Duke Libraries Data Privacy and Retention Audit Report

Joyce Chapman & Angela Zoss
Assessment & User Experience Department[1]
Duke University Libraries[2]
January 10, 2020

Contents

Duke Libraries Data Privacy and Retention Audit Report 1
Key concerns 2
Background 2
Summary of systems 3
Logging systems 3
User account systems 5
Manual data entry systems 6
The Wild West 7
Next steps 7
Appendix A: Interview questions and personal data fields 9
Personal data fields 9
Interview questions 9
Appendix B: Systems for which interviews were conducted 10
Appendix C: Systems reviewed by the Audit team for personal data 11

[1] Assessment & User Experience Department website, https://library.duke.edu/about/depts/assessment-user-experience
[2] Duke University Libraries website, https://library.duke.edu/

Key concerns

Over the course of a review of Duke University Libraries (DUL) data systems that contain personal data about patrons, we identified the following key concerns:

1. DUL lacks policies and consistency around data privacy and retention across all systems.
2. What policies do exist are inadequately documented and inadequately communicated to patrons.
3. We have neither a centralized body to make decisions around data privacy and retention policy, nor a formal requirement that departments engage in decision-making around these issues.
4. For locally hosted systems, our current, largely passive approach to detecting data breaches may be missing security concerns.
5. Even secure systems can facilitate data privacy breaches if staff are not well educated about how to protect patron data.
6. We do not currently have robust staff training around data privacy and retention.
7. Increasing reliance on Single Sign On (SSO) – a requirement for Enterprise software at Duke – has begun to raise security concerns at other libraries because of the potential that unnecessary personal data are being shared with vendors during authentication.

Background

In 2016, the European Union (EU) passed the General Data Protection Regulation (GDPR), which went into effect May 25 2018, becoming the primary law regulating protection and handling of EU and European Economic Area (EEA) citizens' personal data – including the transfer of their personal data outside EU and EEA geographical areas. Institutions that are not compliant with the law can face stiff penalties and fines, which brought GDPR-compliance to the forefront of the conversation for many universities in the United States in 2018. In the summer of 2018, California passed the California Consumer Privacy Act (CCPA), which also put a spotlight on data privacy issues for many universities.

Additionally, Assessment & User Experience (AUX) department member Angela Zoss represented Duke Libraries at the 2018 National Forum on Web Privacy and Web Analytics.[3] This IMLS-funded grant brought together librarians, technologists, and privacy researchers to produce a "practical road map for enhancing our analytics practice in support of privacy." DUL's participation in the forum, as well as the recent developments with GDPR and the CCPA, prompted DUL Libraries staff to consider our own policies and procedures based on recommendations in the Forum's resulting white paper and action handbook.[4]

AUX decided to undertake a more thorough study of data privacy and retention at DUL. AUX charged a small team, Joyce Chapman and Angela Zoss, to conduct the DUL Data Privacy and Retention Audit. The

[3] Web Privacy Forum website, http://www.lib.montana.edu/privacy-forum/

[4] Web Privacy Forum White Paper, https://scholarworks.montana.edu/xmlui/handle/1/15445 and Web Privacy Forum Action Handbook, https://scholarworks.montana.edu/xmlui/handle/1/15446

team developed a list of data fields that constituted "personal data" for the purposes of the Audit, as well as a list of ten questions about each data system.[5] They identified 70 library systems for review[6] and conducted interviews with technical and functional stakeholders for two dozen systems[7]. Systems that capture personal data fields for *library patrons or donors* were included; personal data related to *staff* were not considered criteria for inclusion.

This report is a synthesis of the Audit's findings. We recommend that the Libraries' Executive Group's next step be to charge a cross-departmental team to review Audit findings and make recommendations for improvements to DUL's current policies and procedures around data privacy and data retention.

Summary of systems

The two dozen systems included in this report are grouped into four areas based on similar characteristics to assist in summary and description. The four areas are "logging systems," "user account systems," "manual data entry systems," and "the Wild West."

Logging systems

Several of our systems use automatic logging to track user behavior. In most cases, these logs do not collect highly sensitive, identifiable personal data. The risk of re-identification, however, is large because of the sheer quantities of data that can be amassed by logging and the fact that it captures user behavior – typically details about interactions with items in our collections.

Logging Activity on our Web Applications

Apache Server Logs

By and large, the web applications we host on Duke servers use the default Apache server logging software to track usage. Server logs help IT staff with activities such as identifying malicious attacks on our sites, or troubleshooting problems when patrons encounter errors in our applications. The systems that fall into this category include our repositories, our catalog software, Drupal, WordPress, and other custom and licensed software we are running on local servers. With this logging system, we typically do not retain personally identifiable data. The main fields captured are IP address[8], timestamp, and the resource accessed. One notable exception is that DukeSpace (an institutional repository) has a logging system that includes user account information, if the user is logged in.

