

Partnership



Canadian journal of library and information practice and research

Revue canadienne de la pratique et de la recherche en bibliothéconomie et sciences de l'information

PARTNERSHIP

"Carefully and Cautiously": How Canadian Cultural Memory Workers Review Digital Materials for Private and Sensitive Information

Avec prudence et délicatesse : comment les travailleurs canadiens de la mémoire culturelle examinent les documents numériques à la recherche d'informations privées et sensibles

Jess Whyte  and Tessa Walsh 

Volume 19, Number 1, 2024

URI: <https://id.erudit.org/iderudit/1115783ar>

DOI: <https://doi.org/10.21083/partnership.v19i1.7180>

[See table of contents](#)

Publisher(s)

The Partnership: The Provincial and Territorial Library Associations of Canada

ISSN

1911-9593 (digital)

[Explore this journal](#)

Cite this article

Whyte, J. & Walsh, T. (2024). "Carefully and Cautiously": How Canadian Cultural Memory Workers Review Digital Materials for Private and Sensitive Information. *Partnership*, 19(1), 1–26.

<https://doi.org/10.21083/partnership.v19i1.7180>

Article abstract

This exploratory study is based on semi-structured interviews with digital preservationists working in Canada. The goal was to understand how participants are reviewing sensitive personal information in their digital materials and identify the challenges they face in this work. Findings include a summary of current practices in place (e.g., risk profiling, using software tools for triage, and consultation with donors and/or community), and a list of challenges (e.g., IT or systems infrastructure, scale, barriers to access for software tools, and funding restrictions). The implications of these findings are that Canadian memory workers require time, staff, support, and training, while developers of assistive-software tools for this work should address tool limitations and prioritize user-friendly interfaces. Grant funding stipulations should also be revised to allow for diverse access options, while investment in controlled access repositories would enable responsible online access to complex collections.

© Jess Whyte and Tessa Walsh, 2024



This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

<https://apropos.erudit.org/en/users/policy-on-use/>

érudit

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

<https://www.erudit.org/en/>

PARTNERSHIP

The Canadian Journal of Library and Information Practice and Research
Revue canadienne de la pratique et de la recherche en bibliothéconomie et sciences de l'information

vol. 19, no. 1 (2024)
Theory and Research (peer-reviewed)
DOI: <https://doi.org/10.21083/partnership.v19i1.7180>
CC BY-NC-ND 4.0

"Carefully and Cautiously": How Canadian Cultural Memory Workers Review Digital Materials for Private and Sensitive Information

Avec prudence et délicatesse: comment les travailleurs canadiens de la mémoire culturelle examinent les documents numériques à la recherche d'informations privées et sensibles

Jess Whyte
Digital Assets Librarian
University of Toronto
jessica.whyte@utoronto.ca

Tessa Walsh
Senior Applications and Tools Engineer
Webrecorder
tessa@webrecorder.org

Abstract / Résumé

This exploratory study is based on semi-structured interviews with digital preservationists working in Canada. The goal was to understand how participants are reviewing sensitive personal information in their digital materials and identify the challenges they face in this work. Findings include a summary of current practices in place (e.g., risk profiling, using software tools for triage, and consultation with donors and/or community), and a list of challenges (e.g., IT or systems infrastructure, scale, barriers to access for software tools, and funding restrictions). The implications of these findings are that Canadian memory workers require time, staff, support, and training, while developers of assistive-software tools for this work should address tool limitations and prioritize user-friendly interfaces. Grant funding stipulations should also be revised

to allow for diverse access options, while investment in controlled access repositories would enable responsible online access to complex collections.

Cette étude exploratoire se base sur des entretiens semi-dirigés avec des préservateurs numériques travaillant au Canada. Le but est de comprendre comment les participants examinent les informations personnelles sensibles dans le matériel numérique et d'identifier les défis auxquels ils font face dans leur travail. Les résultats comprennent un sommaire des pratiques courantes qui sont en place (par exemple, profilage des risques, utilisation de logiciels pour le triage, consultation avec des donateurs et/ou la communautés) et une liste des défis (par exemple, les TI ou l'infrastructure des systèmes, l'ampleur, les obstacles à l'accessibilité des logiciels et les limites de financement). Les implications de ces résultats sont que les travailleurs canadiens de la mémoire nécessitent du temps, du personnel, du soutien et de la formation, tandis que les développeurs de logiciels d'assistance pour ce travail devraient se concentrer sur les limites des outils et donner la priorité à des interfaces conviviales. Les conditions des agences subventionnaires devraient également être révisées afin de permettre diverses options d'accès tandis que les investissements dans des dépôts à accès contrôlé permettraient un accès en ligne responsable à des collections complexes.

Keywords / Mots-clés

preservation, sensitive information, privacy, digital; préservation, informations sensibles, confidentialité, numérique

Introduction

The presence of private and sensitive information in digital collections has been shown to have a chilling effect, leading to materials being stuck in processing backlogs and inaccessible to researchers. This effect is caused by the challenge of identifying such information and the potential legal, ethical, and reputational risks associated with accidentally releasing personally identifiable information (PII) and other sensitive data (Goldman & Pyatt, 2013). Review is needed, but the extent of material in digital collections makes traditional review tactics seem daunting (Hutchinson, 2018). To understand why and document how stewards are working with these materials, we asked Canadian librarians and archivists how they are reviewing digital materials for sensitive personal information, what they are doing with that information, and what challenges they are facing.

Literature Review

To find relevant literature for the topic, the authors conducted searches in several prominent archives and library databases as well as Google Scholar. This was done with the intention of finding literature describing the problems memory workers face with private and sensitive information, as well as case studies and research about how such information might be successfully managed. References in the literature found also

proved fruitful in understanding the breadth and current understanding of the problem in the cultural heritage field.

Defining Private and Sensitive Information

It is perhaps helpful to begin with a definition of what constitutes private and sensitive information. In the context of digital archival records, Lee and Woods (2012) offer the following definition: “Any data that are personally identifying, could be used to establish the identity of the producer, establish the identity or personal details of individuals known to the producer (e.g., friends, family, and clients) or are associated with a private record (e.g., medical, employment, and education)” (p. 299).

The most straightforward category of private and sensitive information is PII, such as Social Insurance Numbers and their international equivalents, credit card numbers, tax records, health records, research data, and student records (Goldman & Pyatt, 2013). Such records are often legally protected by local, provincial, or national statutes, such as the *Privacy Act* and the *Personal Information Protection and Electronic Documents Act* in Canada (Office of the Privacy Commissioner of Canada, 2023).

