IJDC | *Research Paper*

Privacy Impact Assessments for Digital Repositories

Abraham Mhaidli
School of Information
University of Michigan

Jordan Cundiff
School of Information
University of Michigan

Libby Hemphill
School of Information
University of Michigan

Florian Schaub
School of Information
University of Michigan

Andrea Thomer
School of Information
University of Michigan

Abstract

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

Submitted 17 January 2021 ~ *Revision received* 21 September 2021 ~ *Accepted* 21 September 2021

Correspondence should be addressed to Andrea Thomer, University of Michigan School of Information, 105 S. State St, Ann Arbor, MI USA 48109. Email: athomer@umich.edu

An earlier version of this paper was presented at the 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.v16i1.753
DOI: 10.2218/ijdc.v16i1.753

Introduction: Privacy Considerations in Digital Repositories

Digital repositories and collections are often characterized by their *trustworthiness* (Donaldson & Conway, 2015; Colati & Colati, 2009; CRL, 2007; Corrado, 2019). A repository's components that determine its trustworthiness include digital object management, technical infrastructure, security, and organizational infrastructure (CRL, 2007). Audits of trustworthiness (e.g., ISO 16363:2012 (Consultative Committee for Space Data Systems, 2012); the earlier Trustworthy Repositories Audit & Certification checklist (CRL, 2007); CoreTrustSeal (2019)) typically focus on evaluating a repository's stability in service of its data contents: how might a data depositor know that this repository 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 the data of those who use the repository, and how this data handling may affect the privacy of both research subjects and of researchers accessing the repository's content. 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. For example, Europe's General Data Protection Regulation (GDPR) requires systems processing personal information, including repositories, to practice "data protection by design" and "data protection by default" (European Parliament and Council, 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 (CRL, 2007) considers aspects of the repository's organizational structure (e.g., financial sustainability, procedural accountability), it does not address repositories' policies and procedures for managing information that their users and processes may generate.

In this paper, we present a method for conducting a privacy impact assessment of a digital repository, demonstrated through application to the Inter-university Consortium for Political and Social Research (ICPSR) "Researcher Passport" project (Levenstein, Tyler & Bleckman, 2018). This project aims to create a digital credentialing system that 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 grant access to repositories with potentially sensitive information, as well as sharing and storing data about the researchers who would have a passport, a privacy impact assessment was of paramount importance. We describe our workflow, the privacy risks we identified, and discuss design, technical, and policy mitigation strategies and recommendations for digital repository credentialing systems. We believe the described approach and findings could be beneficial to other repositories or credentialing systems.

Background

Trustworthiness in repositories

Most research on trust in repositories has focused on how one might come to trust the information in a repository rather than how a repository might trust a potential user. Research toward this end has involved: developing an idea of a "Designated Community" that can better assess the trustworthiness of data held in a repository (Donaldson, 2016); ascertaining levels of trust in a specific document or dataset as opposed to the repository or fonds (Donaldson & Conway, 2015; Yakel, Faniel, Kriesberg & Yoon, 2013); developing trustworthiness as a function of the assessment of information sources (Hertzum, Andersen, Andersen & Hansen, 2002); defining trust as "lack of deception, determining data validity (or accuracy) and by assessing the integrity of repositories" (Yoon, 2014, p. 31); and identifying correlation between the level of trust with the levels of trust one has in the scientific equipment that produces the data itself (Wallis et al., 2007).

Kelton, Fleischmann & Wallace (2008) focus on trust as helping to inform information-related behavior beyond ascertaining the extent to which information in a repository can or cannot be trusted: "What is needed is a change in focus from the attributes of the information itself to the perceptions of the person who is using that information" (ibid, 371). The authors go on to discuss the relationship and distinction between trust, credibility, expertise, competence, and morality. A further shift in focus is warranted from the perceptions of the person using the information to the perception about that person. This nexus of interrelated concepts is important for digital repositories to consider, who need to better understand how they can trust the expertise, competency, and credibility of a potential user of the information they store.

Tyler (2019) addresses the question of whom a restricted data repository is to trust, and to what sorts of restricted data they might be trusted to access. An important concept carried through this work is the relationship between one's digital identity and one's trustworthiness: an authenticated digital identity must hold sufficient, persistent attributes that allow a repository to assess the trustworthiness of an individual. In this way, a digital identity can be configured and programmed to function as an interoperable, automatic, and distributed shibboleth that facilitates access to restricted data. These four facets of digital identifying trustworthiness are the starting point for the further development of the Researcher Passport project which serves as a use case for our Privacy Impact Assessment. But when it comes to trusting that rights to privacy are protected in restricted data digital repositories and archives, it is not just the privacy of research subjects and respondent data that needs to be better understood; the privacy of researchers who access restricted data digital repositories and archives needs to be better understood as well.

Privacy in Repositories

Privacy aspects in the context of repositories have largely been considered from the perspective of research subjects (the people whose data is in the repository), asking how research data when made available might affect the privacy and safety of these subjects (W. E. Miller, 1969; Bancroft, 1972; Tripodi, 1974; Hofferbert, 1976; Bond et al., 1978; Geda, 1979; Edsall, 1981; Clubb, Austin, Geda & Traugott, 1985). More recent research focuses on the risks and rewards of archiving digital social science research, especially qualitative social science research. Responding to the recent "deluge" (Borgman, 2012; Jeng, He & Chi, 2017) of digital scientific research data, much research addresses similar ethical problems of protecting research subjects' privacy and confidentiality in the context of providing the long-term storage and archiving of that data to facilitate future access (Cliggett, 2013; Bishop, 2005, 2009; Parry & Mauthner, 2004; Goldman & Pyatt, 2013; T. Miller, Birch, Mauthner & Jessop, 2012).

This research that takes seriously the privacy and confidentiality concerns of social science data research archiving and looks to advance concrete, pragmatic solutions that attempt to balance risks and rewards to privacy and confidentiality is closest to the research presented in this current study. Striking a balance between the responsibility to participant privacy and the responsibility to make social science research data available to the research community is at the forefront of the ICPSR Researcher Passport project (Levenstein et al., 2018). However, a second class of privacy risks in the context of data repositories, namely those for researchers who access repositories with sensitive or contentious datasets, have been considered less. We argue that privacy risks of digital repositories and respective systems should be analyzed systematically and holistically in order to ensure that those risks can be appropriately mitigated for all stakeholders in the design, development and deployment of digital repositories.

