Reader Privacy: The New Shape of the Threat

Clifford A. Lynch, Executive Director, Coalition for Networked Information

Introduction

This essay briefly summarizes the current range of threats to reader privacy and makes some high-level suggestions that research library leadership might consider to address them. It is not comprehensive, and does not go into much technical detail; those interested in a place to start might see my paper "The Rise of Reading Analytics and the Emerging Calculus of Reader Privacy in the Digital World,"[1] keeping in mind that it's now two years out of date.

I also note recent projects funded by the Institute of Museum and Library Services intended to provide guidance for libraries of all types: Library Values and Privacy in Our National Digital Strategies[2] and the National Forum on Web Privacy and Web Analytics,[3] as well as a very recent and welcome statement of principles in a posting by Mimi Calter of Stanford University on *The Scholarly Kitchen*.[4] These may also be helpful.

Fundamental Reader Privacy Threat Scenarios

Threats to reader privacy fall into three major categories.

The first is eavesdropping on the interactions between a reader and various systems that help the reader to discover and obtain information. To a first approximation, in the digital world this can be effectively addressed by routine (but properly configured!) encryption of such interactions. Surprisingly, as recently as, say, four years ago, implementation of this strategy was relatively rare and libraries had been slow to demand it from vendors. Today such encryption is becoming increasingly commonplace, particularly in research libraries. I shall not consider this further here.

The second threat is disclosure of information that the library is holding about what a patron is reading, perhaps through legal mechanisms (subpoenas or national security letters, for example), or because the library is hacked; it might even be due to accidental misconfiguration of a library system. Aspects of this threat have been a concern since long before libraries computerized their operations—and that was a long time before digital content became dominant.

The third category of threat, which is new to the age of digital content, involves data that is collected by **external** vendors who provide licensed content to libraries (and indeed, also external suppliers of content that is "freely" available, subject to click-though terms and conditions). This is, in my view, the least understood and most dangerous threat to reader privacy today.

Disclosure by the Library

Libraries have addressed this threat on several levels. The first is a recognition that they can't disclose information that they don't have, so they have typically collected as little as possible, and retained it for as short a period as possible (for example, only while a book is out on loan is the loan tracked)—notably, often, with the exception of special collections. The second is to be as rigorous as possible in defending disclosure of information that they do hold, particularly legally. I am less confident that library systems are subjected to the same kind of periodic and rigorous security requirements and audits that are now commonplace for various kinds of enterprise administrative systems; these are expensive and time-consuming, and also tend to increase the overhead costs of running systems. This is an area where at least an exploratory conversation with your IT leadership may be informative.

It can be very challenging to be confident that you are collecting as little information as possible and retaining it for as short a time as possible. There are backups, and there are logs at various levels. Even if you are anonymizing logs it may be possible to re-identify them in various ways, so one should be very cautious about relying on anonymization.

Finally, it's important to recognize that taking an absolutist approach to information collection, as opposed to more nuanced, transparent, and opt-in collection of data about user activities and interests, has meant that library systems appear to the user as far inferior to commercial offerings; they are unable to make recommendations to users, or to remind them of past history. I believe that re-assessing these choices is long overdue, but doing so will further demand that libraries carry out a much more complex and subtle risk assessment; it will also challenge them to convey the implications and risks of various choices to their patrons.

> "I am less confident that library systems are subjected to the same kind of periodic and rigorous security requirements and audits that are now commonplace for various kinds of enterprise administrative systems."

Collection by Third Parties

Third-party platforms offer access to various databases or collections of content such as journal articles. The platform providers know a tremendous amount of information about **what** is being read, and the patterns of reading. Particularly to the extent that they can associate this information with **who** is doing the reading, they have frighteningly detailed data. Even if they can't associate it with a given individual by name, there is still potential power in knowing that someone at a specific institution, or perhaps in a specific department within that institution, is following a specific trail of information over time.

These platform operators can do various things with the information they collect, potentially: in addition to using it for their own purposes, they could share it with others, or resell it. Furthermore, it is subject to disclosure—by legal means, by hacking, or by human error. There are no a priori limits to how long this data can be retained, and normally, if control of a company changes (through acquisition or bankruptcy, for example), the data is just one more corporate asset.

It's also worth noting that there are really two layers of attack on privacy on these external platforms. Not only can the platform operators collect data themselves, but they also sell advertising in most cases, which means that they are also contributing participants in the gigantic internet surveillance apparatus for monetizing users, in much the same way as online newspapers (which also use a mixed advertising and subscription model).

For licensed resources, language in contracts can address all of these issues: limit or forbid data collection, retention, reuse, and redistribution; include criteria about how that data is protected, both legally and technically; forbid third-party advertising.

However, some limited surveying suggests that, as of at least two years ago, contract language dealing with these points was relatively rare. It is also unclear how resistant content providers are to such language. An examination of the posted privacy policies of some of the major content providers does not inspire confidence in the absence of specific overriding contractual stipulations. It would be very useful to have some more current data about research library contractual practices in this area, and perhaps also to have model language available. In situations where vendors will not accept contract language regarding reader privacy, institutions will need to make choices about what minimal levels of privacy assurance are acceptable before they walk away.

"In situations where vendors will not accept contract language regarding reader privacy, institutions will need to make choices about what minimal levels of privacy assurance are acceptable before they walk away."