Other information may be sensitive without being legally protected. As Ry Moran (2016), founding Director of the National Centre for Truth and Reconciliation and current Associate University Librarian – Reconciliation at the University of Victoria, writes, records of Residential Schools and other trauma inflicted upon Indigenous peoples by settler states additionally “contain information that could cause discomfort, harm or embarrassment to individual survivors” (p. 1). Krista McCracken and Skylee-Storm Hogan (2021) further explain that

archival materials relating to residential schools document the abuses, neglect, systemic racism, and other atrocities that occurred as part of the residential school system. Additionally, even archival records such as colonial administrative correspondence, which might seem benign, can actually be representative of historical trauma and be triggering to Indigenous archival users and archival staff. (p. 6)

Indigenous records may also be subjected to cultural protocols that limit who can access particular information and when (Kim, 2019).

In their article “From Human Rights to Feminist Ethics: Radical Empathy in the Archives,” Michelle Caswell and Marika Cifor (2016) document a similar case study of private information that was deemed too personal to be digitized in the South Asian American Digital Archives (SAADA) despite not being legally protected:

While digitizing a collection of papers related to Vaishno Das Bagai, an early Indian immigrant to the United States, [the first author] came across Bagai’s personal suicide note, dated 1928, addressed to his wife and sons, marked at the top with red ink, underlined, and in capital letters: “NO ONE ELSE SHOULD READ THIS.” Although Bagai had been dead for nearly 85 years, and his granddaughter who was donating the collection may have granted permission to digitize the note, the first author felt an affective responsibility to maintain Bagai’s privacy. Out of a sense of empathy with and

care for Bagai, developed over the course of processing his collection, the first author did not digitize the private suicide note. (p. 34)

Caswell and Cifor (2016) contend that affective responsibility should extend not only to the creators of archival records but to their subjects, users, and related communities: “In a relationship of caring, we must balance our desire to capture histories that would otherwise be silenced in the archival record with the privacy, desires, and needs of the subjects of our records” (p. 37). Their article gives examples of affective care with records that have a relationship to marginalized peoples who are subject to “more subtle, intangible, and shifting forms of oppression” (p. 27), including LGBTQ individuals who may be inadvertently outed or be caused discomfort by language used to describe and provide access to collections and who thus may not want those records to be freely available online. Such a responsibility would expand the definition of private and sensitive information to be contextual to a collection, its creators and subjects, and the communities to which it relates.

Difficulties with Private and Sensitive Information in Digital Collections

The literature paints a clear picture of how identifying and managing private and sensitive information in digital collections is challenging along several axes. These challenges begin with the scale and affordances of digital collections themselves. Goldman and Pyatt identify several case studies of how archivists have struggled with this issue in their 2013 article “Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives.” One example given is the Salman Rushdie papers at Emory University, where archivists had to balance their desire to make Rushdie’s correspondence available for research against the very real risk of threats made to the author’s life. Staff at Emory initially began to review the correspondence with the intention of redacting private information but quickly found that the amount of work required for such an approach was beyond what their staffing levels and schedule would allow, ultimately leading the archivists to restrict the complete set of correspondence. Goldman and Pyatt give another example of the STOP AIDS Project records at Stanford University, where archivists used digital forensic tools and keyword searches to identify private and sensitive information but felt unsure about their results and what constituted due diligence. Such information can be difficult to locate and identify because digital collections tend to be more voluminous than their physical analogs (Goldman & Pyatt, 2013; Hutchinson, 2018). Additionally, private information can often be found in unexpected places on digital storage media, such as log files, system files, unallocated files (which users may have thought were permanently deleted but remain forensically recoverable), and a computer or drive’s swap and slack spaces (Barrett, 2017; Lee & Woods, 2012). A digital collection’s donor and even creators are frequently unaware of such remnants.

Reviewing digital correspondence such as text messages, emails, or direct messages also poses a significant challenge to memory workers. Email inboxes frequently contain hundreds of thousands or even millions of messages, which makes manual review time- and cost-prohibitive. As just one example, a project at the National Archives and Records Administration in the United States to review the collection of Justice Elena

Kagan's email took a team of over 20 archivists and technicians over 6,000 hours to make 75,000 messages available for research (Schneider et al., 2017).

Such concerns are examples of the unexpected difficulties of translating existing workflows for physical collections into those appropriate for digital collections. In her Brodsky Series lecture, "Translating Bits: Maintaining (Born-) Digital Heritage," Monique Lassere (2021) described the challenges she faces doing digital preservation work and placed them under an umbrella of translation. For example, Lassere points to the complexities of translating preservationists' needs to an IT department or translating manual processes to technically assisted ones.

Even when digital collections arrive in such a way that obviates concerns about hidden or deleted data, low staffing levels can be a significant obstacle. According to a Canadian Association of Research Libraries (CARL-ABRC) study by Hurley and Shearer (2019), "There are low staffing levels devoted to digital preservation at many organizations. Looking at total Full Time Equivalent (FTE) values across respondent organizations, positions with responsibilities for digital preservation represent less than 1% of total organizational FTEs" (p. 3). For those working in the field, this finding does not come as a surprise; the current allocation of staff to ethically steward rapidly increasing digital collections is insufficient, and this extends to learning and applying new tools and workflows (Lassere, 2020; Lassere & Whyte, 2021).

Reviewing sensitive records can also be emotionally difficult for staff who must engage with the content of such records. Processing collections can be a form of witnessing, a process that requires empathy and emotional work (Cifor, 2016; McCracken & Hogan, 2021). Experiencing those emotions can be both draining and traumatic for staff, and Verne Harris (2014) notes that burnout is common among staff who work with archival materials documenting human rights abuses and trauma.

Existing Practices

There are several examples of archival workflows that may include the identification of sensitive information (Faulder et al., 2018; Post et al., 2019; Sloyan, 2016). However, many of these either look at the whole intake or accession process or are formatted as individual case studies.

Goldman & Pyatt (2013) give two examples, discussed above, of attempting to apply digital forensic tools to identify sensitive information with limited success at Emory University and Stanford University. Application of forensic techniques is also employed at Dalhousie University, where staff begin a review process of digital archives by filtering out duplicates and software or system files, such as files with checksum hashes found in the National Software Reference Library, to focus their attention on a more limited number of records that require further investigation (Barrett, 2017).

