Security Without Obscurity: Managing Personally Identifiable Information (PII) in Born-Digital Archives

Ben Goldman and Timothy D. Pyatt, Penn State University

Abstract

This article examines current archival thinking and practice surrounding the identification and management of personally identifiable information (PII) found in born-digital collections in an academic repository context. Data from recent surveys and reports, along with examples of how several different repositories have attempted to balance management of born-digital records that require access control, inform the recommendations of the authors.

1. Introduction

In an era of online banking, social media, and identity theft, concerns about managing personally-identifiable information (PII) such as credit card information and social security numbers (in the United States) are ever present in the media and public forums. These issues are nothing new for archivists, who have been managing sensitive data since archival repositories emerged. Protocols for restricting access to certain files, redaction of documents, and staff-only access to stacks where the collections are stored are all standard procedures. Some repositories may require researchers to read and sign special forms so that they understand and are partners in managing the risk of exposing such data. So what makes securing sensitive information in electronic format different? It might be tempting to focus on the material differences -- the wide range of fixed media and storage methodology used for electronic records -- and assume that the
same access and restrictions protocols should work for most digital archives. Until an archive measures its management, storage, and access protocols against institutional and industry PII protection standards, the gaps may not be apparent.

That was the case for the Duke University Archives staff when the data storage servers they used for born-digital archival records underwent a campus information technology security audit in 2010. Information technology audits are much like a financial audit. Information technology security experts review the processes and procedures used to manage, protect, and document the materials under review. However in this case the assets were born-digital records rather than financial transactions. After conducting their review, the information technology auditors were surprised by the volume of potentially sensitive digital materials that the Duke Archives staff had collected and stored on its server. The auditors’ primary concern was for the security of the personally identifiable data that could be part of the files. The Archives staff explained to the auditors that archivists have experience managing PII and that existing archival holdings also contained boxes filled with sensitive data. The auditor replied that paper-based archival records relied on “security through obscurity;”¹ a thief seeking personal data would have to manually sift through the files page by page to find that data, while digital PII data can be compromised using computer technologies such as scripts that look for credit card sequences and similar sensitive content.² Servers with PII are also generally subject to a higher level of review and security under most institutional policies, as they tend to be targets for hackers looking for personal data to exploit. Until this review, the libraries and archives were generally considered outside of such concerns.

The IT audit experience of the Duke University Archives staff served as a catalyst to elevate the security of their born-digital records and their awareness of the need to screen such
records for PII. For the larger archives and library community, this case study demonstrates that we must now develop strategies that not only preserve and manage the born-digital records and personal papers we collect, but also can support the protection of PII without placing an undue burden on already-limited institutional resources. We need to work with our colleagues in information technology and better understand the institutional compliance issues and threats that impact our work. This article will examine what content archivists are required to protect under most laws and regulations as well as look at different options that can be employed to store, preserve, identify, and protect these data. Much as we have done with analog records, we also must use these data to help determine what is acceptable institutional risk for collecting, storing, and making available for research born-digital records. What should become apparent after considering this topic is that the similarities between analog and digital issues are more striking than the differences. The challenges presented by emerging technologies will require archivists to develop new workflows and refine existing policies, but past strategies for providing access to archival collections while limiting exposure of sensitive data remain relevant when dealing with born-digital records.

2. Traditional Archival Approaches to Privacy

When considering how best to manage private and sensitive records, most archivists in the United States look to the guidance offered by the Society of American Archivists (SAA). The Society’s “Core Values of Archivists” states:

Archivists promote and provide the widest possible accessibility of materials, consistent with any mandatory access restrictions, such as public statute, donor contract, business/institutional privacy, or personal privacy.
The SAA “Code of Ethics for Archivists” offers further guidance:

Archivists recognize that privacy is sanctioned by law. They establish procedures and policies to protect the interests of the donors, individuals, groups, and institutions whose public and private lives and activities are recorded in their holdings. As appropriate, archivists place access restrictions on collections to ensure that privacy and confidentiality are maintained, particularly for individuals and groups who have no voice or role in collections’ creation, retention, or public use. Archivists promote the respectful use of culturally sensitive materials in their care by encouraging researchers to consult with communities of origin, recognizing that privacy has both legal and cultural dimensions. Archivists respect all users’ rights to privacy by maintaining the confidentiality of their research and protecting any personal information collected about the users in accordance with their institutions’ policies.3