In the larger field of information privacy, privacy impact assessments have emerged as a procedural approach for identifying, tracking and mitigating privacy risks as part of a larger effort towards privacy by design. Privacy by design is a paradigm for the identification of privacy risks and their mitigation as part of the system design process (Cavoukian et al., 2009; Langheinrich, 2001). Addressing privacy risks early in the design process allows for embedding and integrating privacy protections into a system in ways that provide privacy protections without negatively affecting other system requirements (Cavoukian et al., 2009; Danezis et al., 2015), something that is much more difficult to accomplish once a system exists already (Schaar, 2010). A key aspect of privacy by design is data minimization, i.e., only collecting, processing and sharing data that is actually needed (Gürses, Troncoso & Diaz, 2011). Data minimization can be combined with a range of other privacy-enhancing technologies and organizational measures to protect privacy of a system's stakeholders (Danezis et al., 2015). Privacy impact assessments (PIAs) constitute a systematic, procedural approach for assessing privacy risks of a system and developing respective risk mitigations (Wright, 2013). PIAs are increasingly common components of software and product development processes and privacy compliance efforts (Wright & De Hert, 2012), Wright arguing early on that they should be mandatory (2011). In many countries, PIAs have become a required or recommended step for the design of information systems that process personally identifiable information (PII) (Wright & 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 entities collecting or processing data in Europe to conduct a data protection impact assessment when a new system is likely to pose privacy risks. In addition, the GDPR requires entities to provide data protection by design and by default (European Parliament and Council, 2016). Despite the general prevalence of PIAs, little guidance and examples exist for the assessment of privacy impacts in the context of digital repositories. We aim to fill this gap.

Privacy Impact Assessments for Digital Repositories

As discussed in the previous section, a Privacy Impact Assessment (PIA) is a systematic approach for determining privacy risks of an information system and mitigations for those risks (Wright, 2013). PIA's can be thought of as being similar to other risk assessment tools, including the DRAMBORA framework (a toolkit used to audit digital repositories for various types of risk (McHugh, Innocenti, Ross & Hofman, 2007)), but with a specific focus on privacy. A PIA typically involves mapping a system's information flows, determining what privacy risks exist for users of that system, and recommending strategies to mitigate those risks (Wright, 2013). PIAs are also used to determine legal compliance of the system: i.e., whether it adheres to the data privacy laws of the countries in which it operates. For digital repositories, PIAs can be used to reveal potential risks to both the research subjects who appear in research data managed by a repository and the researchers who use and access a repository. Both research subjects and repository users should be considered as data subjects, i.e., the individuals affected by a system's data processing practices. We adapt Wright's PIA methodology (2013), which synthesizes best practices for PIAs, into a PIA process consisting of five phases tailored for digital repositories.

In this section, we give an overview of the main steps involved in a typical PIA. This informs the following section (Case study: Privacy Impact Assessment of Researcher Passports), where we explain how we carried out each step when conducting a PIA on an existing digital repository and the researcher passport system being developed.

Phase I: Threshold Assessment and Preparation

The first phase of any PIA process is to determine whether a PIA is necessary, and if so, who should conduct the PIA and with what timeline, scope, and budget (Wright, 2013). If a system collects PII about individuals and/or if the information collected by the system could be used to harm these individuals a PIA is typically warranted (Wright, 2013; Wright & De Hert, 2012). In the case of digital repositories that contain information about people, a PIA is almost always warranted.

Phase II: Repository Description and Information Flows

The second phase is to map the system's components and information flows, i.e., what and how information is collected, stored, processed, transferred, and made available in different aspects of the repository or system of interest. This entails understanding and describing the repository's purpose, how the system works or will work, and the system's stakeholders; the information the repository collects from what stakeholder and why; how information is used or processed; how this information is stored and managed (e.g., is it stored in a database, is there a delete policy to delete information, can stakeholders update or delete their information from the system); and which stakeholders can access what information, and under what circumstances. Information flows can be identified through interviews with system designers, review of system documentation, and hands-on use of the system (if it exists already). The goal of this phase is to create a comprehensive overview of potentially privacy-sensitive information in a repository. Importantly, this should include a full consideration of all stakeholders whose data might be collected by the system, and who are therefore data subjects, or who may have access to information within the system. Stakeholders to consider include research subjects represented in research data, researchers and users accessing the repository, data curators integrating data into a repository, and administrative and IT staff with potential access to the data.

Phase III: Privacy Risk Analysis

Once the system, its stakeholders, and its information flows have been mapped, the next step is to identify, analyze, and characterize the potential privacy risks posed by the system. Assessing these risks may entail technical analysis of the system as well as additional interviews with stakeholder groups to understand their perspectives regarding the system, how they used or would use the system, and their privacy concerns and perceived risks regarding the system. Different stakeholders, such as research subjects, researchers, curators, and administrators may contribute different perspectives on potential privacy implications of the system. Information flows and interview findings are then used to develop scenarios in which information might be misused. Risks can come from a variety of places, including how that information is collected, who can access the information, and how that information will be used. The following questions are examples of those that might be used to identify risk scenarios:

- This information is being collected in this way. How might this harm stakeholders?
- What are the ramifications for a stakeholder that their information can be accessed by another stakeholder?
- What are potential ways that outside entities can access the information contained in the information flows? What are the consequences for the stakeholders involved?
- What are harms that might result from unauthorized access either by other stakeholders or unauthorized parties?
- This information is being used in this particular manner. How might this harm stakeholders?
- What are all the possible ways that this information could be used to harm stakeholders?

In developing risk scenarios, Wright (2013) recommends assessing the likelihood a given risk will occur, as well as the potential magnitude and severity of the risk. To assess legal compliance, relevant privacy laws should be examined and then, alongside examination of the information flows, it needs to be determined if the system complies with these laws.

Phase IV: Mitigation Strategies and Recommendations

After identifying risk scenarios and impacts, the next phase is to develop mitigation strategies. These can be technical (changes to the system) or organizational (internal and external policies) in nature (Danezis et al., 2015; Spiekermann & Cranor, 2009; Gürses et al., 2011). The following questions are examples of those that might be used to identify risk mitigations:

- What is the root cause of the identified privacy risk?
- Is this risk caused by the collection of certain information? Can the system function if this information is not collected or collected at a different level of abstraction?
- Is the risk caused by a certain entity accessing this information? Are there ways to prevent that entity from accessing this information without compromising the system's utility?
- Is the risk caused by unauthorized access to certain information? Are there ways to minimize the chances of unauthorized access occurring?
- Is this risk caused by a certain usage of certain information? Are there technical or organizational measures that can be implemented to ensure that information is not used this way?

It is advisable to confer with technical privacy experts and legal privacy experts to identify potential mitigation strategies (Hoepman, 2014) and their appropriateness and feasibility. 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 serve to evaluate and prioritize among possible mitigation steps. An important aspect in providing recommendations is to balance privacy considerations with system needs.

Phase V: Implementation, Publication, and Iteration

The PIA's findings should be documented in a report which should provide clear and prioritized recommendations to guide the implementation of privacy mitigations. The PIA report may also be made public to make transparent the steps taken to consider and protect privacy in a given system, which may positively affect the repository's trustworthiness. Importantly, a PIA report should be a living document (Wright, 2013). 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.

Case study: Privacy Impact Assessment of Researcher Passports

To provide a better sense of how a PIA can be utilized in and improve the design of digital repository systems, we present our privacy impact assessment of ICPSR's Researcher Passport system (Levenstein et al., 2018) as a case study. Early on in the project, we identified the need for a PIA because the centralized management of researcher credentials entails substantial collection, storage, and transfer of PII (Phase I). We formed an interdisciplinary team for the PIA consisting of domain experts in digital repositories and privacy experts. Next, we first provide an overview of the Researcher Passport system and its information flows (Phase II), followed by our analysis of potential privacy risks for different stakeholders associated with the proposed system (Phase III), and the mitigation strategies we developed to address those risks (Phase IV). We finally summarize the recommendations and guidance for the development of the Researcher Passports system (Phase V).

