Access and Assessment: Delivering Privacy Preserving Services

Heather Staines, Tasha Mellins-Cohen, Tim Lloyd, Anne Osterman and Karen Brunsting

Abstract

Easy access to resources is a primary concern for researchers and libraries. Maintaining user privacy is also important. This session presents three services that navigate issues around access and privacy. Counting Online Usage of NeTworked Electronic Resources (COUNTER) has determined methods to provide data while upholding user privacy. SeamlessAccess streamlines the discovery process for users. GetFTR helps users determine which articles in their search results are accessible. Although these services are very different, they all aim to protect user privacy while providing the best user experience possible. Please note this article is a record of the recorded presentation.

Keywords: SeamlessAccess, GetFTR, COUNTER, privacy, user data, federated authentication, Open Access, online access

Introduction

Heather Staines, the moderator of this presentation, introduced the session with the hope that the topics covered would provide extremely valuable information on services that libraries and publishers are using today. She also introduced the speakers, who have a wealth of

experience in services and organizations that attempt to reconcile the issues between access and privacy.

COUNTER and Privacy

Tasha Mellins-Cohen explained the ways in which Counting Online Usage of NeTworked Electronic Resources (COUNTER) ensures the privacy of user data. The COUNTER Code of Practice 10.2 addresses data privacy and user confidentiality. This section of the code of practice states that content providers are barred from releasing or selling individual and institutional usage information without permission. It specifies that institutional data cannot be released without the permission of the institution or consortium to which it belongs. There are a few exceptions, including allowing consortia managers access to usage data of institutions on behalf of a member institution. COUNTER has always required content providers to comply with relevant laws regarding user privacy and data collection when they process data for COUNTER reports. In many cases, this means complying with the General Data Protection Regulation (GDPR) passed by the European Union (EU) in 2018. The GDPR is considered the strictest privacy and security law in the world. Content providers who comply with this law are typically compliant with other privacy and security laws by default. All content providers based in the United Kingdom (UK) and the EU must comply with the GDPR, but many content providers based elsewhere also comply with the GDPR to work with UK and EU institutions.

The total usage data on a content provider's platform is compiled from two types of users. These are attributed users, who belong to an identified institution, and non-attributed users, whose data is allocated under world data. To illustrate how usage falls into these categories, imagine the following scenario: user Sam visits a publishing platform. The platform first checks Sam's internet protocol (IP) address to determine if it matches an institution's IP or a previous login using identity

management software such as Shibboleth. If there is a match, all the data usage created by Sam will be attributed to Sam's institution.

Non-attributed data is, for example, when users like Sam log in with an IP that does not match an identified institution. Users may still have access to some content on the publisher's site, but whatever data usage the platform captures will show up in COUNTER reports as world data use. One of the reasons COUNTER created world reports was to continue providing user data while facilitating the changing methods by which users access content. Content providers have historically relied on IP addresses to authenticate users and determine their access. Content providers have used IP authentication as the main mechanism to generate COUNTER user reports. These reports are limited to users working within an institution's IP range. However, because users increasingly access content from their personal computers, they are also using personal IP addresses, which are considered personal information under the GDPR and are governed by that legislation.

Users also increasingly access content from Open Access (OA) publishers. Most OA platforms do not have attributed users. Fully OA publishers typically do not have records of institutional IP addresses because OA publishers do not require users to authenticate to get access. It is cheaper not to maintain an authentication mechanism. OA publishers can supply total usage data but typically not institutional-level usage data. To provide more granular usage data for libraries that are interested in OA usage data, COUNTER created world reports. COUNTER captures non-attributed use data and provides it in world reports, which can be broken down by geographic area.

COUNTER protects user privacy in multiple ways. If a user's IP is recognized through IP authentication, there is no record of their personal data. However, when a personal login or Shibboleth is used, the content provider may have access to personal information. COUNTER provides a second level of protection. It only reports data from an institution or the world; no personal information is gathered. If the content provider collects a more granular level of data, that information is

stripped out as part of the process of creating COUNTER reports. Final layers of COUNTER security are reporting the number of searches, not the number of users from a particular institution, and not reporting behavioral patterns such as the routes a user takes to navigate through a platform. Both of those types of reports risk de-anonymizing the users from an institution.