Recent approaches include the introduction of machine learning tools. Tim Hutchinson at the University of Saskatchewan has been writing on machine learning in archives since 2017. In a pair of articles published in 2017 and 2018, Hutchinson documented

his efforts to train a topic model to find Human Resources-related documents in the records of a university administrator. In his work, Hutchinson (2018) determined that “supervised machine learning could be a viable approach for a ‘triage’ method of reviewing collections for restrictions. At the very least, it could help measure the risk of whether documents requiring restrictions are to be found in a large document set” (p. 2699). In 2020, Hutchinson broadened his scope and gave us a comprehensive literature review, a list of other attempts to use National Language Processing (NLP) and Machine Learning (ML) techniques, and a review of existing tools such as ePADD and BitCuratorNLP. His 2020 paper also provides some recommendations, arguing “effective NLP and machine learning tools for archival processing should be usable, interoperable, flexible, iterative and configurable” (p. 1).

Machine learning tools rely on human teachers, which introduces variability that can affect results. The University of Illinois and the Illinois State Archives recently reviewed predictive coding tools to support the identification of sensitive information in email accessions. These tools included Microsoft’s Advanced e-Discovery, OpenText’s Recommind, FTI Consulting’s Ringtail, Luminoso Analytics, ePADD, and the TAR Evaluation Toolkit. A challenge in their review process was “the variability inherent in human judgement” and the need for the development of “content-tagging protocols that can be consistently applied by human reviewers” (Kaczmarek & West, 2018, p. 5). Predictive coding relies on iterative training processes. Variability in what is deemed sensitive by different reviewers, or teachers, creates a longer learning process and influences the output. The output can therefore become less reliable to some, increasing the dissonance between what is predicted to be sensitive by tools like Ringtail and the judgement of human reviewers.

Ethics has also been a topic of interest as approaches have matured. In their work on the ethics of online access, Cristela Garcia-Spitz and Noah Geraci (2021) discuss how they address access challenges at the University of California San Diego and University of California Riverside. Their strategies include utilizing click-through warnings or agreements in online access tools, implementing granular metadata to establish access controls, and not including certain materials in online indexes. They also lament the requirements of aggregator services that demand all materials be available publicly, or off-the-shelf repositories that require further development to maintain custom restrictions. These technical policy limitations are fixable but require communication with developers and product owners.

Calls for tiered access controls are also presented by Arroyo-Ramírez et al. (2020) in *Levels of Born Digital Access* and by Baker et al. (2023) in *The Exploration of Access Systems Framework*, the latter of which includes a list of specifications for access systems such as restrictions that “can flexibly accommodate cultural restrictions, ethical and legal privacy concerns, trauma informed access, and donor requests as well as local institutional policies” (p. 2). Some degree of such access controls already exists in digital collections tools. For example, Artefactual’s Access to Memory (AtoM, an open source, standards-based, archival description and access system developed in Canada, allows for granular controls based on user groups and PREMIS rights metadata, features which help the National Centre for Truth and Reconciliation (NCTR) describe

and provide granular access to their digital archives (Artefactual, 2020; NCTR, 2021). The Dataverse Project, an open-source web application to preserve and share research data, also allows for restriction at the file level and the ability to open files and datasets to certain groups of users or require acceptance of a Terms of Access agreement before allowing access (The Dataverse Project, 2022).

Finally, the literature also covers methods of identifying and managing private and sensitive information through consultation and relationship building with the individuals and communities that are themselves the creators and subjects of records. In 2016, Ry Moran wrote, “we need to continue to use materials carefully and cautiously, engaging with the communities affected, consulting relevant stakeholders and challenging the fundamental power structures that lay beneath the questions of who has access to and control over records by and about Indigenous people. This approach is a path to reconciliation itself” (p. 2). Krista McCracken and Skylee-Storm Hogan (2021) document such practices undertaken at the Shingwauk Residential Schools Centre (SRSC) community archive, where outreach practices such as “community-based healing projects centered around sharing personal experiences about the Residential School Project ... regional healing circles, community-driven dialogue sessions and the development of a Survivor network” (p. 9) connect survivors of the residential school system to archival materials. The SRSC also invites the affected community into the process of making policy and decisions related to the archives, often “bring[ing] the voices of the Survivor community to committees and decision-making bodies within the institution” (p. 11). As a result of these deep community connections, the SRSC “provides an example of how a Survivor community has transformed trauma-related archives into a space that serves the needs of Survivors and their families” (p. 14).

Aims of the Study

The literature shows that the presence or even potential presence of private and sensitive information in digital collections can be problematic for memory workers managing and providing access to these collections, and despite interesting and helpful approaches tested in case studies, the problem is far from solved.

To better understand both the extent of the problem and potential approaches, the authors conducted a series of interviews with Canadian librarians and archivists about their and their institutions’ experiences. Namely, we wanted to know what types of private and sensitive information are found in digital collections in Canada, how those collections are being reviewed to locate the sensitive information, how that sensitive information is being managed in relation to preservation and access to the collections, and what challenges Canadian memory workers and their institutions are facing.

Data Collection Method

This qualitative, exploratory, single-method study is based on 11 semi-structured interviews conducted over Zoom, a video conferencing platform, between Sep 2020 and May 2021 (interview guide included as Appendix A). Ethics approval was obtained from the University of Toronto’s Research Ethics Board.

Interview participants are digital preservationists working in Canada. Participants were recruited directly as known colleagues who may work with sensitive information. This purposive sampling approach was chosen due to the small number of people in Canada able to serve as participants. We also tried to represent a range of institution types, locations, and sizes when recruiting participants. Participants came from The ArQuives, Concordia University, Dalhousie University, Library and Archives Canada, the National Centre for Truth and Reconciliation records at University of Manitoba, the Shingwauk Residential Schools Centre at Algoma University, Simon Fraser University, University of Saskatchewan, University of Victoria, the City of Vancouver Archives, and York University. Participants were also representative of different roles and experience levels, including at least one archivist, librarian, director, and technical specialist. One potential participant declined our invitation because they felt uncomfortable sharing not-yet-formalized work processes. Our initial goal was to interview 8-12 participants based on our representation needs and the size of the field, with the final number being 11 participants. The scope of materials discussed included special and general collections, closed and open records, research data, and digitized content. Interviews were conducted with both researchers present. We treated the first interview as a potential pilot interview and were prepared to adjust the interview script, but the questions worked well, and no subsequent changes were made.

We obtained informed consent from interview participants, shared the semi-structured interview questions in advance, disclosed how recordings and transcripts would be used, allowed participants to review and edit the interview notes, and offered them an opportunity to review the manuscript before submission. Admittedly, this placed more labour upon participants, but it was also up to them how much, if any, review they wanted to do. Our goal was authentic, reflected-upon responses, and we felt getting participants' confirmation of our understanding was important to the process. We also wanted to give participants the opportunity to examine and comment on our findings.