Researcher Passport: Overview and Information Flows

With more than 55 years of service to the social sciences, ICPSR is the largest archive of digital social and behavioral science data in the world. ICPSR curates, preserves, and disseminates original social science data for research, instruction, and policy evaluation. The organization archives over 10,000 data collections comprising 250,000 files of data and documentation, with millions of downloads each year. Membership has grown from 21 founding institutions in 1962 to more than 775 educational and research institutions worldwide in 2019.

ICPSR stores these datasets in ICPSR's Archonnex system, a proprietary Digital Asset Management System (DAMS). While most of ICPSR's collection is publicly available, more than 1,500 datasets are designated "restricted use," meaning that because of the proprietary, sensitive, or personally identifiable data they contain, stringent requirements are imposed for accessing these datasets to protect confidentiality on research subjects.

ICPSR has established policies and best practices around the use of and access to restricted data (by restricted data we mean the data in these restricted datasets). In the current environment, restricted data are often available to the research community only after individual researchers undergo an application and vetting process. Our prior analysis of the application process for 23 repositories finds that these processes are inconsistent, not only in what they require of researchers but even in how they define restricted data, modalities of access, and responsible and trusted data users (Levenstein et al., 2018). The conditions under which researchers access data depend on the interaction between characteristics of data, researchers, and institutions. This process is burdensome both for repositories who try to make data available responsibly and for researchers trying to use data. The complexity of the process creates opportunities for people and institutions to hoard research data and refuse to share, under the guise of protecting confidentiality, or to claim quite legitimately that it is simply too costly to share data safely.

The above has motivated efforts to develop the "Researcher Passport" system: a durable and transferable digital identifier for researchers that stores information that data repositories need to know, or need to verify, in order to make decisions about whether and how to provide access to their data.[1] 1. Researcher Passports contain information such as a researcher's education, training, institutional affiliation, the datasets they've used appropriately in the past, and their data security experience (Levenstein et al., 2018). The passport will represent and identify both a level of trust a given researcher has earned in responsible data use as well as provide an indication of which types of restricted data has earned requisite trust to access. Researchers will use their passports to apply for access to restricted data at ICPSR, and specialized ICPSR staff (gatekeepers) will review and update the passport when making access and dissemination decisions.

At the time the PIA was conducted, the passport system was in a prototype stage. Users could create a passport (done so through an online portal); this passport was stored in an internal database and could be accessed by ICPSR staff. However, the passport was not being used to access data repositories.

To begin our PIA for the Researcher Passport system, we first interviewed key stakeholders to begin identifying information flows. These stakeholders included the responsible project manager; a senior data project manager; an application manager; and two Freedom of Information Act (FOIA) officers, as some information in the system may be subject to Freedom

[1] Research Passport by ICPSR: https://radius.icpsr.umich.edu/radius/passport

Activity	Data source
Initial stakeholder interviews (n=3)	Project manager (1); Senior data project manager (1); Application manager (1)
Expert interviews (n=4)	FOIA officers (2); Privacy researchers (2)
System interaction	Use of Archonnex; interaction with ICPSR website; res ulting memos on information flows
Second round of stakeholder inter- views (n=8)	Potential users (3); Repository manager/"gatekeeper" (1); Official Representative (1); ICPSR staff in charge of managing data agreement infractions (3).

Table 1. Table summarizing the data collected in conducting the PIA process to determine the information flows and privacy risks of the Researcher Passport system.

of Information requests.[2] We also reviewed the project's description (see Levenstein et al. (2018)) and security assessments of ICPSR's broader data repository management system, Archonnex. Finally, we familiarized ourselves with the use and functionality of the Archnonex system, and in doing so documented what information had to be provided during use, what information was displayed, and any options for users to manage that information. See Table 1 for details.

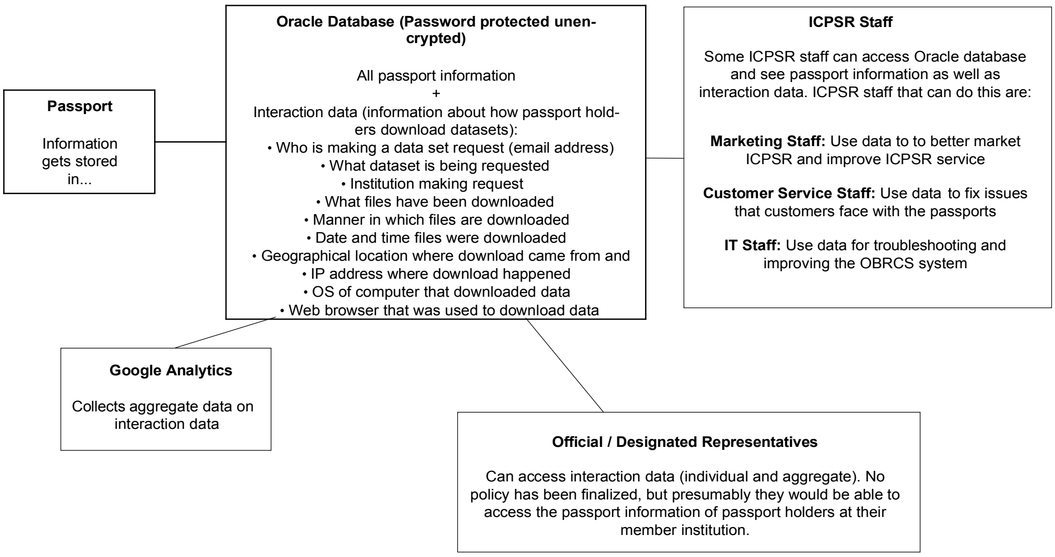
Passport
Information gets stored in...
Oracle Database (Password protected unen-crypted)
All passport information
+
Interaction data (information about how passport hold-ers download datasets):
• Who is making a data set request (email address)
• What dataset is being requested
• Institution making request
• What files have been downloaded
• Manner in which files are downloaded
• Date and time files were downloaded
• Geographical location where download came from and
• IP address where download happened
• OS of computer that downloaded data
• Web browser that was used to download data
ICPSR Staff
Some ICPSR staff can access Oracle database and see passport information as well as interaction data. ICPSR staff that can do this are:
Marketing Staff: Use data to to better market ICPSR and improve ICPSR service
Customer Service Staff: Use data to fix issues that customers face with the passports
IT Staff: Use data for troubleshooting and improving the OBRCS system
Google Analytics
Collects aggregate data on interaction data
Official / Designated Representatives
Can access interaction data (individual and aggregate). No policy has been finalized, but presumably they would be able to access the passport information of passport holders at their member institution.

Figure 1. Example information flow. Researcher information (e.g., name) will be transferred from a researcher into a passport (since the passport collected researcher names). The passport would transfer data into the database where this information would be stored and made accessible to other stakeholders (e.g., ICPSR staff).