The guidance offered by SAA lines up with that offered by other professional organizations. The Association of Canadian Archivists states that “Archivists encourage and promote the greatest possible use of the records in their care, giving due attention to personal privacy and confidentiality…”4 The Institute of Certified Records Managers state that “They should protect confidential, proprietary and trade secret information obtained from other and use it only for the purposes approved by the party from whom it was obtained.”5 Only the International Council on Archives directly addresses electronic records in their 1996 statement: “Archivists should take care that corporate and personal privacy as well as national security are protected without destroying information, especially in the case of electronic records where updating and erasure are common practice.”6 Therein lies the fundamental tension for archivists -- collecting records and facilitating access while striving to protect private and sensitive information that may be part
of the records acquired. Much is left for the archivist to determine and measure against applicable laws, regulations, institutional policy, and donor wishes.

Depending on the collecting mission of an archival repository, there are numerous types of sensitive data a repository might receive that may require extra security as proscribed by state or U.S. federal laws, as well as local policies. OCLC defines sensitive data as “confidential information such as social security numbers, financial information, research data with names of participants, or other confidential information that must be restricted or redacted.” Lee and Woods define private data in the digital context as “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).” Donors and records creators often knowingly or inadvertently transfer such data to an archival repository.

Legislation protecting personal data and privacy (much like the separate but almost mirrored definitions above) makes little distinction between analog and digital data. The most relevant laws in the United States that impact archives are:

- Family Educational Rights and Privacy Act (FERPA): relevant for most academic archives that may have some form of student records.
- Health Information Portability and Accountability Act (HIPAA): the privacy rule of HIPAA applies to archives that are part of covered entities (i.e., an institution that provides health care services in some form and maintains records of patient treatment).
- Public Records Laws (sometimes known as Freedom of Information Acts or FOIA): Depending on the source of funding and the nature of the records collected, an archival institution may or not be subject to a state’s open records laws or the federal FOIA.
Other relevant laws impacting archives may include those pertaining to the protection of credit card information and social security numbers. The work, *Navigating Legal Issues in Archives*, gives an extensive breakdown and explanation of laws impacting privacy in its chapter on “Access and Privacy Issues in Archives.”

While it is important for the archivist to know whether the records they are acquiring contain these types of sensitive data and the institutional context and requirements that protect it, these elements are not always in our control. Archivists often find it impractical to identify the existence of PII as collections have grown in size and resources have diminished. These realities have led to the development of certain strategies for dealing with PII, which are explored in the next section.

a. Institutional Policies

Archivists, whether as an institutional records manager, or as a collector for a manuscripts repository, are used to reviewing and compiling institutional policies for the managing of records that are subject to FERPA, HIPAA, and similar policies. For archivists working within an academic context, there may be relevant university-level resources, such as compliance tutorials, but institutional policies on information security may not be as visible. Most academic institutions also have policies and websites devoted to this subject as well information security officers and or related staff that archivists have not previously worked with as they protected analog personal data. As archivists develop infrastructure for acquiring and preserving born digital records, not just institutional records, but also in manuscript collections, these staff can be great resources as well as advocates for the secure storage needed, as well as the new policies and procedures needed to ensure the data is secure over time.
b. Working with Donors and Records Creators

Archivists both have significant experience and a wide body of literature that discusses working with donors/records creators in transferring their records to an archival repository. Best practice includes making sure the donor has clear ownership of the materials and completing a donor agreement or deed of gift, which typically identifies, in advance, potential restrictions on access to content in the collection. At the very least, some level of pre-acquisition appraisal has been recommended widely in the profession. Absent this level of pre-transfer engagement with collections and donors, archivists may be left with unknown quantities of sensitive data in their collections. Though many collections may ultimately prove to be free of PII, the worst-case scenarios -- large, messy collections that are likely to have PII present -- have been a source of constant worry in professional practice.

c. Processing Challenges

These worst-case scenarios have also been a source of tension in the professional literature, especially in the years since Mark Greene and Dennis Meissner articulated the value of minimal processing strategies. “More Product, Less Process” (MPLP) is the phrase used to describe arrangement and description philosophies that seek to pragmatically apply institutional processing resources in order to support broader collections access. Greene and Meissner recommended defining a baseline level of arrangement and description, deviating from the baseline as needed, based on the specific needs of each collection. In practice, this often results in an archivist processing a collection to a file or box level, since MPLP typically abstains from the kind of item-level appraisal that would facilitate comprehensive identification of sensitive
data in complex collections.