Participants knew that they would not have anonymity, as the field is simply too small, and their names and institutions would be published. We also wanted to acknowledge their participation and provide context for the reader where needed. In this paper, most participant quotes are anonymous, but some are attributed to specific people or institutions where relevant, such as where the context of the collection and/or its associated risks might affect the readers' understanding. Participants were able to withdraw at any time up to a month after reviewing their transcripts and interview notes.

Data Analysis Method

Following the methodology presented by Ryan and Bernard (2003), analysis of the interview notes and transcripts to identify themes relied on cutting, sorting, and looking for similarities and differences between participants. Memo writing was also used to develop themes using the processes outlined by Emerson et al. (1995). Analysis involved the following steps:

1. Compiling the research data (interview notes and transcripts) into physical documents for review

2. Initial, independent reading of the research data
3. Discussion of potential themes to create a shared list
4. Creation of a shared comparison table to identify similarities and differences between questions asked of all participants
5. Second independent reading using the early themes from step 3 as a lens to identify relevant moments, annotating them with comments and colour coding
6. Writing memos or summaries by theme and by participant
7. Discussion to decide on most relevant themes to include
8. Final independent reading, focused on relevant themes
9. Discussion to finalize which themes and snippets to include here (for this step, we used a shared Miro board (a virtual whiteboard platform) to organize our discussion and sort our data)
10. Seeking confirmation from participants by asking them to review our findings before submission

Findings

Our findings are organized into two categories (current practices and challenges) and 13 themes within those categories:

Category	Theme
1. Current Practices	
	1.1 Relying on donor or community consultation
	1.2 Creating risk profiles
	1.3 Reviewing on access
	1.4 Using software for triage
2. Challenges	
	2.1 Reviewer trauma and audio-visual collections
	2.2 Higher risk of putting materials online
	2.3 Third-party information

	2.4 Fear of censoring
	2.5 Friction between theory and practice
	2.6 IT or infrastructure challenges
	2.7 Resources
	2.8 Unique challenges of correspondence
	2.9 Lacking trust and confidence in software tools

Participants mentioned finding a wide range of types of private and sensitive information in their digital collections, including legally protected private information such as medical, financial, and student records; private information about third parties; private business and employment documents; images of children; materials such as local in-community zines that were never intended by their authors to be shared online to a broad audience; classified and unreleased government records; culturally protected materials; and records documenting trauma. Participants also mentioned concern that documents which may not be considered particularly sensitive when reviewed individually in a reading room setting might be able to cause harm when made available online in aggregate due to the potential for computational approaches to the data.

Category 1 – Current Practices

Most participants did not have rigid workflows for identifying private sensitive information, partly due to a lack of broadly appropriate tools and partly because the work relies heavily on professional judgement. It's hard to map concepts such as "let our contextual knowledge of the collection guide effort" to a diagram or decision tree. That said, there are clear strategies in place, many of which rely on knowledge, judgement, and relationships. Though they may not be documented in workflow diagrams, they serve these participants well. The following sections further describe these strategies.

1.1 Relying on donor or community consultation

Many participants cited donor or community consultation as a dominant strategy for either identifying or reducing sensitive information. Consultation with donors helps flag culturally sensitive or private information during the transfer process, while community consultation provides guidance on access and restrictions. As one participant said, they "work with the donors to ensure that any sensitive cultural information or private information is flagged during the transfer process." Similarly, another participant told us, "usually the practice is to discuss with [donors] any potential restrictions that they could suggest. So, the more they can flag for us, the better."

Consultation was also used to answer questions about access, particularly when choosing materials for public release. An example of this approach can be found at the Shingwauk Residential School Centre (SRSC):

If we're unsure around access to any documents, particularly if it happens to be culturally sensitive or it might be sensitive to a particular community, we turn to the Children of Shingwauk Alumni Association [the group of survivors associated with the Shingwauk Residential School and their families] for their advice.

1.2 Creating risk profiles

Risk profiling was another commonly used practice, where participants assess the likelihood and potential impact of sensitive information based on factors such as the source, format, or content of the materials. Risk profiles typically focus on the context of digital records rather than the information of each individual document. Given the scale of many digital collections, some participants use risk profiles to guide where they should look for close reading and human review. This practice often relies on the previously mentioned donor and community consultation.

Risk profiles may be based on who is donating the materials: "There [are] particular offices that we know when they transfer records to us that it will likely contain sensitive information and they're processed accordingly." In other cases, profiles may be based on the format, such as non-textual content:

For sure our audio-visual records, those definitely tend to have a higher likelihood of containing this type of information. So, anytime we go to release new audio-visual content we have to review it in full or review any transcript that may be available for that video file or audio file.

Profiles may also be based on the type of content, such as case records. For example, case records from frontline health and social work organizations may be flagged as having a higher likelihood of third-party content or sensitive information. This approach requires knowledge of the community, the source, the record types, and the contents: "It's usually just based on knowing, like having it be a large community but still kind of small and knowing."

For some participants, broad risk profiling is not always adequate. However, the level of review, from a sampling method to line-by-line review, may still depend on the type of record or the level of access: "For that particular process [survivor inquiries], we have to review every document and every page and every word."

1.3 Reviewing on access

Reviewing on access, rather than intake, was also identified as a strategy. Participants using this strategy review digital content for sensitive information only when a request comes in to access it. This practice may be combined with blanket restrictions or policy-based access (i.e., use of researcher agreements or varying the review level depending on who is making the request to access). As one participant described, "We just say

‘pending review’ and then . . . once somebody requests access to it, an archivist will do a – we call it informal access review.”

Another participant described this access review method as working on a sliding scale based on risk profiles: “before we would give the data to a client, it is assessed . . . and the level, the sliding scale of sort, of complexity and effort that’s invested in that is dependent on the material you’re dealing with.”

The strategy to review on access and not intake was a common one, but it was almost always accompanied by an acknowledgment that it was necessary because of the scale of the materials. This strategy may also affect discoverability and access, as it may keep materials restricted.

1.4 Using software for triage

Participants also reported using software tools for triage. Participants using this strategy use software to support an initial assessment but follow up with human labour and judgement. One method was the use of software to remove irrelevant files. This immediately reduces the quantity of objects to review and visual or mental clutter. Here, one participant describes that process:

That particular case had millions and millions of files and so for me to even start the archival appraisal I needed to really get a grip on duplicates and isolating software, just kind of objects that were in the disc images that were clutter, that made it difficult to do that archival appraisal. And so that was helpful, I went from millions and millions to, I don’t know, something in the order of 100,000 files.