The identified information flows were documented through system flow diagrams visualizing how information is transferred between different system entities and stakeholder. An

[2] In the US, a Freedom of Information Request is a written request by a member of the general public to access records from any federal agency. By law, the federal agency must comply with this request, although some exemptions exist (e.g., if disclosure of such information affects national security). Source: FOIA.gov (https://www.foia.gov/)

exemplar information flow is provided in Figure 1. We also created step-by-step process models showing information exchanges in different parts of the system. For example, we described in detail how a researcher creates a passport, how they enter information into the passport (and what information is entered) and thus become data subjects. A separate information flow was created to show how that information is stored in a database, the features of that database, and who can access that database and for what purposes.

The information flows were then used to prompt reflection on potential scenarios in which there may be risks to the different stakeholders and users of the system. We also used these information flows to develop interview protocols for additional stakeholders, to help us better understand potential risks in the system. We then interviewed eight additional stakeholder representatives (three researchers as potential users, a data repository "gatekeeper," three ICPSR staff managing data use agreements, one institution's official representative).

Researcher Passport Privacy Risk Analysis

In Phase III, our analysis of the system's information flows led to the development of twelve risk scenarios in which stakeholders' privacy could potentially be harmed. We grouped these scenarios based on whether they primarily affect passport holders (S1–S4), affect both passport holders and research subjects (S5–S8), primarily affect research subjects (S9–S10), or primarily affect ICPSR (S11–S12). Risk scenarios are summarized in Table 2; see Appendix A for full descriptions.

	Scenario	Harm Caused
S_1	Past dataset infractions lead to (unfair?) blocking	H_1
S_2	Passport holders' research ideas are stolen by other researchers	H_3, H_5
S_3	Negative impact on passport holders' career and reputation	H_3
S_4	Passport holders targeted for working with specific sensitive and politically charged datasets	H_2
S_5	Erroneous information on passport	H_1, H_4
S_6	A numeric score inaccurately judging passport holders	H_1, H_4
S_7	Bias against individual passport holders	H_1, H_4
S_8	Bias against passport holders of specific groups	H_1, H_4
S_9	Bad actors can use passport information to extract sensitive datasets from passport holder machines	H_4
S_{10}	People using other people's passports for applications allows unauthorized access	H_4
S_{11}	ICPSR's use of passport data decreases users' interest in the system	H_5
S_{12}	ICPSR's lack of data retention policy for interaction data decreases trust of system	H_5
S_{13}	ICPSR using passport information as an 'edge' over other repositories	H_5

Table 2. Table summarizing the identified risk scenarios.

Each risk scenario is associated with a potential harm – the specific privacy impact that could befall a stakeholder. These include:

- H_1: Passport holders being unfairly or mistakenly denied access to data. This could impair passport holders' ability to conduct research.

- H_2: 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.

- H_3: Reputational harm to passport holders. If a researcher is denied access to restricted data, unfairly or not, they could be perceived as being untrustworthy.

- H_4: Unauthorized access to research subjects' data. If a researcher is inappropriately granted access to a research subjects' data, this could potentially put that person at risk.

- H_5: Distrust of Researcher Passport, and by extension, institutions that use it. Any of the above harms coming to pass could cause researchers to lose trust in the Researcher Passport.

We estimated the likelihood of these risks to generally be low – however, the severity of any of these scenarios coming to pass could be high. In most scenarios, risks primarily impacted passport holders (e.g., credentialing information could be used against the researcher or may be exposed to others). Therefore, it is important to identify potential mitigation strategies to protect these users, and to thereby build trust in the Researcher Passport system.

Mitigation strategies

We developed a total of 17 mitigation strategies (see Table 3 for a summary; Appendix B for full descriptions). These can be grouped into design solutions (M1–M3), policy solutions (M4–M16), and technical solutions (M17).

Design and technical solutions all entail steps that could be "hard coded" into the Researcher Passport system. Examples include allowing passport holders to annotate past infractions on their passports to provide further context; implementing multi-factor authentication; and encrypting all passport data by default.

Policy solutions, on the other hand, involve changes to processes for working with data or stakeholders both inside and outside of the Researcher Passport system itself. We grouped policy solutions into three subgroups: dataset application solutions (M4–M6) passport holders' rights policy solutions (M7–M11), and internal ICPSR policy solutions (M12–M16). Dataset application solutions include changes that could be made to the process of applying for access to a dataset (e.g. allowing users to opt out of automated decision making). Passport holders' rights solutions entail giving passport holders control over their data or passports (e.g. allowing them to delete past interaction data after a period of time, or providing them with a process to appeal or request changes to records on their passport). Finally, ICPSR-level policy solutions include a number of internal audits that could be run to ensure that information is accurate, and decision making processes are transparent and fair.

	Mitigation	Addresses
M_1	Hide passport holder's name from gatekeepers during application process	S_2, S_7, S_8
M_2	Allow passport holders to add comments to explain infractions on passport	S_1, S_6
M_3	Implement multi-factor authentication	S_{10}
M_4	Opt out of automated decision making	S_1, S_5, S_6
M_5	Inform researchers of why a dataset application decision was made	S_1, S_5, S_6, S_7, S_8
M_6	Have appeals process for decisions related to dataset applications	S_1, S_5, S_6, S_7, S_8

M_7	Allow researchers to delete past dataset interactions from passport and database after retention period.	S_1, S_2, S_3, S_4, S_9
M_8	Enable passport holders to opt out of ICPSR data usage for non-application- approval purposes	S_{11}, S_{12}
M_9	Allow researchers to choose what of their passport information is shared	S_1, S_2, S_4
M_{10}	Allow researchers to remedy past dataset infractions from passport	S_1, S_3
M_{11}	Have process for requesting correction of information on passport	S_1, S_5, S_6, S_7, S_8
M_{12}	Audit policy for passport accuracy and veracity	S_5
M_{13}	Audit policy for internal decisions regarding applications	S_7, S_8
M_{14}	Strict punishments against sharing of passports	S_{10}
M_{15}	Have clear policy in place to know what to do in case of a passport data breach	S_2, S_3, S_4, S_9, S_{10}
M_{16}	Have clear, transparent policies regarding ICPSR data usage, storage, and retention	S_{11}, S_{13}, S_{12}
M_{17}	Encrypt database by default	S_2, S_4, S_9, S_{10}

Table 3. Table summarizing the identified mitigation strategies.

Resulting Recommendations

Some of the identified mitigation strategies are more feasible or impactful than others. Our resulting recommendations take these factors into account, as well as whether the mitigation strategy raises additional concerns. Because the Researcher Passport system is still being developed, some strategies cannot yet be fully implemented; however, they can be taken under consideration in development and future design work.

We recommend that M3 (Implement multifactor authentication), M9 (Allow researchers to choose what of their passport information is shared), and M17 (Encrypt database by default), be directly adopted, given their relative ease or straightforwardness of implementation. We further recommend that M15 (Have clear policy in place to know what to do in case of a passport data breach) and M16 (Have clear, transparent policies regarding ICPSR data usage, storage, and retention) be directly adopted, due to the large number of risk scenarios they address.