d. Working with Researchers

Addressing the argument that minimal processing potentially makes sensitive data accessible to researchers, Greene and Meissner point out there is no such thing as zero-risk when it comes to PII. Instead, they recommend mitigating risk through the practices mentioned above -- creating institutional policies, engaging donors, undertaking appraisal -- and also through working more directly to educate researchers using collections where PII may exist but remains unidentified. At a minimum, this might require reference archivists to talk to researchers about PII and instructing them to inform the archive should they find such material. Some repositories might consider taking this notion a step further and included language in researcher registration forms that define the researcher’s responsibility in protecting PII.\(^{11}\)

3. Challenges Presented by Born-Digital Archives

As already noted, archivists have begun in recent years to intentionally acquire large sets of born-digital material and also examine the legacy holdings acquired in years past. Often these legacy holdings include unexamined fixed media (for instance, floppy, zip, and compact disks) in collection boxes. In some notable cases, new acquisitions of born-digital material are comprised of whole computers or hard drives. Archivists have started wrestling with challenges such as how to mitigate the risks imposed by obsolete media, how best to authentically store and preserve digital objects and environments, and how to effectively and efficiently acquire large sets of digital material from donors. The growing list of recommendations found in case studies, such as those published by the Society of American Archivists, and in reports such as those
issued by OCLC Research, begins to outline some of the PII challenges faced by archivists managing born-digital materials.

Until recently, much of the profession has remained unprepared to administer digital material in a way that supports identifying sensitive information where it exists in collections. The disks incidentally acquired as part of larger, paper-based collections now constitute what has come to be called a fixed media backlog. The 2010 Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives\(^{12}\) demonstrated that born-digital materials were widely unmanaged and by extension, unappraised; it is not a stretch to assume this might partly be explained by the existence of orphaned fixed media in collections. The scale of this backlog is unclear, though assessment performed by twenty-three institutions participating in the Society of American Archivists Manuscript Repositories Section’s “Jump In Initiative” identified thousands of disks holding data that scaled well into terabytes.\(^{13}\) Unless we take action to better appraise the data found on fixed media, we will be unable to take the appropriate steps to preserve the content as well as remain unaware of potential PII risks.

It perhaps goes without saying, but the nature of digitally-produced content itself presents some immediate obstacles to identifying and mitigating the risks of PII. The computers and storage devices acquired by archives often “obsolesce at several levels” including “file format, file system, operating system, application, and hardware.”\(^{14}\) A further exploration of just one of these levels -- hardware -- helps explain the challenges an archivist working with born-digital material faces. Two once-common forms of storage media, 5.25 inch and 3.5 inch floppy disks, can no longer be read using modern desktop computers because the drives and controllers required to mount these disks have been discontinued by manufacturers. Despite the wide availability of CD and DVD drives in office computers today, cloud storage is already shifting
the landscape of storage option preferences for consumers. When Apple announced a new line of
Macs that did not include optical media drives in the fall of 2012, some analysts pronounced a
death sentence on CDs and DVDs.¹⁵ In order to access the digital files on such obsolete storage
mediums, archivists are modestly collecting historic computing devices. To address the other
levels of obsolescence, some archivists are maintaining “antique” software applications,
exploring tools that migrate files from older environments to modern ones, and even emulating
applications and operating systems to render files as they originally appeared at the time of their
creation.

Best practices for dealing with these issues of obsolescence -- indeed, for managing and
preserving born-digital archives broadly -- are coalescing around a process referred to as disk
imaging. Disk imaging was born from digital forensics, best known for its use in criminal
investigations and the recovery of digital evidence. It has emerged as a method for transmitting
digital files from fixed obsolete media to more sustainable storage environments. Such strategies
lead to the creation of bit-for-bit copies of disks and whole computers alike, in the form of a
single file called the disk image. Forensic approaches also help facilitate, among other things,
capturing and transferring digital material from archival donors, ensuring and documenting the
authenticity and provenance of digital objects, and extracting metadata about digital objects.

