. For the purposes of discussion, we will call this system or systems the "client computer system".

Not only is this client computer system not under the control of the library staff, some or all of it may not be under the control of the patron either. A patron who is connecting from a hotel, cafe, school or other facility must depend upon the integrity both of the computer she is using and the staff who maintain it. A patron who uses a third-party e-mail service such as AOL, Hotmail, or

an e-mail account provided by their Internet Service Provider must also depend upon the integrity of this system.

In fact, even the patron's own computer system, or one made available honestly for public access, may not be safe. At the present time, improperly secured computers are often infected within minutes of being connected to the Internet by viruses, "trojan horses" and other malicious software. Many of these are designed to capture information being sent to and from the computer, and forward it elsewhere. Others are capable of fooling the user about what website they are actually connecting to, substituting a connection to a different web address<sup>v</sup>. These risks are very real, and unfortunately there is little that the library can do to about them, other than urging their patrons to be careful. As noted above, once information is sent to the client computer system, there is no way for the library to control what happens to it.

Another potential for integrity violation comes from the fact that web browsers are typically designed to retain copies of the information they are given, such as web pages, attachments, passwords, and so on. This is called *cacheing*, and it is generally used to improve performance. If a user asks the browser to display a piece of information again (i.e. going back to a previous web page), the browser can quickly display its local copy rather than requesting the information from the server again and waiting for it to arrive over the network. Cacheing is almost universally used, which threatens the integrity of any information viewed by a computer that the patron does not control. In order to protect the integrity of any information they receive or send on a public computer, a patron must "clear the cache" when they are done. Unfortunately, there is little that librarians can do about this other than remind their patrons to do so.

#### 4.3.2 In transit

In addition to the risks of integrity violation inherent in the use of many client computers, there are also risks involved in the use of public and private computer networks to communicate between patron and librarian. Again, except in the special case where the patron is using one of the library's own computers, all information sent to and from the patron must go by way of the public Internet.

In the case of a patron who is using a computer in a location such as a hotel, the information is sent through the local area network. During this transit, the information is potentially subject to interception, unless encryption is used. Thus, media that cannot be encrypted, such as Instant Messaging, are especially at risk. Fortunately, most people who find themselves in such situations check their e-mail via a web-based interface. Most e-mail services provide secured (i.e. encrypted) web access for this very reason. Similarly, if a library provides secured web access to their digital reference service, it can be used safely even over networks whose integrity is not certain.

#### 4.3.3 Library server

Maintaining the integrity of the information stored on the library's own systems is primarily a matter of access control. As long as only authorized users are able to access the system, the integrity of its information is preserved. On one hand, the primary responsibility for maintaining a secure computer system rests with the library staff. On the other hand, as we noted in section

3.4 above, it is impossible to guarantee that a system cannot be broken into. The library staff must decide how much of their resources to put into maintaining computer security, based on the level of risk they are comfortable with. This is, unfortunately, a difficult and uncertain decision. In the modern computing environment, threats and responses can change on a daily basis, and many kinds of risk have never been quantified. It is difficult or impossible to determine that actual level of risk that a given system will be compromised.

There is a potential point of information integrity violation that is often overlooked when security procedures are being designed, and this is the existence of back-ups. These are, by design, a copy of the information stored by the library's computer systems. If they fall into the wrong hands, the integrity of that information is very easily compromised. It is important in general to make sure that these are either kept physically secure, encrypted, or preferably both.

One interesting strategy that is used to maintain the integrity of personal information stored on a library's computer system is to destroy that information before any violations can occur. This strategy is used especially for personal information about patrons, which many libraries discard within a few days of when the patron's question is answered. This strategy is useful even when the questions and answers themselves are kept, stripped of personal information and anonymized. Overall, it is a very useful procedure, since the only guaranteed way to prevent disclosure violations is to delete the information completely. However, one must be careful to do this properly. As with any other kind of information, even a single remaining copy defeats the purpose. Purposely deleting some of the information goes against the usual procedure of making thorough back-ups. The system must be very carefully designed to make sure that everything but the personal information gets backed up, but the personal information never is.

It is also important to note, again, that control can be maintained over information by technical means only so long as it remains on the library's computer system. Once it is sent to a client computer, there is no way to prevent the user of that system from copying it and using it as they choose.