Several mitigation strategies require further development of policy and auditing procedures by ICPSR. We thus suggest that ICPSR clearly articulate its stance on several key questions:

1. What information about past DUA (data use agreement) infractions should the passport contain?
2. Should ICPSR allow passport holders to delete past dataset interactions after a set period of time? Or perhaps, automatically delete data about past interactions? Offering this option would not likely hurt ICPSR, but would increase the privacy, and therefore trust, in the system.
3. Should passport holders be able to opt out of certain aspects of the Passport system, such as automated decision making, etc.?
4. Will passport holders be allowed to remedy past DUA infractions? If so, how?

Discussion

Our work makes several contributions for the digital curation community. First, other repositories can use the workflow we adapt from Wright (2013), particularly as more institutions adopt policies or laws intended to protect user data. Though our PIA focused on privacy risks to researchers using a credentialing system for a digital repository, this method is appropriate for identifying and mitigating privacy risks for any sensitive information, including personally identifiable information (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 that host sensitive or restricted data. Other repositories relying on credentialing services should consider mitigations for the risk of unfairly denying researchers access to data – or to providing undue access to data. For instance, several of the mitigations proposed give users control over what aspects of their passport are shared or enable them to annotate past infractions. Allowing users control over their personal data and the ways in which it is shared will likely be important for other repositories, particularly those in countries with legislation in place like GDPR. Other mitigations, such as multi-factor authentication, can help repositories ensure that only authorized users are accessing their materials. Especially when they host PII and sensitive data, repositories have responsibilities to protect that data and manage access. MFA reduces risks of data leakage or breach that result from password sharing or hacking.

The information flows and harms related to those flows that PIAs produce make it effective for identifying privacy risks in distributed computing contexts. In the Researcher Passport context, a researcher’s PII is attached to a digital credential, which is passed between systems. However, there are other systems in which research data is passed between different servers for analysis or processing (e.g. multiparty computing), or in which research data stays in one place, but other algorithms or analytical processes are given access to the dataset for computation (e.g. the “non-consumptive research" paradigm as by the Hathi Trust Research Center (Jett, Cole, Maden & Downie, 2016)). There are also numerous repositories in which sensitive data is passed to a virtual digital enclave; though the enclave itself may be "safe" for the research data, it is important to ensure that the enclave is safe for the user as well. In each of these contexts, a PIA may help data and repository managers identify and mitigate potential risks to both researchers and research subjects.

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 disseminates. Ensuring the security of repository user data will become more important as research data volume increases, funding agency data management requirements develop, and repositories adopt credentialing systems like the Researcher Passport or develop new digital “enclaves” dependent on user profiles.

Conclusion & Future Work

Digital repositories need to consider the privacy of both research subjects and researchers. We demonstrated how digital repositories can conduct a privacy impact assessment via a case study of a researcher credentialing system. Three possible future directions for research arise. First, considering ethical social science data archive and repository behavior from a social justice lens (Punzalan & Caswell, 2016); as privacy protection challenges become increasingly complex, the focus on equitable agency, representation, restorative justice, and community could prove helpful to ensuring continued balance is struck between privacy and accessibility of research data, for both participants and researchers. Second, there is a need to continue to address the risks and rewards of “Big Data” for social science research (Mills, 2018), especially the extent to which increased computational computing capabilities undermine statistical-method-grounded methods for protecting data privacy and confidentiality. Finally, there are possible avenues of research in understanding of restricted social science data archives from a "socio-technical framework" perspective (Plale et al., 2019). The Researcher Passport described in this paper is an example of such a "socio-technical framework" through which one can understand restricted social science data archives. By paying attention to, and finding novel solutions to, complex privacy, confidentiality, and policy relationships in-and-between the data, the enclave, and the users, we hope to provide some contribution to the future of proper and equitable restricted social science (re)use.

Acknowledgements

This research was funded by the National Science Foundation under grant number 1839868.

References

Bancroft, T. (1972). The statistical community and the protection of privacy. *The American Statistician*, 26(4), 13–16.

Bishop, L. (2005). Protecting respondents and enabling data sharing: Reply to Parry and Mauthner. *Sociology*, 39(2), 333–336. doi:10.1177/0038038505050542

Bishop, L. (2009). Ethical sharing and reuse of qualitative data. *Australian Journal of Social Issues*, 44(3), 255–272. doi:10.1002/j.1839-4655.2009.tb00145.x

Bond, K., Berreman, G. D., Carroll, J. D., Coser, R. L., Douglas, J. D., Freidson, E., . . . others (1978). Confidentiality and the protection of human subjects in social science research: A report on recent developments [with comments and rejoinders]. *The American Sociologist*, 144–177.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059–1078. doi:10.1002/asi.22634

Cavoukian, A. et al. (2009). Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario, Canada.

Cliggett, L. (2013). Qualitative data archiving in the digital age: Strategies for data preservation and sharing. *Qualitative Report*, 18(1), 1–11.

Clubb, J. M., Austin, E. W., Geda, C. L. & Traugott, M. W. (1985). 'Sharing research data in the social sciences', In Fienberg, S. E., Martin, M. E., & Straf, M. L. Sharing research data. National Academy Press pp. 39–88.

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) (Magenta Book No. CCSDS 650.0-M-2). Retrieved from 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. (2007). Trustworthy repositories audit & certification: Criteria and checklist (Tech. Rep.No. 1). The Center for Research Libraries.

Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.-H., Metayer, D. L., Tirtea, R. & Schiffner, S. (2015). Privacy and data protection by design-from policy to engineering (Tech. Rep.). ENISA. Retrieved from https://www.enisa.europa.eu/publications/ privacy-and-data-protection-by-design

Donaldson, D. R. (2016, October). The Digitized Archival Document Trustworthiness Scale. *International Journal of Digital Curation*, 11, 252–270. Retrieved 2019-12-08, from http://www.ijdc.net/article/view/11.1.252 doi:10.2218/ijdc.v11i1.387

Donaldson, D. R. & Conway, P. (2015, December). User conceptions of trustworthiness for digital archival documents. *Journal of the Association for Information Science and Technology*, 66(12), 2427–2444.

Edsall, J. T. (1981). Two aspects of scientific responsibility. *Science*, 212(4490), 11–14.

European Parliament and Council. (2016, may). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved 2019-07-08, from http://data.europa.eu/eli/reg/2016/679/oj/eng

Geda, C. (1979). Social science data archives. *The American Archivist*, 42(2), 158–166.

Goldman, B. & Pyatt, T. D. (2013). Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives. *Library & Archival Security*, 26(1-2), 37–55.

Gürses, S., Troncoso, C. & Diaz, C. (2011). Engineering privacy by design. *Computers, Privacy & Data Protection*, 14(3), 25.

Hertzum, M., Andersen, H. H. K., Andersen, V. & Hansen, C. B. (2002, October). Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with Computers*, 14(5), 575–599. Retrieved 2019-12-08, from https://academic.oup.com/iwc/article/14/5/575/709406 doi:10.1016/S0953-5438(02)00023-1

Hoepman, J.-H. (2014). Privacy design strategies. In In IFIP International Information Security Conference (pp. 446-459). Springer, Berlin, Heidelberg.