Digital forensics tools also provide the ability to automate the identification and
remediation of private or sensitive information found in digital files, even where it is not obvious
to the naked eye. While forensically imaged computers and removable media may contain
obvious PII in the files and folders, such media might also contain hidden deleted files that
persist and are recoverable, unbeknownst to their creators. Archivists and donors alike need to
realize that forensic tools are capable of surfacing and preserving such files. Likewise, creators
may be fully aware of the PII to be found in a particular computer directory, or in an email account, but may be less aware of the kind of identifying information that can be gleaned from metadata associated with particular types of digital information, such as global positioning system (GPS) data found in the exchangeable image file format (EXIF) metadata attached to digital photographs. In light of recent news surrounding the collection of public phone metadata by the National Security Agency, archivists should also be aware of the fact that even seemingly innocuous metadata can be combined with other sources of data to reveal sensitive information.\textsuperscript{16}

Lee and Woods provide a comprehensive inventory of the kind of hidden files that may be found in whole computers preserved using forensic approaches. These include, among others: log files, user account information (including usernames and passwords), cached network traffic, web-browsing activity, deleted files (“de-allocated but not overwritten”), and “slack space - data existing in blocks associated with allocated files that have not been completely overwritten”\textsuperscript{17}. Such files may contain recognizable PII like credit card information, tax records, medical records, and social security numbers, or sensitive information that creators may wish to have restricted. A 2003 forensic study of over 150 discarded hard drives purchased from the secondary computer market found an astonishing amount of private and sensitive data, including several examples of credit card numbers retrieved from log files.\textsuperscript{18}

The tension between privacy challenges and preservation opportunities created by digital forensic tools clearly manifests itself in the example of archiving Salman Rushdie’s computers. The Manuscripts, Archives, and Rare Books Library (MARBL) at Emory University acquired the papers of Rushdie, the internationally known author of \textit{The Satanic Verses}, in 2006. The collection ultimately included four computers (a Macintosh Performa and three Powerbooks) no longer in use by the author. As part of a visionary approach to preserving and providing access to
these literary papers, Emory staff produced a forensic image of one of the computers and made it available as an emulation of the original computing environment to on-site researchers.

Owing to the threats on Rushdie’s life that resulted from the publication of *The Satanic Verses*, the restrictions agreed upon by Emory and the author extend beyond the routine (financial and legal data) to include redaction of personal information related to Rushdie’s friends and family, including the phone numbers, fax numbers, and addresses found in his correspondence. As staff at Emory began processing the computers, however, they encountered difficulty in the redaction of such PII: “the time, resources, and development needed to effectively redact sensitive information from the correspondence proved too great for the work schedule and resources established for the first phase of processing.” Consequently, Emory staff were forced to limit access and completely restrict all digital correspondence. The tension Emory describes is not unlike the one that archivists must consider in applying minimal processing methodologies to large collections that may contain PII: how best to balance privacy concerns with researcher access and the realities of institutional resource constraints.

Redaction may ultimately become easier through the ongoing development of forensic tools. The BitCurator project, with its effort to develop a suite of forensic tools specific to the archival context, is advancing the profession’s technical capabilities considerably. Yet redaction of PII in digital files proceeds from the identification of PII, and identification remains problematic, despite the availability of many powerful tools. Archivists at Stanford University have described their experience identifying PII in 30,000 unique digital files in the STOP AIDS Project records using forensic software called Forensic Toolkit (FTK). While the tool supports keyword and phrase searching within documents, as well as the identification of character strings reminiscent of credit card or social security numbers, the archivists found the results to be
imprecise at best, and misleading at worst. Inundated with false positives, they realized that files would have to be individually examined in order to limit access to PII. Like Emory, they struggled to find a balance between access, privacy, and resource constraints: “We also struggled with how much time and how many searches constitute ‘due diligence’ in searching for restricted information. We ran what we thought was an appropriate number of pattern and keyword searches, but it is always possible that some files containing private information may have slipped through the cracks.” Ultimately, the need to manually review the PII identified by forensic tools limited their ability to explore more nuanced arrangement and description methods for the overall collection.

Further complicating matters, the existence of PII, even restricted PII, may present new institutional risks. For repositories housed in larger institutions, especially academic ones, strategies related to storage and processing of files that contain private information may be affected by institutional network and data security policies that evolve in response to the continued threat of cyber intrusion. As computers and web technologies have proliferated in our lives and workplaces, the field of cyber security has grown to become a sixty-seven billion dollar industry with news stories of data breaches becoming a common occurrence. The University of Maryland recently suffered a data breach that exposed sensitive information for over 300,000 students, staff, and faculty, despite an overhaul of the University’s security infrastructure just prior to the breach. It is important for archivists to realize the policies developed in response to these security risks will not necessarily account for the ways in which archival repositories may wish to preserve and secure PII within their collections.

