Report on the CNI Authentication and Authorization Survey 2016

Clifford Lynch

Published August 2016

In June and July 2016 the Coalition for Networked Information (CNI) conducted a brief e-mail survey of its college and university members on authentication and authorization practices related to sharing user information with library-licensed external content providers (publishers, platform providers and aggregators). The survey (a copy is here: cni.org/authentication-survey-2016) was sent to member representatives at some 188 institutions representing both the library and information technology organizations. We asked about both technical and contractual approaches to the control and management of this data. We wanted to ascertain the extent to which information about individual users of licensed content was being passed to content providers, and if so, what measures were being taken to control this data, to determine the extent to which privacy concerns in this area were or were not being addressed by higher education institutions.

These results should be read with a number of strong caveats. We had responses from about 60 institutions, and we claim no statistical rigor in this work. This is a complex and nuanced area, and often the answer is "it varies from content supplier to content supplier." Not all responses were entirely clear or comprehensive, and I had to use my best judgment in how to interpret them. These results are best viewed as giving a sense of what's actually being done at present, and perhaps as offering some insight into trends and underlying thinking. Indeed I have deliberately rounded the numbers to discourage overly quantitative pronouncements based on the data. Note that our agreement with the responding institutions precludes sharing of the individual responses.

Basically, content suppliers to research libraries authorize users either by origin IP address, or by obtaining and examining user attributes from a trusted source (using Shibboleth as a mechanism and the InCommon Federation managed by Internet2 as a business framework in the United States; there's a lot of complex activity with inter-federation trust frameworks taking place on a global basis, but the majority of CNI's members are in the United States, and issues about resource providers *across* trust federations are still very much a moving target, so we didn't explore those questions). There were about six responses from Canadian higher education institutions, only a couple of which reported using attribute-based authorization.

Slightly over half of all respondents had implemented Shibboleth, with larger universities outnumbering smaller ones by about two to one. However, very few

reported that they were using Shibboleth for content resources; most of the applications were in other areas. Even those using it for content resources said that it was only used very selectively, with JSTOR, Project MUSE, and HathiTrust most commonly cited as the examples. It is worth noting that the list of InCommon sponsored partners (see incommon.org/participants/) does include a number of major commercial and nonprofit publishers such as Elsevier and the Association for Computing Machinery, and while we did not specifically ask respondents about these, none of our respondents explicitly mentioned them as examples. (I do know, anecdotally and outside of the survey, that attribute-based authentication is in fact being used with Elsevier. It may be worthwhile, as a follow-on to this survey, to ask some of the major scholarly publishers for their perspective on the state of play of attribute-based authorization.)

About a dozen responses indicated that they were passing personally identifiable data (names, email addresses, etc.) in attributes. Note that there is a much shorter list of research and scholarship (R&S) service providers that are part of the InCommon infrastructure (see incommon.org/federation/info/all-entity-categories.html#SPs) and at present there are no publishers on this list; most respondents say that they will pass personally identifiable data to these services. These R&S service providers operate under common rules rather than making bi-lateral agreements with individual institutions. Up until now, the primary driver for the R&S work has been the needs of multi-institutional scientific collaborations.

All but about five respondents use EZproxy or some variant (the remaining few are using VPN, or virtual private network, based solutions). This seems to be the main (or only) way to handle access to off-campus content resources at most institutions, and with IP-based authentication no personally identifiable data is passed to the content suppliers. A number of respondents, elegantly, use Shibboleth to manage access to the EZproxy system. Note that there are many content suppliers who seem to have no plans to support Shibboleth, so EZproxy or something similar is clearly going to be required on an ongoing basis; recognizing this reality, some institutions simply went with EZproxy as a standard mechanism for *all* external resources. Several respondents also noted that it was easy, with a proxy solution, to ensure that no personal data was passed to content suppliers, and that this was entirely within the library's control, avoiding complex discussions and potential lack of clarity about attribute release policies.

We asked if contracts with content suppliers contained language limiting collection, retention and reuse or resale of data about users and their activities. About 15 institutions made at least some effort to include language limiting retention or resale, though often this is inconsistent from one contract to the next even at these institutions. Several respondents noted that they felt this wasn't much of an issue because they weren't passing any personal data to the content suppliers in the first place. Re-identification of users by the content supplier (by soliciting email addresses, for example, so that users could get notifications of new content), and the subsequent reuse of that re-identified data, does not seem to be much of a consideration by responding institutions, contractually or otherwise, except that a few do make some effort to educate users about the privacy implications of choosing to disclose data to external content providers.

Finally, a number of respondents mentioned contractual provisions for content providers to provide usage data back to institutions, most commonly following the

NISO COUNTER (Counting Online User NeTworked Electronic Resources) work. Given the apparently very limited use of attribute passing to content providers, however, it seems that little is being done in terms of either content vendors returning usage data faceted by user attributes passed to them, or very detailed usage logs that include anonymized unique identifiers passed from the institutions and returned to the institution, where they can be de-anonymized at various levels of detail and aggregation.

While the point of this work was to gather data and insights, I will risk a few tentative conclusions based on the data, comments in the survey responses, and a few subsequent conversations with respondents.

Right now, IP based authentication and proxies are the dominant approach. This is tried and true, and, at least from a privacy point of view, relatively safe and simple.

What are the arguments for a shift to attribute-based authorization in this specific context? I think that this approach offers the best options for rich data analytics at institutions licensing content access, with the best privacy protections being, perhaps, to only pass a random opaque identifier which the content provider reports back in transactional usage data. These IDs are then de-anonymized locally when doing analysis based on the externally reported usage data. This process is technically complicated, and it also places a lot of responsibility on the institution to behave responsibly in protecting and retaining data, and in what it does analytically with that data. This is probably a most realistic scenario for large, sophisticated institutions that negotiate high-absolute-value complex contracts ("big deal" bundle licensing agreements) with content providers. To the extent that this future actually unfolds, it also creates interesting prospects for data sharing and pooling across institutions to gain a better understanding of the use of various content resources, and of the emergence of third-party services that might lower barriers to participation.

Finally, for those institutions doing attribute-based authorization, the survey sometimes functioned as a wake-up call, and certainly underscored the importance of all stakeholders, including libraries, to fully understand and, where appropriate, participate in the ongoing development and deployment of attribute release policies. This can be complex, and it's also vital to recognize that the faculty, staff and students at our universities have a key stake in this issue, so effective communication and education are also essential. The responses made it clear that attribute-based authorization is deploying much more rapidly in areas other than licensed library content resources.

My thanks to the member representatives who took the time to respond to this survey; in many cases this involved considerable coordination within their institutions. I am particularly grateful for those who not only shared what they were doing, but some of the reasoning behind the choices that they had made. Thanks also to the CNI steering committee members who helped refine the questions we asked, and to Joan Lippincott and Diane Goldenberg-Hart of CNI for their help with the process.