---

[5] See Appendix A

[6] See Appendix C

[7] See Appendix B

[8] Note that IP address can be somewhat sensitive, as a device may retain an IP address over one or more sessions, and some devices accessing our resources might have a fixed IP addresses. For devices using wireless access points on campus, especially, IP address can be reassigned over time, even within a single session if connections are dropped. There is always a chance DUL would be subpoenaed to provide information about an IP address during a certain period of time.

Apache server logs can be retained indefinitely, and for some DUL systems this is the case. For example, the logs from the older version of DukeSpace are retained for analysis purposes. For most of our newer systems, the logs are set to be purged at regular intervals to save disk space.

The IT Security Office at Duke has developed a Duke University Standard for Logging[9] that might be instructive for establishing DUL policies around logging. Another model for responsible logging for some systems might be the technique employed for EZproxy logging. Relevant information is extracted from the logs daily and aggregated, and then the logs are immediately purged.

Google Analytics and Google Custom Search

Another major type of logging for usage of our web applications is Google Analytics (GA). GA helps us assess how patrons use our sites and services so we can make informed decisions about how to improve them. With GA, DUL registers a web site (or "property") with Google, then adds code to our web site that allows Google to track some of the usage of the site. The default code sends the IP address of the user to Google, but additional code can be added by DUL to anonymize the IP address before the information is sent to Google. Some of DUL's web sites anonymize the user's IP address, but not all do.

Using GA also opens our web site users up to Google's data blending and aggregation efforts. For example, if a DUL web site user is logged into a Google account while browsing our site, or if they have other browser tabs open, Google can associate usage of our site with usage of other sites and other highly personal information.

None of the data collected by Google for this service is stored on Duke servers. Duke can access certain information about usage of our web site using the GA web portal and other analytics tools that can connect to GA data, but the data accessible to DUL excludes personally identifiable information. Even the full IP address is masked for the properties that send IP addresses to Google. Certain fields, however, have the potential to identify a user in some cases. For example, if a user accesses the website from a very unusual device, their activity might be tracked uniquely. Google, of course, collects and retains much more data than can be accessed by web site owners.

DUL also uses Google Custom Search to embed a site-specific Google search box in our web site. This provides highly relevant DUL web pages as results matching patron search queries. Search terms are very personal, and libraries are highly motivated to protect these data and prevent them from being connected to personal information. We set up our GA to prevent us from connecting search terms to specific user characteristics, but Google is certainly making these connections, both for the search terms used to find DUL web sites and for the search terms used in the Google Custom Search boxes on our sites. While moving away from GA entirely would best protect patron privacy, the services offered by GA are currently unmatched by other services.

[9] Duke University Standard for Logging, https://security.duke.edu/secure/policies/log-standard

Logging Activity on our Machines

Another type of usage logging employed by DUL is the Duke-wide LabStats software that logs usage of our public machines. It tracks netID in connection with timestamp, computer ID, and information on the software applications used during the session. The data are kept indefinitely, unless a department chooses to purge data from its assigned machines itself.

User account systems

Many of the services we offer allow or require users to create accounts – or create accounts by default on users' behalf. Accounts carry inherent data privacy concerns because personally identifiable data can be connected to historical usage data. Thus, these account-providing systems constitute the primary source of data privacy and retention concerns.

Circulation-tracking systems

One of our largest systems for user accounts is our primary Library Management System, Aleph. Aleph creates an account for everyone in Duke's identity management system, automatically, each night. Aleph tracks circulations of print materials from our non-Rubenstein[10] collections, but because of the sensitivity of that data, DUL has established retention policies and procedures for Aleph. Specifically, the patron identifier associated with a loan is scrubbed approximately one month after the loan is closed, leaving only basic demographic data. The parallel system tracking circulation of materials from Rubenstein (Aeon) retains personal data indefinitely to ensure the security of our rare materials. A third set of systems governs interlibrary loan circulations (ILLiad and D2D). These systems are hosted and do not have good support for purging personal information while retaining some information about circulation history.