Granular string searching is another example of software tools supporting practice. String searching refers to searching a digital object for terms or patterns, like looking for the expected pattern of a social insurance number. Among those using proprietary software, Forensic Toolkit (FTK) was a common choice for identifying this type of information. One participant also specifically called out Quick View Plus as a proprietary tool that assists with manual review of a wide range of file formats. Among participants developing workflows using open-source tools, BitCurator and bulk_extractor were commonly mentioned by name. Almost all participants do or have done some type of string searching in their review processes. However, some found the approach too aggressive, leaving them wishing for options including confidence ratings, the ability to ignore entities, and the ability to set granularity levels.

Category 2 – Challenges

Many of the challenges covered in our literature review were shared by participants. However, most were not Canadian examples. We recognize that we learn from our colleagues in other countries, but there are differences that impact outcome. Our laws differ. Our relationships with and obligations to Indigenous Peoples differ. Our funding structures differ. Our multilingual identity differs. As a result, our challenges and needs may differ. Understanding and documenting those challenges was our goal here.

2.1 Reviewer trauma and audio-visual collections

Reviewing audio-visual collections was found to be particularly difficult and emotionally taxing due to the labour required and the sometimes personal and heavy nature of the material. Eight of the 11 participants identified audio-visual as being difficult to review. Two also noted how challenging this material can be for reviewers because of how personal it is, with one saying, “audio-visual material can be very heavy to process, or to look at,” and the other explaining that

the amount of vicarious trauma on other staff members is fairly high, [and] it’s actually the review of the audio-visual testimonies, when people have to sit down and watch them from beginning to end and listen to every word . . . it can be taxing on people.

Another noted that their process encourages reviewers to take breaks, and they sometimes flag potentially sensitive content for them before their review. One participant supported reviewers by providing access to elders and/or mental health professionals.

The review of audio-visual materials also tends to require more staff time and labour. String-searching is not possible without transcripts and does not cover images. Reviewers must view or listen to the content in its entirety. One participant considered using image analysis machine learning tools to identify nudity in image collections as a way of triaging content but did not have the staff time to pursue the idea and had concerns about inadvertently censoring valid forms of self-expression.

2.2 Higher risk of putting materials online

Participants expressed concerns about the higher risk of harm when putting materials online, including legal consequences and the lack of infrastructure for secure access. There are two consequences to this, the first being that access tends to be online or not at all. As one participant noted when it comes to access, “the window’s open or the window’s shut.” Another lamented,

our model is sort of like access [means] online dissemination, but does it have to be? Are there other ways around that? . . . even the most banal information now can be aggregated, analyzed, put into different contexts. That makes it problematic . . . But we haven’t figured that out.

Most memory institutions in Canada do not have the infrastructure, such as secure virtual reading rooms or repositories with escalating permissions, or staff capacity to redact digital materials at a granular level. As a result, the options left are often to provide online access, restrict the content completely, or not collect at all.

A second consequence of this challenge is the effect funding for online access can have on what is collected. Digitization or digital collection grants are typically only available in Canada for public and online materials. This means valuable content may not be selected or prioritized for attention if it contains sensitive information. As one participant noted, “you end up cherry picking or avoiding more complex content.” This is because,

as another participant suggested, “our current funding programs don’t build in that nuance that’s needed to make these collections more accessible.”

Grant funding models tend to support time-limited projects, not infrastructure. Funding is provided for digitization, but not for a full review for sensitive content or the building and maintenance of a secure virtual access system. The result may be a version of our history that continues to leave out voices or records deemed unsuitable.

2.3 Third party information

The presence of third-party information in digital collections raised privacy concerns and the potential for harm, especially when such information is easily shared and indexed. Third party information refers to information about a person who is not the principal subject of a record. For example, in a collection of letters, the principal subject may be the donor of the letters, but a third party would be someone discussed within those letters, or the original sender. The presence of third-party information is a risk that exists in any format, but the amount or type of harm may be amplified with digital collections. For example, digital formats allow for easier searching and sharing: “In isolation they’re not that sensitive, we don’t really care. But it’s the aggregate, we don’t want people disseminating that aggregate as an aggregate.”

We also saw concerns about outing people—that is, disclosing information about their intimate or professional lives that they may not have chosen to disclose widely or no longer wish to disclose—and similar concerns about calling someone by a prior name that they no longer wish to be called. This can be harmful when a deadname (former name) does not affirm someone’s current identity or puts them at risk by disclosing a transition without their consent. Releasing content that includes third parties carries a higher risk of harm when that content is easily shared and indexed digitally. For example, the University of Victoria is home to the Transgender Archives, which acquires documents, publications, personal records, and memorabilia of persons and organizations associated with activism by and for trans people. To address third party challenges, they rely on access restrictions and researcher agreements. They also work with donors to identify potentially sensitive information or include restrictions and lean heavily on their own professional judgement and human review. In our interview, they raised concerns about the digitization of small, local publications or newsletters. These are important research resources that tell the story of local trans communities and activism, but the archives also recognize people (third parties) may have contributed to these publications not thinking of the future of broad digital access. That is, someone may have intended to write a letter to a small, local newsletter with a circulation of a few hundred, not for it be published online and indexed. The ArQuives takes a similar approach, relying on staff expertise, donor consultation, and human review to balance access with care. The ArQuives collections also include materials from frontline youth organizations which undergo closer review due to third party concerns.

2.4 Fear of censorship

Fear of censorship was another challenge, as participants grappled with how to review without excessive restriction. Care does not always mean restriction or redaction. As one participant asked, “How do we do it [review potentially sensitive content] in a way that protects our donors and gives them the right to that expression without it being something that is kind of policed?” They were referring to the inclusion of gender or sexual expressions of donors with their consent. How do you review, whether manually or with the use of software, in a way that respects the content of that expression and does not restrict too aggressively? This is likely not a challenge software can address.

2.5 Friction between theory and practice

Friction between archival theory and digital practices was also observed, with the formats of digital collections often necessitating a file-by-file review rather than a collection-level approach. Due to the scale of digital collections, archival theory often encourages or requires working in the aggregate (Greene & Meissner, 2005). However, the tools for working with digital content may force users into a file-by-file review. This tends to result in increased focus on individual objects rather than taking a preferred collection or fonds level approach. For some participants, such as the NCTR and the SRSC, file or item-level review and description is necessary regardless of the format. For others, it may be a burden. When describing using FTK computer forensics software and Archivematica, an open-source digital preservation system, one participant said, “You're forced into a very granular perspective when . . . maybe that's not normal for print material.” In this case, they were unable to translate their traditional, paper-based method to digital because of the file-by-file approach these tools take and their inability to adjust the granularity depending on context.