Hofferbert, R. I. (1976). Social science archives and confidentiality. *American Behavioral Scientist*, 19(4), 467–488.

ICPSR. (2018) Retrieved from https://www.icpsr.umich.edu/files/about/researcher/ICPSR_ResearcherCredentialingWhitePaper_May2018.pdf

Jajodia, S. Abou El Kalam, A. & Sans, T. (Eds.), (2014) Ict systems security and privacy protection (pp. 446–459). Berlin, Heidelberg: Springer Berlin Heidelberg.

Jeng, W., He, D. & Chi, Y. (2017). Social science data repositories in data deluge: A case study of ICPSR's workflow and practices. *The Electronic Library*; Oxford, 35(4), 626–649. Retrieved from https://search.proquest.com/docview/1949952800/abstract/BB935487EFCB454DPQ/1

Jett, J., Cole, T., Maden, C. & Downie, J. (2016, March). The Hathi Trust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data*, 2(0), e1. Retrieved 2019-12-15, from http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.3/ doi:10.5334/johd.3

Kelton, K., Fleischmann, K. R. & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology*, 59(3), 363–374. Retrieved 2019-12-08, from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20722 doi:10.1002/asi.20722

Langheinrich, M. (2001). Privacy by design—principles of privacy-aware ubiquitous systems. In International conference on ubiquitous computing (pp. 273–291). Springer, Berlin, Heidelberg.

Levenstein, M. C., Tyler, A. R. & Bleckman, J. D. (2018). The Researcher Passport: Improving Data Access and Confidentiality Protection (Tech. Rep.).

L'Hours, H., Kleemola, M. & de Leeuw, L. (2019). CoreTrustSeal: From academic collaboration to sustainable services. *IASSIST Quarterly*, 43(1), 1–17.

McHugh, A., Innocenti, P., Ross, S. & Hofman, H. (2007). Drambora: The digital repository audit method based on risk assessment.

Miller, T., Birch, M., Mauthner, M. & Jessop, J. (2012). Ethics in qualitative research. Sage. Miller, W. E. (1969). The development of archives for social science data. *Quantitative Ecological Analysis in the Social Sciences*, 521–31.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? *Qualitative Research*, 18(6), 591–603. doi:10.1177/1468794117743465

Parry, O. & Mauthner, N. (2004). Whose Data Are They Anyway? Practical, Legal and Ethical Issues in Archiving Qualitative Research Data. *Sociology*, 38(1), 139–152. doi:10.1177/0038038504039366

Plale, B. A., Dickson, E., Kouper, I., Liyanage, S. H., Ma, Y., McDonald, R. H., . . . Withana, S. (2019). Safe open science for restricted data. *Data and Information Management*, 3(1), 50–60. http://dx.doi.org/10.2478/dim-2019-0005

Punzalan, R. L. & Caswell, M. (2016). Critical directions for archival approaches to social justice. *The Library Quarterly*, 86(1), 25–42. doi:10.1086/684145

Schaar, P. (2010). Privacy by design. *Identity in the Information Society*, 3(2), 267–274.

Spiekermann, S. & Cranor, L. F. (2009, Jan). Engineering privacy. *IEEE Transactions on Software Engineering*, 35(1), 67-82. doi:10.1109/TSE.2008.88

Tripodi, T. (1974). Uses & abuses of social research in social work. Columbia University Press.

Tyler, A. (2019). Facilitating access to restricted data: Operationalizing trust in data users. *International Journal of Digital Curation*. https://doi.org/10.2218/ijdc.v15i1.602

Wallis, J. C., Borgman, C. L., Mayernik, M. S., Pepe, A., Ramanathan, N. & Hansen, M. (2007). Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. In L. Kovács, N. Fuhr & C. Meghini (Eds.), Research and Advanced Technology for Digital Libraries (pp. 380–391). Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-74851-9_32

Wright, D. (2011, August). Should privacy impact assessments be mandatory? *Commun. ACM*, 54(8), 121–131. Retrieved from http://doi.acm.org/10.1145/1978542.1978568

Wright, D. (2013, October). Making privacy impact assessment more effective. *The Information Society*, 29(5), 307–315.

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

Yakel, E., Faniel, I. M., Kriesberg, A. & Yoon, A. (2013, June). Trust in Digital Repositories. *International Journal of Digital Curation*, 8(1), 143–156. Retrieved 2019-12-08, from http://www.ijdc.net/article/view/8.1.143 doi:10.2218/ijdc.v8i1.251

Yoon, A. (2014). End-users' trust in data repositories: Definition and influences on trust development. *Archival Science*, 14, 17–34. doi:10.1007/s10502-013-9207-8

Appendix A: Risk Scenarios

Risks to passport holders

S1: Past dataset infractions lead to (unfair?) blocking. Any Data Use Agreement infractions would appear in a researcher's passport. Applications for new datasets could be unfairly or prematurely rejected based on those infractions if details about the infraction are unavailable. The likelihood is moderate: the mere presence of an infraction could influence the gatekeeper's perception of the passport holder. The severity is moderate: passport holders may not access datasets important for their research. However, passport holders with egregious data misuses should not be allowed to access sensitive data.) The harm it leads to is H1.

S2: Passport holders' research ideas are stolen by other researchers. The passport stores what datasets a researcher has previously used and is currently working on. If a rival researcher accesses a passport, they could know what datasets the researcher is currently working on. The rival researcher may try to work with the same dataset and 'scoop' the original researcher. The likelihood is moderate: research rivalry exists in academia. The severity is high. The harms it leads to are H3 and H5.

S3: Negative impact on passport holders' careers and reputations. Any Data Use Agreement infractions would appear in a researcher's passport. People in charge of hiring decisions could request to see a job applicant's researcher passport and use any negative information to deny the passport holder a job or role. Similarly, leaked passports could harm a researcher's reputation if they contain negative information. For instance, researchers may be reluctant to collaborate with that passport holder. The likelihood is low: the chances others will be able to access the passport information is low; and moreover, chances are low people will weigh passport credentials over other factors, such as number of publications, when evaluating job applicants. The severity is moderate. The harm it leads to is H3.

S4: Passport holders targeted for working with specific sensitive and politically- charged datasets. The passport stores what datasets a researcher has previously used and is currently working on. Some of these datasets could be of politically charged or otherwise sensitive topics (e.g., the U.S. Trans Survey). Entities could target researchers working on specific topics, and information from the passport (such as address) could leave passport holders vulnerable to attack if it is leaked. The likelihood is low: few passport holders work on politically charged topics, and it is unlikely that entities could access this information. The severity is high. The harm it leads to is H2.

Risks to passport holders and research subjects

S5: Erroneous information on passport. Information on the passport may be inaccurate, whether it was incorrectly entered or verified or has gone out of date. Incorrect passport information could mean that either (1) passport holders are barred from accessing datasets they should be able to or (2) passport holders are granted access to datasets they should not be able to. The likelihood is low. The severity is high. The harms it leads to are H1 and H4.