SeamlessAccess and GetFTR

Tim Lloyd provided overviews of two services that improve users' online access experience. He explained what these services are as well as how they preserve user privacy. SeamlessAccess is a community-driven initiative and service that enables a more streamlined online access experience. Information providers implement SeamlessAccess to reduce the challenges that users encounter when trying to access paywalled content. One friction point is the varied terminology providers use to guide users to articles and other content. Users are expected to recognize terms such as login, access full text, access PDF, Shibboleth, federated authentication, as well as others. Content providers also employ many different access workflows and visual layouts to direct users through their sites. Users must keep navigating and learning new pathways to access content. This increases the cognitive load for users, which may result in frustration and an attempt to find content elsewhere. When users were primarily on campus, they had a more seamless access experience because the content provider automatically authenticated the IP address being used. However, as users increasingly try to access content outside campus networks, they are more likely to go directly to a provider site instead of using a library portal, bypassing proxied resource links that would otherwise authenticate them. One solution is federated authentication, such as Shibboleth. This enables users to get access at the point of discovery—while off-campus and with a single username and password—rather than requiring them to access resources via the library website.

Created in July 2019 as a community-driven initiative, SeamlessAccess builds on the federated authentication system. The goal of SeamlessAccess is to deliver an access experience that is as seamless as possible; specifically, it is designed to be intuitive, consistent across varied platforms, and work perpetually regardless of the user's starting point, affiliation, and device. SeamlessAccess builds on the existing federated authentication structure that allows users to log in using their existing institutional credentials. Federated authentication is preferred over IP authentication because it offers delivery at the point of discovery. There is no need to route users through a proxy URL. SeamlessAccess supports personalization without requiring the user to register personal credentials with each vendor. It protects user privacy by enabling pseudonymous access, which addresses one of the libraries' major concerns regarding individual logins. Federated authentication is more secure and requires less ongoing administration than IP address authentication.

SeamlessAccess works at two levels. First, it is a service that simplifies the user's access experience. It provides a consistent call to action on all the participating platform's sites. The service also remembers a user's institutional affiliation across participating publisher websites, removing the need to re-select an institution each time the user accesses a resource. SeamlessAccess implements best practices around access experiences, such as user interfaces and workflows. Second, SeamlessAccess is developing standards and best practices for federated authentication, as well as working on access issues such as the impact of changes to browser security settings and privacy.

SeamlessAccess protects privacy in multiple ways. It only records the user's institutional affiliation, which is stored on the user's device. SeamlessAccess only shares that information with participating service providers for one reason: to simplify the user's institutional login. The user can opt not to have it stored and can delete it at any time. No other personal information is stored.

SeamlessAccess has also led two cross-industry projects related to privacy. One project has developed standards to protect the privacy of

library users who access resources through federated authentication. Until recently, there have not been any standards to address this issue. To this end, the SeamlessAccess working group has developed two new entity categories, anonymous authorization and pseudonymous authorization, which were formally adopted by the identity federation community in March 2021. Anonymous authorization can be used by library service providers to grant access based on authentication and allows the site to make authorization decisions based on affiliation. It does not require any other user data, only institutional affiliation. There is no personally identifiable information in anonymous authorization. Pseudonymous authorization is the same, except for an added opaque, unguessable subscriber identifier that enables content providers to personalize the user experience. Again, there is no personally identifiable information, and usage cannot be tracked across content providers. Lloyd likened this to wearing a different Halloween mask to each house you visit.

The second project led by SeamlessAccess is a toolkit for contracts between libraries and service providers. It will also serve as a reference for service providers on library requirements. The toolkit provides model language that can be used to update contracts and documentation to help libraries and service providers choose the appropriate entity category for desired results and outcomes. This working group released a model license agreement for public comment in March 2022 and are using the feedback received to develop a revised edition.