2.6 IT or infrastructure challenges

IT and infrastructure challenges emerged as participants highlighted gaps in resources, including limited IT support, lack of funding for storage and access systems, and restrictions on the use of open-source software. Four participants described a gap existing between the resources needed and the IT resources available to them. For example, IT departments may have different expectations regarding the meaning of “long-term” storage: “IT don't really have their head around [the concept that] . . . we're going to need it forever and it's going to [keep] increasing.”

Participants also mentioned having policies that require the use of locally managed storage for content suspected to contain sensitive information. This can complicate the use of hosted digital preservation services and storage and means that for several institutions, digital files can end up sitting in network storage rather than being put through their full digital preservation workflows. One participant specifically mentioned institutional IT policies that prevent the use of open-source software, limiting the tools available for detecting and managing PII and sensitive information and hindering their ability to develop staff expertise:

I have a vested interest in hiring highly skilled staff, building expertise, being able to use open-source products to do some development toward our own unique specifications and requirements, to be able to contribute that back to the international community. And to be able to sort of evolve based on your ability to sustain your own operations. That's not really possible at the moment and that's what I would change—to engage more deeply [with] more open-source-oriented tools and development.

2.7 Resources

Participants consistently expressed the need for more resources. We asked each participant, “If you had a magic wand and an unlimited budget, what changes would you make to help you do this work?” Overwhelmingly, the responses included more people, more time, and/or more training. Every participant described current levels of funding or staffing as a challenge. As one participant said, “the challenges come like 20 steps before, where we're trying to convince people that this is going to cost money if they want this to be a part of what we do.”

Several participants also mentioned being the only or one of only a few staff members at their institution assigned to digital curation, leading one to comment they “feel alone a lot.” Low staffing levels contribute to keeping complex collections in backlogs rather than making them accessible for research: “There's just not enough time for me to do all of the work in a kind of real, meaningful way.”

2.8 Unique challenges of correspondence

A common theme was concern for how to handle private and sensitive information in correspondence such as text messages and emails. Among the challenges noted by participants are the scale of the messages, the ubiquity of private and sensitive information, and the difficulty of separating personal or private correspondence mixed in through an inbox. As one participant said,

Even if we take a writer or an artist or a trans activist, their phone collections will be filled with correspondence . . . authored by other people, their incoming correspondence, it might be full of hundreds of different correspondents . . . and that's a real challenge.

Despite its challenges, email remains important as a modern form of correspondence. One participant whose organization is currently unable to acquire email described the impact of this, explaining that “we're only getting pieces of the total fonds.”

Participants rightly noted that this is a space where software tools have been developed in recent years to assist in review. Several participants mentioned experimenting with ePADD specifically for review of email. However, issues remain:

But it just struck me even with like a moderate or even small sized email archive, it's just not worth it to try to make those decisions item by item. So, it's almost in those cases, we're just saying the entire email archive is restricted, [and] access under a research

agreement comes in, right? In this case [the researchers] would have to come in, because we haven't really got a way to provide offline delivery.

2.9 Lacking trust and confidence in software tools

Lastly, participants expressed concerns about the trust and accuracy of software tools for detecting sensitive information, leading to cautious reliance on these tools and a preference for human review. Multiple participants mentioned having low confidence in the results of software tools designed to detect private and sensitive information. This is particularly true of software tools which rely on string searching or regular expressions to identify such content. These tools were seen as aggressive, or as one participant said, "very, very, very overzealous," with another adding, "I'm not sure how accurate it could possibly be when it comes to really important identifying information."

Lexicon-based NLP or Artificial Intelligence (AI) approaches can share the same issue:

Just a few times we've used ePadd and taken a look at some of those lexicons that are supposed to flag sensitive information—they're almost so way over the top . . . you've got a body of email and it says there's 200 records [relating to] heavy drug use. But when you go, it's just because it's flagged words crack or dope . . . which in another context have nothing to do with drugs . . . I just don't have that much confidence in that sort of word searching for the qualitative data.

Given the risks of releasing sensitive information, such software tools may be more appropriate for initial triage than as a solution in and of themselves: "We've looked into some different solutions, but we've never pursued it fully just because of the risk level of this type of information being missed by some type of algorithm and then being released. The risk is too great."

There are also barriers to entry for such tools, including a lack of commercially available AI-based sensitivity review software tailored for cultural institutions and the labour involved in training these models and evaluating their results. One participant commented:

Are we going to be able to build models that help other institutions or particular types of formats like, you know, university records or HR and so on? . . . I fear some of it's going to be fairly institutional context-based and we'd need to kind of have the tools to develop our own processes at different institutions.

Another remarked on the constraints of programs that aren't intuitive to use and require technical expertise:

We shouldn't have to learn machine learning techniques to be able to do this . . . it needs to be in a user-friendly way so you're not relying on the techie to be doing all the sensitivity reviews or all the processing.

Discussion

In our findings, we discovered that what is perhaps most needed is time and support for Canadian memory workers. Throughout this research, we also learned that many of the strategies and challenges were not new; they were simply being translated to the digital realm. Tactics for reviewing sensitive information still heavily rely on human knowledge, judgment, and relationships, which in turn rely on informed and engaged cultural memory workers actively working with records producers and subjects. These people then rely on support, time, and training for when they must engage with traumatic and difficult materials. While software tools can assist, the stakes of certain collections may be too high and the nuances too subtle to rely solely on these tools without human oversight. Having more people and more time for training would also allow for experimentation. New technology is of little use if you don't have the time to try it and learn how it may or may not be helpful.

These findings may also be of interest to developers or designers who work on discovery tools. Based on our findings, we suggest being upfront about limitations, defining capabilities and jargon at the point of use, and striving for easy installation and graphical user interfaces when possible. If appropriate review requires expert knowledge of the content and its community, it should be easily performed by individuals with that expertise. As one participant suggested, this work should be done by those who work with collections, not systems. The time required to learn tools and workflows and then review content should not be overwhelming and is a labour issue, as mentioned by Lassere (2021). Review tools should also be designed to be assistive rather than authoritative, acknowledging a reviewer's judgment while allowing flexibility in their decisions.