S6: A numeric score inaccurately judging passport holders. One of the proposed features of the passport is to have a numerical 'score' associated with the passport, calculated by taking into account a passport holder's credentials, past uses of datasts, etc. This score would be used to determine access to datasets (the higher the score, the more sensitive datasets a passport holder could access). Scores may be inaccurate measures of a passport holder's trustworthiness – however, the presence of the score may mean that gatekeepers default towards the score and ignore other useful data that would be important to take into account. This in turn could mean passport holders are unfairly granted or denied access to certain datasets. The likelihood is moderate. The severity is moderate. The harms it leads to are H1 and H4.

S7: Bias toward individual passport holders. When reviewing an application, gatekeepers can see an individual passport holder's name. Gatekeepers could have personal relationships with applicants that influence their decision. The likelihood is low: presumably, there would be few cases of passport holders interacting directly with gatekeepers who have a personal relationship with them. The severity is high. The harms it leads to are H1 and H4.

S8: Bias toward passport holders of specific groups. When reviewing an application, gatekeepers can see an individual's passport holder's identity and may infer group characteristics of the applicant (such as race, gender, religion, institutional affiliation, etc.). Gatekeepers could have an explicit or implicit bias in favor or against certain groups that impacts their decisions when approving or denying applications. The likelihood is moderate. The severity is high. The harms it leads to are H1 and H4.

Risks to research subjects

S9: Bad actors can use passport information to extract sensitive datasets from passport holders' devices. The system stores interaction data (e.g., time at which dataset is accessed, geographical location, IP address, browser used, email, name of passport holder). If an actor gains access to this interaction data, the actor could use this data to help them attack a passport holder's device and extract dataset the passport holder possesses. The likelihood is low. The severity is high. The harm it leads to is H4.

S10: People using other people's passports for applications allows unauthorized access. Individuals could steal account details of passport holders, or passport holders might lend their passport credentials to others. Either scenario would allow unauthorized users to access datasets. The likelihood is low. The severity is high. The harm it leads to is H4.

Risks to ICSPR

S11: ICPSR's use of passport data reduces users' interest in the system. ICPSR may use passport data and interaction data in a number of ways (such as improving the service or marketing ICPSR to institutions). Passport holders whose data is collected may not view these data uses as legitimate, and may be uncomfortable with how their data is used. The likelihood is low: ICPSR uses data only for marketing, troubleshooting, and improving ICPSR service; uses that are probably not going to be seen as controversial by passport holders. The severity is low. The harm it leads to is H5.

S12: ICPSR's lack of data retention policy for interaction data decreases trust of system. The Researcher Passport system collects interaction data, and the time of analysis there was no apparent data retention policy or ability for passport holders to delete past interaction data. Passport holders may be uncomfortable with the idea that their interaction data is stored indefinitely. The likelihood is low. The severity is low. The harm it leads to is H1.

Appendix B: Mitigation Strategies

Design Solutions

M1: Hide passport holder's name from gatekeepers during the application process (addresses risks S2, S7, and S8). When passport holders apply to access certain datasets, information that could be used to discriminate against passport holders should be hidden from gatekeepers. The system would need to provide an anonymous channel for gatekeepers and applicants to communicate while protecting identities. This design solution mitigates risks of bias and could help protect researchers from being targeted for "scooping" attempts.

M2: Allow passport holder to add comments about infractions on passport (addresses risks S1 and S6). Although all data agreement infractions are serious offenses, not all data agreement infractions have the same severity: an accidental infraction (e.g., losing a laptop) is different than a purposeful infraction (e.g., selling a dataset to a corporation). To highlight the nuance of the different infractions, passport holders should be able to add comments to contextualize their infraction. Passport holder comments should be prominent and visible to ensure gatekeepers notice them (e.g., display comments in-line with infraction information, in a clear, easy-to-read format and font size). This design solution does not guarantee that information is accurate or that gatekeepers notice it.

M3: Implement Multi-Factor authentication (addresses risks S10). Multi-factor authentication (MFA) requires users to present multiple pieces of evidence (e.g., password, PIN number) to login to systems and could provide additional security overall. MFA would require the collection of additional information (e.g., phone numbers), adds a step for passport holders, and increases technical requirements (e.g., often requires a smartphone).

Policy solutions – dataset application process

M4: Allow passport holders to opt out of automated decision-making processes (addresses risks S1, S5, and S6) . The passport system could automate access decision-making for low-risk datasets. Passport holders may feel uncomfortable about having their access to a database be determined by an algorithm or confused about why an algorithm rejected their application. To alleviate their concerns, passport holders could opt-out of automated decision-making. The option to opt-out could help increase trust with the system but would decrease the efficiency of system.

M5: Inform passport holders of why a dataset application decision was made (addresses risks S1, S5, S6, S7, and S8). If a passport holder's request for access to a restricted dataset is denied, care should be taken to explain why. Either the gatekeeper or the automated system should produce a detailed report that passport holders could then access. Passport holders should also receive information about how to remedy the problems with their application and reapply, if applicable. This would increase the accountability and trust of the system and surface biases or unfair approval/rejection practices.

M6: Have an appeals process for application decisions (addresses risks S1, S5, S6, S7, and S8). When applications are denied, passport holders should be able to appeal the decision. Passport holders could include the decision report and present their case. A second gatekeeper would weigh both arguments and make a final decision. In case the decision was automated, passport holders could request human review. This would add accountability to the system and reduce the chances of unfair rejections. It also generates additional work for gatekeepers.

Policy solutions – passport holders' rights

M7: Allow researchers to delete past interaction data and dataset interactions from passport and database after retention period (addresses risks S1, S2, S3, S4, and S9). Currently, information on past dataset uses and interactions is stored in a database, without an apparent data retention policy. Passport holders could be given the option to delete such interactions after a substantial amount of time has passed (e.g., 5 years). Passport holders could either check a global option under account settings (e.g., "Delete all past data interactions after 5 years"), or could review individual interactions with datasets and delete each interaction one by one. Alternatively, this could be an automatic policy whereby all data interactions are deleted after a certain period of time, with sensitive (or politically charged) datasets being deleted sooner. This would allow passport holders to delete past dataset interactions which could be sensitive or put passport holders at risk, mitigating potential harms to them, and would increase trust with system. On the other hand, Past interactions never 100% fully gone (some evidence of interactions could remain, such as email conversation between gatekeepers and passport holders); interaction data could be valuable to ICPSR for improving service purposes, and if enough people delete past data, this could impact ICPSR.

M8: Enable passport holders to opt out of ICPSR data usage for non-application- approval purposes (addresses risks S11 and S12). Passport holders should be given the option to opt out of (or opt into) their data being used by ICPSR for non-application related purposes (e.g., marketing, or improving ICPSR service). Passport holders should be informed of these options when they first create an account, and be allowed to conveniently change their choice (for example, by having a button to opt in / opt out under 'Settings'). Moreover, this option should not limit or prevent passport holders from using the Researcher Passport system as normal. This would grant passport holders autonomy over how their data is used – which, in turn, would increase trust with the system. On the other hand, if too many passport holders opt out, then ICPSR may find it difficult to market or improve their service.