Academic institutions, in particular, have begun more formally to approach the issues associated with data security. In addition to being subject to such laws and regulations as FERPA
and the HIPAA Privacy Rule as mentioned previously, universities are increasingly being required to meet the security and access requirements of the U.S. federal agencies that fund academic research. This trend has led libraries to take a more active role in providing tools to help researchers manage, preserve, and provide access to research data, but a less visible trend has been the effort across higher education to create data categorization and protection policies. The authors’ parent institution provides a useful example. In 2012, Penn State adopted a data categorization policy “to establish the categories of sensitivity that apply to data processed, stored or transmitted within the University and its systems.” This data categorization policy identifies the responsibilities of university employees to identify and protect sensitive data, and includes language reflecting the potential sanctions that may result from violation. But some of the aforementioned strategies for preserving born-digital materials will not necessarily make it simple for archivists to comply with these policies. Institutions that acquire whole computers and preserve them using forensic strategies may in fact be preserving data that an institution may have a legal mandate to purge. Yet, in creating a bit-for-bit image of a computer, an institution would be preserving all files contained on the computer, even those that fall into restricted categories.

Further, archives may ultimately be required to apply their limited resources in ways they would not normally, simply to meet the requirements established by institutional data security policies. Such policies could be interpreted in a way that requires archivists to immediately process born-digital materials after acquiring them, regardless of whether resources exist for such work. It may not be viable to simply create disk images and deposit them into storage to await further analysis later, despite the fact that our resource limitations, evolving practices, and the still-developing tools to support these practices, often dictate this kind of staggered approach.
4. Recommendations

Despite the evolving challenges presented by born-digital material and the emerging landscape of solutions we employ, there are a number of activities that can be undertaken to begin building a foundation to support working with these materials. Such steps -- which include engaging different stakeholders and developing policy frameworks -- will not only be significant for helping to mitigate the risks associated with PII, but will also contribute to establishing formal programs for managing, preserving, and providing access to these holdings.

a. Develop policies that include more specific language about born-digital materials

The Association of Research Libraries SPEC Kit 329 examined approaches to managing born-digital collections. The authors reported that 71% of institutions surveyed had some form of established policy addressing whether “files with PII should be retained with restrictions or destroyed,” yet only 43% of respondents had developed policies indicating whether such born-digital collections could be accessed without being screened for PII.26 These results highlight some policy gaps that institutions might seek to fill in. Policies addressing how to manage the existence of PII might naturally evolve from existing policies related to PII in analog materials; but in order to accommodate born-digital materials, institutions should seek to clarify the strategies they intend to employ, including whether or not they intend to preserve obsolete media and computers as forensic images.

Institutions that adopt forensic approaches to preserving born-digital material will need to determine how to handle hidden files, slack space, and other ambient data (such as browser caches) that might contain PII or other forms of sensitive information. Some have developed
policies explicitly stating their intention not to preserve hidden or deleted files accidentally transferred to an archives, believing such steps to be a violation of donor privacy and intent. Such policies need not take a firm stance on such data, however. Some institutions may instead decide to take a more nuanced approach and simply acknowledge the likelihood of encountering such data and give it an appropriate amount of attention. In explaining their decision not to recover deleted files on Salman Rushdie’s computers, Emory noted their intent to make decisions about hidden files and similar data on a “collection-by-collection basis,” based on factors that equally benefit both the donor and the repository.

Some of this policy language should make its way into gift agreements. The 2013 CLIR report, *Born Digital: Guidance for Donors, Dealers, and Archival Repositories*, provides an example:

“All parties should be aware that repositories may discover materials not intended to be included in a collection (as indeed has long happened with paper archives), such as files created by a user other than the donor or items that fall outside of the scope of the acquisition.”

Doing something as simple as adding an addendum to your standard agreement asking their assistance in letting the repository know that PII is potentially present in the records can be a good start. An example of such an addendum can be found as Appendix A.

b. Engage collection donors/records creators about their born-digital materials

Addressing the attendant privacy issues through policy statements is one way to begin engaging potential donors on the potential risks associated with transferring born-digital materials. More generally, archivists may find it prudent to engage donors in a conversation
about PII very early in the negotiation process. Doing so can help alleviate worries about access (including the potential concern that their files will end up on the World Wide Web), raise awareness about data persistence and archival recovery possibilities, and even trigger donors to self-appraise their records for PII in advance. The authors of the 2013 CLIR report call on archivists to “make sure donors and dealers are aware of the different kinds of deleted information that may be present in their digital materials” and recommends that archives inquire about the presence of private or sensitive information before the materials are transferred to the repository.

