SSO's Utopian Promise Is Based on Flawed Assumptions

Heather N. Shipman
Cornell University Library, heather.shipman@cornell.edu

Heather N. Shipman, "SSO's Utopian Promise Is Based on Flawed Assumptions" (2018). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/https://doi.org/10.5703/1288284317009

SSO's Utopian Promise Is Based on Flawed Assumptions

Heather N. Shipman, Cornell University Library, Heather.shipman@cornell.edu

Abstract

Single-sign-on authentication (SSO) for licensed library e-resources is growing in popularity, touted as a valuable tool supporting the personalization of user experiences while maintaining user privacy. Such proposals, however, are based on assumptions that are not well supported by evidence. This paper addresses three such flawed assumptions: that SSO assures privacy; that all authorized patrons have SSO credentials; and that personalization is desirable to libraries and their patrons. In reality, privacy is merely one possible SSO configuration, not a guarantee; walk-in library patrons do not have SSO credentials; and there is a growing body of evidence that existing personalization algorithms, and the data collection practices that feed them, cause great harm to users and to society. In this paper, I present Cornell University Library's experiences and concerns surrounding these particular issues, which lead us to oppose SSO authentication for our licensed e-resources unless certain conditions are met.

Context

This paper reflects a portion of the Charleston 2018 presentation "Authentication, Identity Management, Privacy and Personalization: How Can Libraries Strike the Right Balance and Avoid the Growing Dystopian Dangers?" The panel, which addressed various issues associated with this broader topic, consisted of Kari Paulson, Molly Rainard, Steven Harris, Heather Shipman, and Josh Howlett. In my portion, I addressed the "dystopian dangers" posed by single-sign-on (SSO) authentication, in which a library patron signs into an external online service by logging in to their library's institutional account; the institution then sends data about the patron to the service provider to verify that the patron is entitled to use the service.

We used the word "dystopia" in our title, description, and during our presentation; Kari observed that its purpose is to draw attention to the severity of the problem. However, the word itself comes from philosophy and fiction; it implies a situation so extreme that it's difficult to believe that it could become reality. Because the dangers inherent in exposure and collection of user data *are* a reality, I've chosen to turn the burdens of disbelief and proof around: SSO has been proposed as a utopian solution to a variety of problems, but the evidence shows that it has, thus far, failed to prove that it works as promised.

Contemporary culture frequently describes a dystopia as a utopia gone wrong: assumptions on which the utopia were built contained fatal flaws. In this panel and paper, I outline the three most egregious flawed assumptions inherent in SSO proposals.

Assumption #1: SSO Is *Anonymous*[1]

It's true that SSO *can be* configured to preserve user anonymity. It is *not*, however, the only possible configuration, nor is it the default. At many institutions, the SSO configuration is not under the direct control of the library, limiting our ability to protect our patrons' data.

Such is the case at Cornell University Library (CUL): Cornell University's central IT department (CIT) uses Shibboleth as its SSO implementation for the entire campus and is responsible for its configuration. Because the library world's discussion of SSO assumed and promoted the benefits of its anonymity, CUL didn't closely examine CIT's configuration.

We discovered our mistake mid-2016, when I downloaded a detailed non-COUNTER usage report from the ProQuest Ebook Central (PQEC) platform and discovered that title- and session-level information was associated with usernames, and the usernames were our personally identifiable university netIDs. There are two major problems that led to this scenario.

> First: PQEC shouldn't have been collecting usage data at this level.

Our discussion of this issue with ProQuest branched out into further discussion of libraries' concerns and practices regarding patron privacy, and ProQuest has taken our strongly worded opinions under advisement. ProQuest's data collection practices have since changed with the introduction of the European Union's General Data Protection Regulation (GDPR).

https://doi.org/10.5703/1288284317009

We continue to work with ProQuest on these issues; this panel—organized by ProQuest's Kari Paulson—is a part of that work. The dangers of widespread patron data dissemination and collection cannot be addressed by merely creating a mutual understanding between CUL and ProQuest. No one can solve this alone.