M9: Allow researchers to choose what of their passport is shared (addresses risks S1, S2, and S4). Passport sharing with passport holders and gatekeepers is currently a binary choice – either all the passport is shared, or none of it is. This could worry passport holders who want to share some parts of their passport (such as completion of a certain badge) but hide others (such as CV, or past datasets worked with). Providing passport holders with the options to choose what parts of their passport to share with others could be a step to solving this, so that passport holders could share useful information while withholding sensitive information they do not want to share. There are many ways this could be implemented. One way could be to allow passport holders to select in their profile what information is 'visible' information and what information is 'private': when a passport is shared, only the 'visible' information is shared. Another approach could have passport holders, when sharing their passport with someone else, fill a checkmark list where they can select what items of information are to be shared and which ones are not to be shared. This ability to limit what information is shared or kept private should not extend to gatekeepers: otherwise, passport holders could withhold information that gatekeepers need to make an accurate decision about an application. This would increase trust with system, and allow sensitive or otherwise reputation-damaging information to be kept hidden.

M10: Allow researchers to remedy past dataset infractions from passport (addresses risks S1 and S3). To avoid the consequences a past data infraction might have on a passport holders career, passport holders could be offered the chance to expunge past dataset infractions from their record and remove it to their record permanently. This "expunging" could happen automatically after a substantial time has passed (e.g., 10 years), or done by a passport holder after taking remedial action to prove that they are capable of handling data responsibly (e.g. undergoing extensive training on how to handle data securely). Additionally, the remedial action could change and adapt to the context and circumstances of the infraction: for infractions of the type of accidentally losing a laptop computer with sensitive data, one type of training is required; but for other infractions a different training is used. For more severe infractions (e.g., deliberately leaking sensitive data; attempting to re-identify research subjects), it could be that there is no expunging mechanism. This would prevent remediable infractions from the past from permanently affecting passport holder. On the other hand, it is unclear if remedial courses or trainings completely 'makes up' for past dataset infractions; additionally, some extremely sensitive datasets might be too sensitive even for people who have expunged their record.

M11: Have process for requesting correction of information on passport (addresses risks S1, S5, S6, S7, and S8). A lot of the information in the passport is entered manually by passport holders. Some information, such as credential information, needs to be verified or pulled in by outside entities before being present on a passport. Other information gets added by the system not the passport holder (such as past dataset infractions). If there is incorrect information on a passport, the passport holder should be able to request correction of that information. In the case of credential information that needs to be verified, passport holders could ask it to be verified again. In the case of information that is not added by the passport holder, passport holders could appeal for it to be corrected, in which case the information would be reviewed by a human entity (e.g., dedicated ICPSR staff). This would reduce the chances of incorrect information being present in passport. On the other hand, it could add extra burden for staff and verification services.

Policy solutions – internal to ICPSR

M12: Audit policy for passport accuracy and veracity (addresses risk S5). To ensure that passport holder entered data on the passport is accurate, ICPSR should have audit policies that review the information in a passport and ensure it is up-to-date. This could be done in several ways, for instance, by contacting the institution the passport holder claims to belong to; verifying that publications exist; or interviewing the passport holder directly and contrasting their answers with what appears in their passport. The verification mechanisms should be periodically reviewed and audited. If verification mechanisms are found to be faulty (e.g., exhibit high false positive rates), then alternate verification methods should be found. Auditing the verification processes could involve having designated staff interview and carefully examine the verification mechanisms; alternatively, staff could try to create a fake passport with fake credentials, and see whether the verification mechanisms are able to detect the fraud. This would ensure that passport information is accurate. On the other hand, any auditing process is likely to be expensive and time-consuming, placing extra burden on staff; and it could be privacy invasive or otherwise inconvenience passport holders.

M13: Audit policy for internal decisions regarding applications (addresses risks S7 and S8). With the advent of the passport system, it will be easy to track who has made what requests, and whether these requests have been approved or rejected. The passports could facilitate internal auditing to check and see whether applications are approved or rejected in a fair fashion. For example, the data collected about what datasets passport holders have requested could be used to check if disparities exist between rates at which passport holders from community colleges get rejected vs. Ivy leagues – if a disparity is found, then the matter could be investigated more in depth to see if there is an unfair bias in the approval process. To perform this, an auditing policy could be in place where once a year all the application decisions are reviewed, and the application approval rates of different groups could be compared (e.g., ivy league colleges vs. community colleges). More ambitiously, 'fake' passports could be created that are exactly the same in terms of credentials but differ in group affiliation (e.g., a female sounding name vs a male sounding name) and seeing whether applications are approved at the same rates. This is an approach that has been used to evaluate differences in job offers for resumes of individuals belonging to different groups, such as white sounding names vs black sounding names. This approach would require gatekeepers to take time evaluating false applications, so it might not be the optimal approach. This would increases accountability and inclusion of system, as well as increase trust with system. On the other hand, such an auditing process is likely to be expensive and time consuming, and it could place additional burden on gatekeepers and ICPSR staff.

M14: Strict punishments against sharing of passports (addresses risk S10). To address issues of passport holders sharing their passport with others to allow them to access datasets, there could be a policy in place to punish and mitigate this behavior. For example, passport holders who share their passport could have a mark on their passport. SImilarly, passports that have been found to be shared could have to go extra verification steps before being approved for requests. These policies should be made explicit to passport holders, to discourage passport sharing. This would reduce the chances of passport sharing and associated harms. On the other hand, this strategy does not address the issue of detecting when passport sharing is occuring (this mitigation only addresses passport sharing after it has been discovered); it would be difficult to differentiate between voluntary sharing of passport versus unauthorized access to passport; and the addition of this negative mark on passport can raise new risks (e.g., S1 (Past dataset infractions lead to (unfair?) blocking) or S3 (Negative impact on passport holders career and reputation)).

M15: Have clear policy in place to know what to do in case of a passport data breach (addresses risks S2, S3, S4, S9, and S10). In cases that passport information is leaked to external sources, there should be policies in place to inform actions ICPSR should take so as to mitigate harms. Such policies should include mechanisms so that the passport holders whose information is leaked are informed of the breach, and passport holders are given information regarding what data has been exposed, the risks that the exposure has for the passport holders, and what steps (if any) passport holders can take to mitigate the consequences of the breach. This would help reduce the fallout in case of a data breach or data exposure.

M16: Have clear, transparent policies regarding ICPSR data usage, storage, and retention (addresses risks S11, S12, and S13). ICPSR should have clear and transparent policies regarding its data practices. Additionally, ICPSR should inform what these policies are to passport holders in ways that are clear, readable, and understandable. Care must be taken to ensure such policies comply with relevant state and country laws. This transparency can help reduce the uncomfortableness of passport holders over how their data is used, stored, and managed, and the increased transparency would increase trust in the system.

Technical solutions

M17: Encrypt database that stores passport information by default (addresses risks S2, S4, S9, and S10). Currently, the database that houses passport information and interaction data is password protected, but not is encrypted. We advocate encrypting it by default, since this reduces the chances of unauthorized access to information within the database. This would reduce the chance of unauthorized access to passport information and interaction data.