Repositories

DUL hosts three separate repositories – DukeSpace, the Duke Digital Repository (DDR), and Research Data Repository (RDR). The accounts that can be created for these repositories allow users to make submissions, view restricted items, and in some cases save searches. The repositories tie into Duke's netID system (using Single Sign On via Shibboleth) to create accounts on demand. DukeSpace, our institutional repository for sharing Duke publications openly, has a tight connection to a licensed service called (Symplectic) Elements, which automatically generates accounts for all Duke affiliates and stores basic user data indefinitely. By and large, the repositories and related services do not store highly sensitive data, with the exception of DukeSpace. DukeSpace has implemented its Apache server logs in a way that connects a user's account to the web site viewed, if a user is logged in. Additionally, if a repository allows a user to save a search, the connection between the user's account and the saved search is stored in the repository's database.

---

[10] David M. Rubenstein Rare Books & Manuscripts Special Collections Library https://library.duke.edu/rubenstein/

Other accounts

Only a few other systems used by DUL include account creation.[11] The Development department interacts with the Duke Alumni and Development Database (DADD), a Duke-wide system tracking alumni and donor data across the entire university. This system is centrally managed and has extensive security protocols. An account is automatically created for every past and present Duke affiliate (as well as any family members for which Duke has data from other sources) and patrons themselves never see or interact with this system. Finally, AskTech is a local ticket tracking system for reporting and dealing with problems related to Technical Services. The system associates tickets with user email addresses, but no other identifying information is required, and the system is hosted locally with limited outside access.

Manual data entry systems

Some of our systems provide fairly unstructured platforms in which data fields are defined by staff members and the majority of data is manually entered by staff or patrons. These systems include KnowledgeTracker, the Gifts Database, the Materials Purchase Request Form, Qualtrics, Airtable, and the Springshare products LibInsight and LibCal. All of these except the Gifts Database and Materials Purchase Request form (homegrown Django applications) are licensed and hosted by vendors. All of the systems allow the personal data to be easily exported to spreadsheets, and any staff with access to the tools have permissions to export data. Most of the systems allow staff to set up many different projects, each with different staff-defined data fields. None of the systems have policies around data retention except KnowledgeTracker.

The type and extent of personal data contained within these systems therefore varies from project to project. Almost no LibInsight projects include personal information and where they do, it is typically a first and last name. The single dataset in KnowledgeTracker (used for online Special Collections reference) includes an email address provided by a user that is then associated with the reference question they submit. The Materials Purchase Request Form, LibCal, and Qualtrics have the ability to capture netID (and potentially, more personal data fields about a user) using SSO authentication. DUL has not turned this feature on in LibCal to date, but the feature is live in Qualtrics and has been used for library-administered surveys. Currently, when using Shibboleth authentication for a Qualtrics survey, the Duke survey administrators are returned the following fields: netID, first and last name, email address, organizational unit, primary affiliation, and all other Duke affiliations. The Request Form can only be accessed by Duke affiliates and therefore requires Shibboleth authentication, and pre-populates the form that the user must complete and submit with personal data, such as DUID, name, address, and phone. LibCal also contains an email address for the majority of transactions, as an email address is required in order to confirm a room booking, or register for an event.

[11] Note: this excludes systems where staff create accounts but non-staff users do not. For example, both WordPress and Drupal use accounts to track usage and permissions, but both are excluded from this analysis because the data pertain only to staff.

Qualtrics contains the most personal data of any of these systems. Survey data can contain personal information collected via SSO-authentication or any type of personal data a survey requests a user to type in manually. Surveys can also be associated with a "Contact List," which is a CSV file containing an unlimited number of fields of data about people to which a survey will be directly distributed via email. At a minimum, such a list contains an email address, but typically also contains a name, and often contains demographics as well. Contact List data can be associated with survey responses without participants' knowledge when a survey is distributed via the Contact List, causing a unique survey link to be sent to each respondent that identifies them in the dataset. Many DUL staff members have Qualtrics accounts, but Qualtrics requires each survey to be shared with either individual accounts or with a pre-determined Group and its members before it or its data can be viewed by anyone other than its creator. DUL does use Groups to some extent, but the majority of surveys exist in individual's accounts and are shared with a limited number of other staff members.

Large numbers of staff have LibInsight and LibCal accounts, including interns and student workers who are only here for short times. Every account holder can view and export data for all of DUL's projects in either system, and there is no way to limit what data is accessible to whom. While Duke Qualtrics accounts are linked to a netID and an individual will lose access once they leave Duke, accounts must be manually deleted from LibInsight and LibCal, requiring departments to remember to contact the technical stakeholder for the system and request that an account be deleted each time an intern, student worker, or staff member leaves. Airtable is only available to a handful of staff who have thus far mostly used it to track information without personal data. KnowledgeTracker access is mostly limited to Rubenstein Research Services and curatorial staff.