> Second: Cornell's SSO configuration exposed personally identifiable information.

Cornell is a member of the InCommon federation, a centralized service facilitating the transfer of SSO data between participating identity providers and service providers. Cornell's membership in InCommon predates the library's use of SSO authentication; the identity data that Cornell sends to InCommon is the data any member service provider would receive by default. CIT has configured it such that personally identifiable information (PII) (such as name, netID, and e-mail address) are provided to InCommon.

This is great for research collaboratives who use InCommon to facilitate communication between collaborators across the world, but it's disastrous for the preservation of library patron privacy. CUL has a good working relationship with CIT—a luxury not enjoyed between every library and their IT department, unfortunately—and so, moving forward, we're working with them to address this issue:

- We judge IP authentication to be far less problematic than SSO, and are avoiding SSO implementations.
- We have cancelled a subscription whose platform replaced IP authentication with an SSO authentication more invasive than we were willing to accept.
- When a newly added, much needed e-resource requires SSO, we're working with CIT to write an overriding SSO configuration for that e-resource to release as little patron data as possible.
- We're launching a broader CUL investigation into privacy issues at large, with intent to craft a more comprehensive privacy policy, as well as tools and workflows to support it.

Ideally, we want the default SSO authentication configuration to submit the smallest amount of data possible, but because the current configuration has been in use for a very long time and is used by a large number of service providers, retroactive cleanup will be extensive and must be conducted carefully to avoid breaking many authentication processes simultaneously. Additionally, the responsibility for this project would lie entirely within CIT, not CUL, and thus initiating it will require building significant political will at the highest levels of both CUL and CIT. It is our hope to foster this as part of our work on the broader privacy initiative.

Assumption #2: All Authorized Patrons Have a Login

For SSO to function, a library patron must enter their institutional username and password. However, not all authorized patrons have such a login—nor should they need one. Walk-in patrons should not be expected to create an account in order to use library e-resources onsite; this is an especially critical issue for public libraries that serve entire communities. In many cases, library funding may include a requirement to provide such access.

Cornell University is a hybrid public/private university; although some of our colleges are endowed, others are state-funded and associated with the State University of New York (SUNY) system. SUNY policy states that "the public is given access to University libraries insofar as possible."[2] CUL requests walk-in access for licensed e-resources in all license negotiations, largely successfully.

Theoretically, SSO services like Shibboleth can support walk-in access, but this is far more complicated to implement than on-campus IP authentication is. Furthermore, CUL is not aware of any example of successful SSO implementation with viable walk-in access. We have, however, seen discussion of problematic implementations via the ERIL-L mailing list.

Assumption #3: Libraries and Their Patrons Want Personalization

"Personalization" comes at the price of data collection, and libraries are widely believed to be one of the last institutions still protecting user privacy;[3] active defense of patron privacy is part of the library mission, even including defense against our own governments.[4]

Regarding data collection practices, Cornell has more questions than answers, and at the forefront of them all, we ask: *what happens to the data?* There is an extreme lack of transparency on this front.

Sometimes this is because an organization doesn't have the infrastructure in place to communicate this information effectively. CUL's initial efforts to investigate our SSO data leaks were hampered merely by no one—CUL or external—knowing who the right person to ask was, and staff turnover compounds that problem. Continuing investigations are also often hampered by the lack of a common vocabulary between programmers, privacy advocates, product managers, and other stakeholders. We have made inroads on this problem, but it is far from solved, and remains a major issue across the library e-resource industry.

But sometimes the lack of transparency is deliberate. It should go without saying that this is unacceptable.

Furthermore, the precedents for personalization proposals are appalling: companies like Google, Facebook, and Amazon are making money hand over fist by using Big Data and AI to "personalize" their platforms, but there is a growing body of evidence that these algorithms are doing very bad things.