For resources that the library does not license (but that their patrons may rely on for various purposes), there's not much that the library can do other than help their users to understand terms and conditions, privacy policies, and risks. But doing this is an important part of improving digital and information literacy.

It's also important to explicitly recognize a large class of online learning materials and electronic textbooks as potential environments for massive data collection as well. Here, historically, the library hasn't been involved in the licensing process or terms, and it's often extremely unclear whether student privacy is even being considered, much less protected, or whether data that would be helpful to the university for various reasons is actually being made available to the institution (or what happens to it if it is made available). Unlike research materials, students often have no choice about using these educational resources. My expectation is that, for many reasons (escalating costs, privacy liability, the uptake of open educational resources, etc.), libraries are going to become much more involved in these arrangements going forward. They have a great deal of expertise to bring, not just in privacy but also in other areas, such as preservation and archiving. The current ways in which most institutions select and contract for these resources is deeply problematic and overdue for re-examination.

One of the biggest questions in understanding data collection by third parties is whether they can identify individual users accessing their platform. Even today, a great deal of authentication of users is done via proxy servers, which verify that a user is a member of a given university community and, once validated, pass that user (and all other validated users) on to external services from a common address known to the external service as only sending validated traffic.

Ever since proxy servers came into use in the 1990s (indeed, there are forms of proxies that pre-date the web), there has been a belief that this process effectively anonymized traffic to external services and, hence, rendered the reader privacy issue largely moot as long as proxies were employed. This was probably at least generally true in the early days of the web. More than 20 years later, the various technologies for user tracking and re-identification have advanced greatly, fueled by the demands of various advertising and data collection platforms. It would require a very careful, determined, and sophisticated user to have much hope of avoiding tracking and re-identification today, with or without

the intercession of a proxy service—hence the need to address the problem contractually.

> Libraries are going to need to think very carefully about what data they want to collect and what risks it represents.

There are other authentication technologies in use. Most notably, there is Shibboleth, which many universities use with major content suppliers. Here, the data that the institution passes to the third party about a given user is determined by the institution's attribute release policies. There have been instances where institutions were releasing very specific, individually identifying information to external platforms as standing policy. If your institution is using Shibboleth to handle authentication for licensed content, it's vital that you understand the details of this attribute release policy and that your users understand it as well. If you've not had this conversation with your institution's IT policy leadership, it's past due.

Note that the experimental RA21 initiative is really, as I understand it, just an effort to make Shibboleth a bit less cumbersome to use. From a reader privacy perspective, it's no better and no worse that the local Shibboleth implementation, though I know it's been viewed by some with considerable suspicion.

Library Analytics: An Emerging Dilemma

As the costs of digital content continue to increase and library budgets are stretched, it's very valuable for libraries to have good data about what's actually being used, and who (individually or demographically) is using it. Libraries are also being pressured to demonstrate impact, particularly with regard to student outcomes. Indeed, there have been some uncomfortable conversations between institutional leadership determined to develop the most powerful analytics for predicting student outcomes, and library leadership unwilling to collect and supply some of the data that the analytics developers would like to have.

Libraries are going to need to think very carefully about what data they want to collect and what risks it represents. Then they will need to consider how to inform their users about what is being collected, how it is being used, and where the data collection is going to happen. They may need to share some additional information (role, school, or departmental affiliation, for example) about users with external platforms if they want the platform to return usage data faceted by those attributes. Emerging techniques and technologies, such as differential privacy, may ultimately prove very helpful here.

The Most Important Steps to Take Now

This brief paper suggests many issues that library leadership needs to consider with regard to reader privacy, but three stand out to me as most urgent:

1. If you are using Shibboleth for authentication to external content platforms at your institution, be sure that you understand your institution's attribute release policy and the governance around the development and maintenance of that policy.

2. As new licenses for content and services are established, or as existing ones are renewed, add language dealing with reader privacy as a routine matter.

3. Develop a strategy and a program for informing and educating the university community about reader privacy issues broadly. In my view, this is ideally done by the library in partnership with other organizations (such as information technology, general counsel, registrar, instructional technology, etc.) in a coordinated and holistic way. In any event, it's essential that this communication be put in place sooner rather than later, even if the library must act alone for a time while an effort is being made to develop a more strategic institution-wide conversation about the issues.

Endnotes

1. Clifford A. Lynch, "The Rise of Reading Analytics and the Emerging Calculus of Reader Privacy in the Digital World," *First Monday* 22, no. 4 (April 3, 2017), https://doi.org/10.5210/fm.v22i4.7414.

2. "Project Report: 'Library Values & Privacy in Our National Digital Strategies: Field Guides, Convenings, and Conversations,' " Center for Information Policy Research, University of Wisconsin–Milwaukee, August 2, 2018, https://cipr.uwm.edu/2018/08/02/project-report-library-values-privacy/.

3. "National Web Privacy Forum," Montana State University Library, accessed April 28, 2019, https://www.lib.montana.edu/privacy-forum/.

4. Mimi Calter, "Guest Post—Protecting Patron Privacy in Digital Resources," *The Scholarly Kitchen*, March 13, 2019, https://scholarlykitchen.sspnet.org/2019/03/13/guest-post-protecting-patron-privacy-in-digital-resources/.

To cite this article: Clifford A. Lynch. "Reader Privacy: The New Shape of the Threat." *Research Library Issues*, no. 297 (2019): 7–14. https://doi.org/10.29242/rli.297.2.