Lastly, changes to grant funding stipulations that require public access may be needed most. This recommendation is based on the findings regarding the higher risk of putting materials online and the difficulties around third-party information. Access platforms that allow for granular controls and digitization funding models that enable more nuanced forms of access (e.g., offline access or limited access) should be considered. Funding agencies could enable access to more diverse, representative, and underserved collections by either funding digitization work without requiring that all materials are made available online and by allocating funding for memory workers to spend time carefully reviewing collections that may contain sensitive and private information during digitization projects. An example of this type of work is the Indigitization collaborative initiative (indigitization.ca), which enables community-led digitization projects through grant funding and training without requiring control of or access to the digitized content. Further investment in controlled access repositories would also enable responsible online access to complex digital collections. For example, the Dataverse repository (known as Borealis in Canada) is a research data management platform that permits role-based access controls at the file, dataset, collection, or Dataverse levels. While this system is designed specifically for research data, the concept of a national deposit and access platform allowing for depositor-managed access controls is appealing based on the challenges expressed by participants. This would allow for greater access to more complex or sensitive collections rather than keeping them hidden.

Limitations of This Study

We are two insider researchers who are or were digital preservationists at two different Canadian academic libraries. Participants were known to us as peers. There are challenges associated with insider research. According to Coghlan and Brannick (2005), an insider researcher may be perceived as being too close to the data, potentially making assumptions during the interview process and lacking the objective distance to challenge responses. Additionally, Lofland and Lofland (2006) warn that being too close may prevent researchers from seeing novelties. However, there is also value in having a rich understanding of a topic and an existing relationship with participants (Chavez, 2008; Taylor, 2011). In their work, *Gaining Access: A Practical and Theoretical Guide for Qualitative Researchers*, Feldman et al. (2003) argue that access is enriched over time, and being an insider has advantages such as trust, access, and awareness. Stanley and Wise (1993) similarly encourage researchers to embrace their insider perspective and trust their emotional responses as valid research tools. An insider perspective and a deeper understanding of an experience allows for empathy and response, which points the researcher to topics of interest an outsider may not identify (Stanley & Wise, 1993). In her article on researching within pre-existing friendships, “The Intimate Insider: Negotiating the Ethics of Friendship When Doing Insider Research,” Jodie Taylor (2011) reflects on the stresses inherent to researching peers, such as the fears of getting it wrong or betraying trust, and provides mitigating strategies such as seeking participants’ validation of the researcher’s observations and interpretations. Taylor served as inspiration for our decision to share live notes during the interviews and ask participants to review and edit their transcripts and this paper before submission.

We acknowledge this familiarity and have tried, throughout this process, to examine and reflect on it while also putting it to good use.

Tessa Walsh, one of the authors of this paper, is also the creator and developer of Bulk Reviewer. Bulk Reviewer is a desktop application that aids in identification, review, and removal of sensitive files in directories and disk images. It’s likely most participants have either used Bulk Reviewer or are familiar with it. We didn’t mention or ask about Bulk Reviewer specifically, but we were concerned some participants might hold back on criticism of available tools in the field because of this connection. This did not end up being a concern, however, as only one participant mentioned Bulk Reviewer and did not withhold criticism about its limitations.

At the time of the interviews, the same author was also employed as a software developer by Artefactual Systems, a company that develops and maintains the open-source Archivematica and Access to Memory (AtoM) projects. These software projects are widely used by archivists, librarians, and other memory workers in Canada, including many of the participants in this study, but are not focused specifically on the issue of private and sensitive information.

Conclusion

Through this work, we explored how Canadian librarians and archivists currently review digital materials for sensitive personal information. We discovered a range of strategies already in place that heavily rely on consultation, contextual understanding of the content, risk profiling, and the translation of existing approaches to digital materials. Challenges identified include reviewer trauma, risks of online dissemination, third-party information, fear of censorship, friction between theory and practice, IT and infrastructure challenges, limited resources, unique correspondence challenges, and concerns about the accuracy and trust of software tools. Addressing these challenges and supporting existing effective strategies will contribute to the responsible and ethical management of digital collections, ensuring access while safeguarding sensitive and private information.

Acknowledgements

We would like to thank our gracious and generous participants:

Creighton Barrett, Dalhousie University

Jesse Boiteau, University of Manitoba and the National Centre for Truth and Reconciliation

Richard Dancy, Simon Fraser University

Glenn Dingwall, City of Vancouver Archives

Alex Garnett, Simon Fraser University

Tim Hutchinson, University of Saskatchewan

Krista McCracken, Algoma University and the Shingwauk Residential School Centre

John Richan, Concordia University

Tom Smyth, Library and Archives Canada

Anna St. Onge, York University

Raegan Swanson, The ArQuives

Lara Wilson, University of Victoria

We would also like to thank the Canadian Association of Research Libraries (CARL-ABRC) for their support of this project through the Research in Librarianship Grant.

References

- Artefactual. (2020). [Make rights actionable on digital objects](#). *AtoM documentation* (Version 2.6).
- Arroyo-Ramírez, E., Bolding, K., Butler, D., Coburn, A., Dietz, B., Farrell, J., Helms, A., Henke, K., Macquarie, C., Peltzman, S., Tyndall Watson, C., Taylor, A., Venlet, J., & Walker, P. (2020, February). [Levels of born-digital access](#). Digital Library Federation.
- Baker, D., Butler, D., Clemens, A., Velazquez Fiddler, C., Gentry, S., Riddlesperger, L., Wachtel, J., Williams, C., & Wisner, J. (2023, December 26). [An exploration of access systems: A framework for access systems for born-digital archival materials](#). The Digital Library Federation Born-Digital Access Working Group.
- Barrett, C. (2017, May 16). [Digital forensics tools and methodologies in archival repositories \[research seminar presentation\]](#). Faculty of Computer Science, Dalhousie University.
- Caswell, M., & Cifor, M. (2016). [From human rights to feminist ethics: Radical empathy in the archives](#). *Archivaria*, 81, 23-43.
- Chavez, C. (2008). [Conceptualizing from the inside: Advantages, complications, and demands on insider positionality](#). *The Qualitative Report*, 13(3), 474-494.
- Cifor, M. (2016). [Affecting relations: Introducing affect theory to archival discourse](#). *Archival Science*, 16(1): 7-31.
- Coghlan, D., & Brannick, T. (2005). *Doing action research in your own organization* (2nd ed.). SAGE.
- Dataverse Project. (2022, June 13). [Dataset + file management](#). *User guide*. Dataverse.org.
- Emerson, R., Fretz, R., & Shaw, L. (2011). *Writing ethnographic fieldnotes* (2nd ed.). University of Chicago Press.
- Faulder, E., Annand, S., DeBauche, S., Gengerbach, M., Irwin, K., Musson, J., Peltzman, S., Tasker, K., Jackson, L. U., & Waugh, D. (2018). [Digital processing framework](#). Cornell's Digital Repository.
- Feldman, M., Bell, J. & Berger, M. (2003). *Gaining access: A practical and theoretical guide for qualitative researchers*. AltaMira Press.
- Garcia-Spitz, C., & Geraci, N. (2021). [Negotiating online access: Perspectives on ethical issues in digital collections](#). In E. Arroyo-Ramírez, J. Jones, S. O'Neill, & H. Smith (Eds.), *Radical empathy in archival practice* [special issue], *Journal of Critical Library and Information Studies*, 3(2), 1-20.