#### 4.3.4 Third party server

An interesting situation occurs when a library contracts out its technology operations to a third party. Many digital reference services are structured in this way, with a central server that is used by a number of different libraries. This may be organized as a single collaborative service, with the library staffs sharing question-answering duties, or a single server may be used to support multiple independent service points. In either case, the staff from each library logs into the central server, which acts as a switchboard connecting patrons to librarians.

The same kinds of vulnerabilities occur as with a library's own systems, but the risks are multiplied. If someone is able to compromise the central server, the integrity of everyone's information is at risk. The collaborative model also introduces another degree vulnerability. If the userid and password of any member of any library's staff is compromised, that also puts at risk the integrity of the entire collaborative's information. In this regard, the independent-service model is safer.

In fact, the use of third-party servers and collaborative service arrangements represents yet another tradeoff between security on one hand and cost and usability on the other. By making such arrangements, libraries are able to lower their costs, and patrons are given greater access to the service. The down side is an increased risk that an information integrity violation will occur, and a greater potential for harm if one does happen.

#### **4.4 Agents of concern**

##### **4.4.1 Outsiders acting illicitly**

The stereotypical “bad guy” in most computer security scenarios is an outsider. Whatever their motivations, they are intent on committing violations of information integrity. We have already discussed the means by which they may attempt to carry out their nefarious plans, and the ways by which the champions of truth and justice can attempt to thwart them (although as in all epic battles the outcome is always in doubt until the last moment). In all seriousness, the concept of an unknown malefactor acts simply as a placeholder in most discussions of computer security. Since the possible set of people and motivations is so vast, there is no way to characterize the most likely type of attacker, or the most likely methods for them to use. The biggest variable is this: does a potential malefactor have any kind of advantage that would help them violate the integrity of our system? Possible advantages include: knowledge of our systems and procedures, possibly gleaned from public information or from searches of our trash (yes, really!), personal relationships with our staff, physical access to our building, and numerous others.

##### **4.4.2 Patrons acting illicitly**

By contrast, the potential set of integrity violations that can be committed by patrons is much smaller (this assumes that the people in question are actually acting as patrons and not trying to break into the system; in that case, we would class them as outsiders). Patrons cannot by definition violate the integrity of their own personal information, since they already control it. And, in the course of an ordinary interaction with a digital reference service, they do not have access to anyone else’s. What is left, is the possibility of illicitly copying the information they receive. We have already discussed these vulnerabilities in section 4.2.3 above.

##### **4.4.3 Staff members acting illicitly**

The biggest unknown in any discussion of computer security is the possibility of compromise by an insider, i.e. a member of the organization’s own staff. One would hope that this is rare or nonexistent in libraries, and I think that this is a reasonable assumption. The potential financial rewards for violating the integrity of digital reference information are tremendously smaller than those of compromising the information held by a bank or a store, for example, while the risks of getting caught are just as high. This alone should be a sufficient deterrent in almost all cases.

##### **4.4.4 Government agencies acting with legal authority**

In addition to those we have discussed so far, another potential for the compromise of information integrity exists which is unique to libraries and similar organizations. According to many in our profession, our standard of ethics implies that the release of personal information about our patrons to the government may in some circumstances constitute a violation of

information integrity. This may well include cases in which the government has a legally valid subpoena. Given that the library has no legal option but to comply with such subpoenas, the best measure that can be taken to avoid this possibility is to completely delete the information in question as soon as is practicable, as was discussed in section 4.3.3 above.

At least two Federal agencies, the FBI and the NSA, are known to have the capability of scanning e-mail as it is sent across the public Internet. Both agencies refuse to reveal the extent to which they are actually using this capability, and it is not known whether they are also scanning other kinds of Internet traffic (or even whether they have the technical ability to do so). E-mail is particularly vulnerable because it is almost never encrypted, and because e-mail traffic is a very small percentage of total internet use. This latter means that the total amount of information to be scanned is much smaller than for other kinds of traffic such as web page accesses or peer-to-peer file sharing.

## **5 Tools for digital reference**

A number of different technologies are in common or potential use for the provision of digital reference service. Each of these are vulnerable of information integrity violations in ways peculiar to their own nature, and they also share a number of common vulnerabilities. It turns out that the vulnerabilities to disclosure violations (especially those relevant to patron confidentiality) vary quite a bit from technology to technology, whereas the vulnerability to copying violations is consistent.