GetFTR

GetFTR is a service that addresses friction for users, which includes not knowing which of the relevant search results are accessible at the point of discovery. Some results may be OA while some may be paywalled, and users must click on each result one by one to determine a resource's access level. Another friction point is not knowing whether the version linked is the version of record or not. These barriers push

users away from scholarly discovery, and therefore away from scholarly content.

GetFTR accelerates the discovery of and access to authoritative, trusted research. It is a free-to-use solution for discovery services and scholarly collaboration networks. It is sustained through financial contributions by participating publishers. GetFTR prevents users from having to determine accessibility by clicking on results one by one. It maximizes the number of accessible articles by combining different sources and entitlement systems. It also provides users access to the best version of the article available through their institutional subscriptions. GetFTR delivers an improved, cohesive user experience across multiple sites and platforms. It removes access friction in the authentication process by supporting multiple access methods and removing unnecessary steps. GetFTR currently supports more than forty-five million Digital Object Identifiers (DOIs), which represent more than fifty-four percent of scholarly content, and is enabled through several participants including sixteen integrators, discovery services, and scholarly collaboration networks as well as twelve publishers and two publishing platforms.

GetFTR works by enabling real-time access to participating publisher and aggregator entitlements. First, the user generates a set of articles using their standard discovery workflow. Then GetFTR's application programming interface (API) sends the user's institutional affiliation to participating publishers to check access. Next, the API responses confirm the user's article-level entitlements and provide links to the content. Finally, GetFTR incorporates those links into the search results to identify which articles the user can access. A "Get Full-Text Research" icon is used to identify which articles the user has access to. Users have this consistent visual identifier across research solutions that use the GetFTR service.

GetFTR has several methods for protecting user privacy. Data such as sensitive personal information and search terms are not passed to GetFTR. It only receives the DOIs and the information necessary to identify the user's institutional affiliation. GetFTR does not track

individual users, analyze data at the user level, or make user-level data available to any third party. Participating publishers must commit to not tracking users with the information received from GetFTR and not combining this data with any other data they may have. GetFTR has no visibility of the links a user clicks on in the discovery service. The links go directly to the content provider's site and redirect to the user's institution for authentication when relevant. GetFTR is not involved in article retrieval in any way.

Libraries and Consortia

Anne Osterman contextualized the information in this presentation by providing a library and consortia perspective. Osterman reminded us that the right to privacy and information freedom is a foundational library value and must be remembered as the framework from which we approach these issues. Libraries are not immune to pressures surrounding privacy, and we all must demonstrate relevance in value and return on investment. The projects presented in this session can help us demonstrate value and relevance.

Osterman highlights OA as an interesting and growing area for everyone in the scholarly ecosystem. OA articles and journals do not require affiliation with an institution for access and therefore require different methods of determining usage data. There is value for libraries in collecting the local usage of OA journals if it can be determined. There is also value for libraries in understanding how the OA projects that libraries and consortia fund are being used globally. Traditional user data is sought and valued for many reasons and by various groups. Libraries must find ways to provide or collect this data while protecting our users' privacy. Libraries are increasingly interested in collecting and analyzing user data but understand the need to maintain privacy. User data is helpful when trying to understand how different groups use different resources and is essential in relating the stories of libraries and consortia.

Conclusion

User access barriers are a significant challenge to libraries' continued relevance. If users go outside of the library's resources, the library loses relevance. The services such as SeamlessAccess and GetFTR are ways to help streamline access to library resources while reducing barriers. They provide a way to operate within library values of protecting privacy while giving users what they want. As we navigate issues of personalization and protecting privacy, the fact that publishers and libraries are working on these projects in close collaboration is valuable.

Contributor Notes

Heather Staines is Senior Consultant at Delta Think.

Tasha Mellins-Cohen is Executive Director of COUNTER.

Tim Lloyd is CEO of LibLynx.

Anne Osterman is Director of VIVA.

Karen Brunsting is Acquisitions & Collection Development Librarian at the University of Memphis.