- 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.
- Greene, M.A., & Meissner, D. (2005). [More product, less process: Revamping traditional archival processing](#). *American Archivist*, 68(2), 208-263.
- Harris, V. (2014). [Antonyms of our remembering](#). *Archival Science*, 14(3-4), 215-229.
- Hurley, G., & Shearer, K. (2019, November 29). [Final report of the survey on digital preservation capacity and needs at Canadian memory institutions, 2017-18](#). Canadian Association of Research Libraries.
- Hutchinson, T. (2017, December 14). [Protecting privacy in the archives: Preliminary explorations of topic modeling for born-digital collections](#) [Paper presentation]. *Proceedings of the 2017 IEEE International Conference on Big Data*, 2251-2255.
- Hutchinson, T. (2018, December 10-13). [Protecting privacy in the archives: Supervised machine learning and born-digital records](#) [Paper presentation]. *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, 2696-2701.
- Hutchinson, T. (2020). [Natural language processing and machine learning as practical toolsets for archival processing](#). *Records Management Journal*, 30(2), 155–174.
- Kaczmarek, J., & West, B. (2018). [Email preservation at scale: Preliminary findings supporting the use of predictive coding](#) [Paper presentation]. *iPres 2018 Proceedings*.
- Kim, E. T. (2019, January 30). [The Passamaquoddy reclaim their culture through digital repatriation](#). *The New Yorker*.
- Lassere, M. (2021). [Translating bits: Maintaining \(born-\)digital heritage](#). *Brodsky Series for Advancement of Library Conservation*, Lecture 15. Syracuse University.
- Lassere, M. (2020, May 18). [The archeology of maintenance: The role of information maintenance in sustaining digital archives](#) [Plenary session]. *Best Practices Exchange*.
- Lassere, M., & Whyte, J. (2021). [Balancing care and authenticity in digital collections: A radical empathy approach to working with disk images](#). *Journal of Critical Library and Information Studies*, 3(2), 1-25.
- Lee, C. A., & Woods, K. (2012, September 26-28). [Automated redaction of private and personal data in collections](#). *Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation International Conf*, 298-313.
- Lofland, L., & Lofland, J. (2006). *Analyzing social settings*. Wadsworth Publishing Company.

- McCracken, K., & Hogan, S.-S. (2021). [Residential school community archives: Spaces of trauma and community healing](#). *Journal of Critical Library and Information Studies*, 3(2), 1-17.
- Moran, R. (2016). [Indigenous people should decide on matters of access to archival information](#). *International Journal of Circumpolar Health*, 75.
- National Centre for Truth and Reconciliation. (2021, Mar 29). [NCTR launches a new website and archive database](#).
- Office of the Privacy Commissioner of Canada (2023). [Summary of privacy laws in Canada](#).
- Post, C., Chassanoff, A., Lee, C., Rabkin, A., Zhang, Y., Skinner, K., & Meister, S. (2019, June 2-6). [Digital curation at work: modeling workflows for digital archival materials](#) [Paper presentation]. *Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries*, 39-48.
- Ryan, G. W., & Bernard, H. R. (2003). [Techniques to identify themes](#). *Field Methods*, 15(1), 85-109.
- Schneider, J., Chan, C., Edwards, G., & Hangal, S. (2017). [ePADD: Computational analysis software facilitating screening, browsing, and access for historically and culturally valuable email collections](#). *D-Lib Magazine*, 23(5/6).
- Sloyan, V. (2016). [Born-digital archives at the Wellcome Library: Appraisal and sensitivity review of two hard drives](#). *Archives and Records*, 37(1), 20-36.
- Stanley, L., & Wise, S. (1993). *Breaking out again: Feminist ontology and epistemology* (2nd rev. ed.). Routledge.
- Taylor, J. (2011). [The intimate insider: Negotiating the ethics of friendship when doing insider research](#). *Qualitative Research*, 11(1), 3-22.

Appendix A: Semi-Structured Interview Script

Part 1 - Introductory Probes

1. Tell us about your institution - its size, maybe some of the things it's known for...
2. And what's your role there?
 - a. Have you been doing that long?
 - b. And are you on a team or are there other people that help you with this?
3. What types of digital content are you typically working with?
 - c. What domain? (e.g. research data, special collections, artworks)
 - d. How consistent is this content, e.g. in terms of its source or format?
4. What types of private or sensitive information are you most concerned about?
 - e. Has this come up before?
5. Would you say it's common? /OR/ and so, what types of private or sensitive information are most common in your collections?
6. Do your donor agreements, collection development policies, or other policies address the issue of private or sensitive information in your collections?

Part 2 - Workflow Specifics

7. How do you know it's there? As in, how do you identify or find <provided example of sensitive information>?

Note: Questions below may change depending on institutional context and response.

8. And who is responsible for that? <probe to find out who conducts the review process? Is it multiple people?>
9. When does the review process happen?
 - f. Would you say it's ongoing? Or more of a formal one-time step in your overall procedure?
 - g. Follow-up if answer is one-time step: At what stage in your management workflow does this step happen? Prior to materials being ingested? Prior to providing access?

10. Do you find sensitive or private information tend to be concentrated in particular document types/file formats/collections?
11. Are you using any software tools to help with that process? Which ones?
12. Have you experimented with other tools?
13. What steps are involved in your review process?
 - h. Do you mind if we sketch this out a bit while we talk? [dependent on context]
 - i. Does this change depending on the content/collection/donor?
14. And how long does that review typically take?
15. Would you say your digital review process is similar to your paper-based review process?
 - j. Follow-up re: if differences, how so?
16. Have you identified gaps in your current workflows and tooling? If you had a magic wand and unlimited budget, what changes would you make to help you do this work? E.g. software tools to do _____, staff, etc.

Part 3 - Wrap Up

17. Are there dimensions of the review process you'd like to mention that we haven't already discussed?