### **5.1 Vulnerability to violations of patron confidentiality**

#### **5.1.1 E-mail**

This is the oldest method used for digital reference, and the simplest. Everyone has access to e-mail (through free services such as Hotmail if necessary) and it is the mechanism least likely to be interrupted by technical problems. It is also by far the best method for *asynchronous* digital reference, for example when the librarian must take some time to do research before sending an answer to the patron. In fact, nearly every digital reference tool in use currently includes an e-mail component for this very reason. In many cases, e-mail functions as a medium for information storage as well as transmission. Most of us keep a lot of important information in our stored e-mail, and so do many organizations.

Unfortunately, the very simplicity and ubiquity of e-mail makes it weak at protecting the integrity of the information it transmits and stores. Because it is almost never encrypted, it is vulnerable to integrity violations while in transit. Most people rely on third-party systems to store their e-mail, either an e-mail account made available by their Internet Service Provider or a widely available service such as a Hotmail or Fastmail. While these have a reasonably good record so far for maintaining the integrity of client e-mail, there are certainly no guarantees.

As noted in the previous section, government agencies (the NSA and FBI, possibly others) are known to have the capability to monitor public e-mail traffic, although they decline to reveal the extent to which they are using that capability. Another interesting vulnerability is that “spam-blocking” software could mistakenly prevent specific patrons from sending and/or receiving e-

mail from a library. This latter does not pose any risk to information integrity, but it does constitute a potential “denial of service” problem.

Far worse than these problems, however, is the fact that standard e-mail client and server software, especially the widely used ones provided Microsoft, is terribly insecure. They are vulnerable to viruses, and also provide avenues which can be used by attackers to break into the systems they are running on. E-mail server software which runs on Unix or Linux is much more secure, but there are still vulnerabilities reported in such servers on a regular basis.

In any case, the only way to avoid these potential violations of information integrity is to avoid the use of e-mail in digital reference transactions. This would make asynchronous digital reference much more difficult, and would make the reference service unavailable to people who lack access to more sophisticated technology. Once again, we face a tradeoff between ease of use and security.

In the course of being transmitted, received and stored, each e-mail message is copied many times. This is done automatically by the different pieces of software that are involved in the process of handling e-mail, and there is really no way to avoid it. Most of these copies are destroyed immediately, but it is important to pay attention to the back-up policy that is used on your organization’s e-mail server. Depending upon when the back-ups are taken, they may include copies of e-mail messages in transit; these copies will then be in existence as long as the back-ups are.

Any digital reference software package that sends and receives e-mail makes use of standard e-mail server software (and of course, of whatever e-mail client software their patrons use). These, then, share the same vulnerabilities noted above.

### 5.1.2 Web-based tools

The digital reference tools most widely used today provide interfaces that work via client web browsers, communicating via the HTTP protocol. They are inherently more secure than e-mail, because of the possibility of using SSL connections that cannot be intercepted (see section 3.1.3). I’m not sure to what extent these products actually have SSL capability, but the fact that the matching capability is built into almost all web browsers means that it could be added to the servers as a seamless upgrade.

It is the case, however, that any of these products that do not use SSL, or which are used by patrons who do not have SSL-capable browsers, are still subject to the vulnerabilities discussed in section 4.3.2 above. In addition, they are all subject to the vulnerabilities inherent in the client computer systems to which they talk, as discussed in section 4.3.1.

### 5.1.3 Instant Messaging

Several libraries are currently exploring the use of the Instant Messaging (IM) systems for digital reference. These are widely available, and especially popular with young people. IM is used in two ways: either via software running on ordinary computers, or via cell phones, palmtops and other specialized devices. Communication using the former mode is potentially vulnerable in the same ways as web-based systems, especially since IM traffic is not encrypted. The second mode

is more secure, since the client devices communicate through proprietary wireless channels rather than the public Internet. In either case, risks of disclosure violations on the client side are less because IM traffic is not typically cached.