In her book *Algorithms of Oppression*, Safiya Noble argues that search engine algorithms reinforce and amplify existing racist and sexist prejudices, while simultaneously presenting an air of neutrality because a *computer* chose the results.

In the keynote sessions of the 2018 ER&L conference,[5] Robyn Caplan, danah boyd, and Siva Vaidhyanathan discuss ways in which Big Data, social media, and other algorithms are impacting our entire culture—not, generally, in good ways. Vaidhyanathan, in particular, focuses on ways in which they're actively undermining democracy.

Facebook has run unethical research studies on its users;[6] Amazon built (and later scrapped) a recruiting tool that showed bias against women;[7] Target was able to identify a pregnant teen and, through their targeted advertising, disclosed this to her family;[8] data breaches are rampant.[9]

Humans have been working on personalization algorithms for years, and these are the results. This is our proof of concept; this is the precedent some service providers are eager to follow.

Cornell University Library seeks to avoid sacrificing patron privacy to personalization proposals.[10] We desire proof that the data is being handled in responsible and ethical ways, and so far, we have seen no such proof.

What Would Make SSO Palatable?

Under what circumstances would CUL reconsider our opinion of SSO authentication? First, all of the aforementioned problems would need to be solved. We desire the following:

- Patron anonymity should be the default, and it should be *difficult* to expose PII by accident.
- Walk-in patrons should be able to use library e-resources without being asked for a login.
- Service providers should be transparent with regard to which fields of data are being collected, what they're being used for, and how long they're stored.
- Contact information for those responsible for these practices should be easy to find and use.
- Service providers should be willing to work with libraries toward mutually acceptable data collection and retention practices.

Once the above issues are solved, we would be more willing to consider limited personalization protocols, under the following conditions:

- Patrons should still be able to use the service without revealing PII.
- Personalization should only be enabled by the patron's request.
- Patrons should be able to choose the level of personalization they desire and be properly informed as to what data would be collected to enable each personalization function.
- Data use should be ethical.

Under no circumstances should the library e-resources industry emulate Google, Facebook, and other companies collecting and mining user data to benefit their revenue streams, to the detriment of the user's privacy. In fact, given the importance of privacy to the library brand, we believe that library e-resource vendors actively protecting patron privacy could have a significant advantage in the marketplace over competitors undermining or ignoring privacy issues.

Notes

1. RA21: Resource access for the 21st century. *FAQ.* Retrieved from https://ra21.org/index.php/what-is-ra21/faq/#Does_RA21_take_into_account_user_privacy

2. State University of New York. *Public access to SUNY Libraries.* Retrieved from https://www.suny.edu/sunypp/documents.cfm?doc_id=330

3. Lehane, C. (2017, December 4). *Libraries and the fight for privacy.* Retrieved from https://www.huffingtonpost.com/entry/libraries-and-the-fight-for-privacy_us_5a258588e4b05072e8b56b44

4. Roberts, D. (2015, June 5). *NSA surveillance: How librarians have been on the front line to protect privacy.* Retrieved from https://www.theguardian.com/world/2015/jun/05/nsa-surveillance-librarians-privacy

5 . ER&L. *2018 Conference.* Retrieved from http://www.electroniclibrarian.org/past-conferences/past-conferences-2018/

6 . Arthur, C. (2014, June 30). *Facebook emotion study breached ethical guidelines, researchers say.* Retrieved from http://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say

7. Dastin, J. (2018, October 9). *Amazon scraps secret AI recruiting tool that showed bias against women.* Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

8. Hellman, E. (2015, June 16). *"Toward the post-privacy library?* Retrieved from https://americanlibrariesmagazine.org/2015/06/16/toward-the-post-privacy-library/

9. McCandless, D., Evans, T., Barton, P., & Tomasevic, S. (2018, December 5). *World's biggest data breaches & hacks.* Retrieved from http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

10. Cornell University Library. *Library practices on the collection, use, disclosure, maintenance and protection of personally-identifiable information.* Retrieved from https://www.library.cornell.edu/practices