The Wild West

A final subset of systems containing personal data are those used every day by staff for a myriad of activities: files on computer drives, paper files, and emails. DUL currently has limited informal policies and no formal policies around the export, transmission, publishing, and printing of data from our systems. Numerous staff members run reports, perform further analysis, combine or annotate files, and save, print, or email the results. Emails with file attachments containing exported personal data are sent back and forth among library staff, or uploaded to third party spaces such as Basecamp. Some staff additionally republish data; for example, in Tableau dashboards. If default Tableau publication settings aren't changed, the raw data read into the Tableau file is available for download, and may contain personal data not used in the dashboards that the author is unaware that they are sharing.

Next steps

The team recommends that the Executive Group charge a cross-departmental task force to review Audit findings and make recommendations for improvements to DUL's current policies and procedures around data privacy and data retention. The task force would review the Audit report and gather any additional

data necessary to inform their work, which may include setting priorities, working with departments and units to create policies where they are lacking, making recommendations for how to communicate policies to patrons, and other tasks determined by the task force.

Appendix A: Interview questions and personal data fields

Personal data fields

For the purposes of this audit, we defined "personal data" to include the following information:

1. Name
2. netID or DUID
3. Email address
4. Physical address
5. Phone number
6. photograph
7. birth date
8. IP address/hostname
9. MAC address

Interview questions

1. Is personal data collected?
2. Describe scope of the personal data
3. Where is personal data held?
4. What is the retention of the personal data? What currently happens? Are there policies around retention (of any sort - oral tradition, written documentation)?
5. Who has access to the personal data while it is being retained?
6. Are there processes in place for informing users about collection/retention? Do they get a notice when registering? Do they have to sign something?
7. Do we get notifications of breaches to our data systems? (If we do, do we inform users?)
8. Can users request copies of their data?
9. Can users be forgotten?
10. Can users opt out of data collection?

Appendix B: Systems for which interviews were conducted

Not all systems for which interviews were conducted were included in the final report; some did not collect any personal data or any data about patrons.

1. Aeon
2. Airtable
3. Aleph
4. AskTech
5. BorrowDirect
6. ContentDM
7. Duke Alumni and Development Database
8. Duke Digital Repository
9. EZproxy
10. Gifts Database
11. Google Analytics
12. Google Custom Search
13. ILLiad
14. Knowledge Base
15. KnowledgeTracker
16. LabStats
17. LibCalendar
18. LibInsight
19. Materials Purchase Request Form
20. MeeScan
21. Online Journal Titles
22. Open Journal System (OJS)
23. Qualtrics
24. Research Data Repository
25. RequestApp
26. Scanners
27. Summon
28. TRLN-Direct
29. WordPress

Appendix C: Systems reviewed by the Audit team for personal data

The team reviewed and discussed the following systems to determine which were relevant to the report and would require interviews.

1. Aeon
2. Airtable
3. Aleph
4. Altmetric Explorer
5. Archive-It
6. ArchivesSpace
7. AskTech
8. Bento Results
9. BorrowDirect
10. Catalog Request System (CRS)/RequestApp
11. CONTENTdm
12. Crossref
13. Dark Archive
14. Data files embedded in other shared work (e.g., Tableau dashboards)
15. Duke Alumni and Development Database
16. Duke Digital Repository
17. DukeSpace
18. DVS Collections
19. Elements
20. EZID
21. Ezproxy
22. Files on shared drives (staff share, sharepoint, Box)
23. Files on staff machines (e.g., files received from IR, HR downloads from SAP, data downloaded from DUL systems like ILLiad)
24. Flickr
25. get it@Duke
26. Gifts Database
27. Google Analytics
28. Google Custom Search
29. Google Scholar
30. HathiTrust
31. ILLiad
32. Internet Archive
33. KnowledgeTracker
34. KnowledgeBase
35. LabArchives

36. LabStats
37. LibCal
38. LibInsight
39. Library Archival System (GFA)
40. Library Catalog UI
41. LOCKSS
42. LTI
43. Materials Purchase Request Form
44. MeeScan
45. Morphosource
46. Omeka
47. Online Journal Titles (A-Z List)
48. Open Journal System (OJS)
49. Open Science Framework (OSF)
50. Paper files from events (e.g., attendance sheets from workshops)
51. Protected Data Network (PDN)
52. Protected Data Repository (PDR)
53. Qualtrics
54. Research Data Repository (RDR)
55. Research Databases (A-Z List)
56. Research Guides (LibGuides)
57. Scanners
58. Scholars@Duke
59. Scriptorium
60. Search TRLN
61. Simile
62. sites.duke.edu
63. Suggest A Purchase Form
64. Summon
65. Tripod2
66. TRLN Direct
67. Warpwire
68. WordPress
69. WorldCat Union Catalog
70. YouTube