#### 5.1.4 Internet Telephony

Internet Telephony (VOIP) technology is still in its infancy, and is not yet used as a primary communication medium for digital reference (except inasmuch as calls to a traditional telephone reference might use Internet Telephony on the patron's side). The legal status of this technology is still in flux, especially with respect to whether the companies that provide these services have "common carrier" obligations, and the conditions under which the government can intercept calls. Until these issues are settled world-wide, the legal implications of this technology are still uncertain.

## 5.2 Vulnerability to copying violations

### 5.2.1 Attachments

Almost all of the technologies used in digital reference allow the inclusion of "attachments", a term which refers to additional pieces of information included with the main text of the answer. These might be photographs, sound files, formatted documents, or anything else relevant to the question and answer.

The legal implications of this technological feature depend upon the nature of the attachments sent. Those that are in the public domain, or are otherwise freely usable, pose no problems. However, librarians may also wish to send selections from licensed resources as attachments. In this case, the potential for copyright violation exists on the part of either the patron or of anyone else who may have access to a copy of the attachment.

As discussed in section 3.3 above, there are no technical solutions that can remove this vulnerability. There are ways of making the copying of information contained in attachments more difficult (which also make it more difficult for the patrons to use this information in legitimate ways), but no means of preventing copying absolutely. This is true of all attachments, no matter what digital reference tool is used to send them, and no matter in what form or medium the information is expressed. To the degree that closing this vulnerability is a priority, legal and social means must be used instead.

### 5.2.2 Co-browsing, database access

Many digital reference tools use "co-browsing" technology, that allows the librarian to guide the patron in exploring information sources that may be relevant to the patron's problem. Under this facility, each web page viewed by the librarian is mirrored on the client's browser, in such a way that both the client and librarian can click and type to navigate. Similar technology that only provides for control by the librarian is known as "page pushing". In either case, the legal implications depend upon what kind of information is being viewed (free or proprietary), whether or not the patron is a member of the library's community of service, and on license terms under which the library has access to the information resource(s) being viewed.

The problematic cases are those in which the information being viewed as proprietary, the patron would not normally have access to these resources (i.e. because the patron is not a member of the library's service community) but the license terms allow the librarian to share limited amounts of information with the patron. In this case, the technological challenge is to ensure that when the digital reference session is finished, the limited access provided to the patron terminates. Otherwise, the patron could continue browsing and get unauthorized access to additional parts of the database, which under the license terms would constitute an information integrity violation.

## 6 Conclusions

- If there's one thread running through this paper it's that there are no magic bullets. One of the axioms of technology is that technology does not solve problems-- it merely creates new and different ones.
- Without the support of non-technical factors such as laws and social attitudes, nobody can guarantee the preservation of information integrity. Every purely technological solution has its points of vulnerability.
- That said, it is important to understand technology and to keep current on the latest identified threats and solutions. Otherwise, you risk being subject to newly identified and exploited vulnerabilities, and thus to potential liability.
- Although technical means cannot completely eliminate vulnerabilities, they can make it difficult to exploit them and easier to catch people who do so.

## Notes

<sup>i</sup> While the simple act of copying information leaves no evidence, any attempt to actually *use* the information, including selling or giving away one or more copies, will create external effects on the world at large. In many cases, the effects are eventually significant enough to be noticed by the relevant enforcement authorities

<sup>ii</sup> Note that the same is true in the physical world: consider a theft from an unlocked building versus a locked one. In both the physical and virtual realms, taking something from a locked place is punished more harshly than taking something from an open one.

<sup>iii</sup> The information may well have been transmitted in some kind of scrambled form, but at some point the client software must unscramble it in order to display it on the screen, play it through the speaker, or otherwise convey it to the user. In other words, at some point either a clean copy of the original information must exist in local memory, or a scrambled version exists which must be unscrambled in order to display it. If this scrambling involves cryptography (see section 3.1.3 above), then the key must be either stored as part of the client software or transmitted with the information in a non-scrambled form. If cryptography is not used, the unscrambling must be

---

done using some series of simple steps which are carried out in order by the program. In any case, it is typically not very difficult to figure out how to replicate this process.

<sup>iv</sup> *Digital Millenium Copyright Act*, passed by the U. S. Congress in 1998.

<sup>v</sup> If, hypothetically, one of these were specifically designed to attack a library website, it could allow a malefactor to masquerade as a member of the library staff. I am not aware of any such attacks on libraries to date; financial institutions seem to be the primary targets of such attacks.