IJDC | *Conference Pre-print*

Privacy Impact Assessments for Digital Repositories

Abraham Mhaidli
University of Michigan

Cundiff Jordan
University of Michigan

Libby Hemphill
University of Michigan

Florian Schaub
University of Michigan

Andrea Thomer
University of Michigan

Abstract

Trustworthy data repositories ensure the security of their collections. We argue they should also ensure the security of researcher and human subject data. Here we demonstrate the use of a privacy impact assessment (PIA) to evaluate potential privacy risks to researchers using the ICPSR's Open Badges Research Credential System as a case study. We present our workflow and discuss potential privacy risks and mitigations for those risks.

Submitted 16 December 2019 ~ *Revision received* 15 May 2020 ~ Accepted 21 June 2020

Correspondence should be addressed to Andrea Thomer, 105 S. State St., Ann Arbor, MI 48109. Email: athomer@umich.edu

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v15i1.692
DOI: 10.2218/ijdc.v15i1.692

Privacy Considerations in Digital Repositories

Digital repositories and collections are often characterized by their trustworthiness (Donaldson and Conway 2015; Colati and Colati 2009; CRL, The Center for Research Libraries 2007; Corrado 2019). A repository's components that determine its trustworthiness include digital object management, technical infrastructure, security, and organizational infrastructure (CRL, The Center for Research Libraries 2007). Audits of trustworthiness (e.g. ISO 16363:2012 (Consultative Committee for Space Data Systems 2012); the earlier Trustworthy Repositories Audit & Certification checklist (CRL, The Center for Research Libraries 2007); CoreTrustSeal (L'Hours, Kleemola, and Leeuw 2019)) typically focus on evaluating a repository's stability in service of its data contents: how might a data depositor know that this repository is a place that can be trusted to host their digital objects?

We recognize that the security of a repository's holdings is paramount, but suggest that there is another aspect of trustworthiness that must also be considered: how a repository handles and manages user data. This data might range from email addresses of user accounts; to institutional affiliation or other biographical information required for access to sensitive data; to search histories, clickstreams, and other trace usage data. In some regions repositories are legally required to consider the privacy of their users' data in their design (e.g. Europe's General Data Protection Regulation requires systems processing personal information, including repositories, to adhere to its "data protection by design" and "data protection by default" requirements ("Regulation (EU) 2016/679 of the European Parliament (General Data Protection Regulation)" 2016)), yet few frameworks or checklists exist to assist repository managers in analyzing the privacy risks for researchers accessing data in digital repositories. For instance, while the TRAC checklist considers aspects of the repository's organizational structure (e.g., financial sustainability, procedural accountability), they do not address repositories' policies and procedures for managing information that their users and processes generate.

In this paper, we present a case study of a privacy impact assessment conducted as part of the development of the "Open Badges and Research Credentials System" (OBRCS) (Levenstein, Tyler, and Bleckman 2018). This digital credentialing system would lessen the administrative overhead of accessing sensitive or restricted data repositories by providing researchers with a "passport" valid for multiple repositories. Because this system would require sharing and storing user data, a privacy impact assessment was of paramount importance. We describe our workflow, privacy risks we identified, and discuss design, technical, and policy mitigation strategies and recommendations for digital repository credentialing systems. We believe this approach and findings could be beneficial to other repositories or credentialing systems.

Privacy Impact Assessment of OBRCS

A Privacy Impact Assessment (PIA) is a systematic approach for determining privacy risks of an information system and mitigations for those risks (Wright 2013). In many countries, PIAs have become a required or recommended step for the design of information systems that process personally-identifiable information (PII) (Wright and De Hert 2012). For instance, the E-Government Act of 2002 requires U.S. government agencies to conduct PIAs in the development or procurement of systems processing PII; Europe's GDPR requires a "Data Protection Impact Assessment" when a new system is likely to pose privacy risks. Many companies have adopted PIAs as part of their design processes (Wright and De Hert 2012). For digital repositories, PIAs can be used to reveal potential risks to both a system's data subjects and its users. We adapted Wright's PIA methodology (Wright 2013), which synthesizes best practices, into five phases tailored for the digital repository environment.

Phase I: Threshold Assessment and Preparation

The first phase is determining whether a PIA is necessary, and if so, who will conduct the PIA and with what timeline, scope, and budget. For the OBRCS project (Levenstein, Tyler, and Bleckman 2018), a PIA was necessary because the centralized management of researcher credentials entails the collection, storage and transfer of PII.

Phase II: Repository Description and Information Flows

The second phase is to map the system's components and information flows: what information is collected, stored, processed, and made available in different aspects of a repository. This entails describing the repository's purpose, functionality, and stakeholders; the information the repository collects from what stakeholder and why; how information is used or processed; and how this information is stored and managed.

For the OBRCS PIA, we identified information flows by first interviewing multiple key stakeholders (OBRCS project manager, senior data project manager, application manager). We also consulted with two Freedom of Information Act (FOIA) officers, as some information in the system may be subject to FOIA requests. We also reviewed the project's description (see Levenstein, Tyler, and Bleckman 2018) and security assessments of ICPSR's broader data repository management system, Archonnex. Finally, we documented what information needed to be provided when using Archonnex, what information was displayed, and any options for users to manage that information. The identified information flows were documented through system flow diagrams visualizing how information 'flows' between different entities; and through step-by-step process models showing information exchanges in different parts of the system.

Phase III: Privacy Risk Analysis

Based on identified information flows, the next step is to identify how the system could impact the privacy of its stakeholders. This may involve additional interviews with relevant stakeholders. Information flows and interview findings are then used to identify scenarios in which information might be misused. These scenarios are documented, including an assessment of how likely each scenario might be, and respective consequences.

We interviewed 8 additional stakeholders (3 potential users, 1 data repository "gatekeeper", 3 ICPSR staff managing data use agreements, 1 institution's official representative). We grouped risks based on the stakeholder impacted (researchers, data subjects, data repository's institution). We estimated the likelihood of these risks to generally be low, though the severity of any of these scenarios coming to pass could be high. These include:

- Passport holders being unfairly or mistakenly denied access to data.

This could impair passport holders' ability to conduct research.

- Physical or emotional harm to passport holders.

Passport holders could be targeted by "bad" actors, e.g., other researchers seeking to "scoop" a project or outside groups targeting researchers working on politically-charged topics.

- Reputational harm to passport holders.

If a researcher is denied access to restricted data, unfairly or not, they could be labeled as being untrustworthy.

- Undue access to human subjects data.

If a researcher is inappropriately granted access to human subjects data, this could potentially put human subjects at risk.

- Distrust of OBRCS, and by extension, institutions that use OBRCS.

Any of the above harms coming to pass could cause researchers to lose trust in OBRCS.

Phase IV: Mitigation Strategies and Recommendations

After identifying risks, the next phase is to develop mitigation strategies. These can be technical (changes to the system) or organizational (internal and external policies). The PIA should present a holistic view of recommended mitigation strategies to serve as consistent guidance for system developers and designers. The PIA's recommendations serves to evaluate and prioritize among possible mitigation steps. An important aspect in providing recommendations is to balance privacy considerations with system needs.

We identified 17 mitigation strategies, including design solutions (e.g., do not include past infractions in a researcher's profile), technical solutions (e.g., applying multi-factor authentication) and policy solutions (e.g., auditing fairness of data access decisions), which we will describe in the talk and full version of the paper.

Phase V: Implementation, Publication, and Iteration

The PIA's findings are documented in a report which guides the implementation of recommended mitigations. The report should be frequently updated to reflect the implementation of mitigations in the repository, and revisited as new features are added to the system, and as new stakeholders, uses, information flows, and data types arise.

Discussion & Conclusion

Our work makes several contributions for the digital curation community. First, the workflow we describe, adapted from , will likely be directly usable by other repositories – particularly as more institutions and countries increase emphasis on user data protection. Though our PIA focused on privacy risks to researchers using this system, this method is appropriate for identifying and mitigating privacy risks for any sensitive information, including PII about research subjects. Methods of restricting or preventing access to PII often rely on the researcher or data depositor; the PIA could instead be used by repository managers for a more centralized assessment of privacy risks. Second, our PIA identified risks that may be present in other digital repositories – particularly for those including sensitive or restricted data. Other repositories relying on credentialing services may similarly wish to consider mitigations for the risk of unfairly denying researchers access to data – or to providing undue access to data. Finally, our work expands notions of trustworthiness for repository managers and users. A trustworthy data repository should ensure the security of all information that flows through a system – not just the data it stores. Ensuring the security of repository user data will only become more important as more repositories adopt credentialing systems like OBRCS or develop new digital "enclaves" dependent on user profiles.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1839868.

References

Colati, J. B., & Colati, G. C. (2009). A Place for Safekeeping: Ensuring Responsibility, Trust, and Goodness in the Alliance Digital Repository. Library & Archival Security, 22(2), 141–155.

Consultative Committee for Space Data Systems. (2012). Reference model for an Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0m2.pdf

Corrado, E. M. (2019). Repositories, Trust, and the CoreTrustSeal. Technical Services Quarterly, 36(1), 61–72.

CRL, The Center for Research Libraries. (2007). Trustworthy Repositories Audit & Certification: Criteria and Checklist.

Donaldson, D. R., & Conway, P. (2015). User Conceptions of Trustworthiness for Digital Archival Documents. Journal of the Association for Information Science and Technology, 66(12), 2427–2444.

Levenstein, M. C., Tyler, A. R. B., & Bleckman, J. D. (2018). The Researcher Passport: Improving Data Access and Confidentiality Protection. ICPSR. https://www.icpsr.umich.edu/files/about/researcher/ICPSR_ResearcherCredentialingWhitePaper_May2018.pdf

L'Hours, H., Kleemola, M., & de Leeuw, L. (2019). CoreTrustSeal: From academic collaboration to sustainable services. IASSIST Quarterly / International Association for Social Science Information Service and Technology, 43(1), 1–17.

Regulation (EU) 2016/679 of the European Parliament (General Data Protection Regulation). (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

Wright, D. (2013). Making Privacy Impact Assessment More Effective. The Information Society, 29(5), 307–315.

Wright, D., & De Hert, P. (2012). Privacy impact assessment. Springer.