Template surveys such as the one created by the AIMS Project provide some insight on the level of up-front technical appraisal that may help identify PII issues in advance. Such advice is sound, but this level of interaction may not always be possible with each and every donor of born-digital material. We should make every effort to have these discussions, but accept that they may not always produce the desired results. Where this level of engagement is not possible, archivists might consider amended donor agreements that address these concerns (as suggested above) or establish certain levels of restrictions on such material until it can be adequately appraised. A strategy of this nature would require an institution to have a firmer sense of the strategies they will apply to preserving and managing born-digital materials.

c. Develop strategies for preserving and managing born-digital archives

Establishing policies and engaging donors on potential privacy risks implies that institutions know what tools and approaches they will use to manage and preserve born-digital materials. These techniques will likely be fluid and continue to evolve as institutions experiment and share their experiences and as tools grow in complexity. It is important for archivists to have
a sense of what options exist for dealing with born-digital material, and what capacity they have for implementing them. Making the institutional commitment more explicit does not mean there will be a single approach. Archivists will likely develop many workflows for dealing with born-digital collections in different contexts. These workflows will be dependent on a range of variables, including the results of interactions with donors and dealers, which will dictate what content gets acquired, how it is transferred, and how restrictions and access will ultimately be handled.

If archivists are unable to limit the amount of sensitive data being transferred as part of larger deposits, or engage donors to help identify where PII might persist in transfers, then it must have strategies in place for assessing born-digital acquisitions. The current backlog of legacy media holdings in archival repositories provides a useful example. For decades repositories have received floppy disks and optical media incidentally, as part of larger analog collections. While these disks have languished in boxes, the content contained within them has been left unappraised by archivists. Future acquisitions may contain both obsolete and modern forms of media, with the newer forms likely consisting of gigabytes or terabytes of data. In either case, a repository cannot address PII without first having an intentional strategy for working with born-digital material.

Archivists now have at their disposal a plethora of strategies for engaging born-digital materials. Literature in recent years has covered everything from grounding our approaches in traditional archival practice to identifying introductory steps that institutions of varying resources can implement to acquiring computers and creating forensic images. This documentation provides any institution with pathways for beginning to manage born-digital holdings. The two most commonly implemented approaches to transferring born-digital files from fixed media
could be generally described as copying files and directories, and creating forensic images. Archivists should explore how these approaches fit with their available resources, how they complement their strategies for working with creators, or within institutional policy frameworks, how each might enable or limit their ability to identify and manage PII. An institution that does not have the resources to employ forensic tools as a long-term preservation and access strategy could simply leverage available data security tools, such as commercial products like IdentityFinder.  

**d. Engage institutional IT staff**

Archival commentators have long advocated for greater collaboration between archivists and other institutional staff that may be instrumental to establishing successful electronic records management programs. Writing fifteen years ago, Philip Bantin asked: “Who are the people who manage the central databases in your institution? Who is involved in decision support? Is there a group of individuals who meet regularly to discuss data and information management issues? Who is involved in risk assessment and management? Who are the individuals auditing the institution’s information systems?” Early recommendations of this nature seem to have been intended to position the archival repository as a stakeholder in institutional discussions related to digital documentation. One early study, for example, identified the poor perception of archives by other institutional stakeholders as a significant barrier to successfully developing an electronic records program. 

The Duke security audit is an excellent example of the benefits realized from inter-institutional engagement, but such interactions could also be the start of a larger strategic interaction between archivists and technologists. With IT assistance, archivists might explore the
implementation of needed software, including software that can scan for PII. Critically, IT staff can also help establish secure server space to store born-digital archives and do so in a way that complies with institutional policies related to PII.

The Penn State Special Collections Library staff have had an ongoing discussion with information technology staff. Originally, this engagement was necessitated by the desire to secure the appropriate hardware necessary to support the management of legacy born-digital holdings. The hardware included both a dedicated PC for working with born-digital materials and a modest amount of network storage to use as a dark archive (i.e., secure, restricted access storage) as recommended by OCLC. These discussions helped IT staff better understand the archival requirements underpinning the request for hardware and network storage, which in turn prompted them to recommend better security arrangements to meet institutional information security protocols. For example, IT staff decided to make not only the network storage dark (accessible only to archival and IT staff), but also limited user and external server access for the PC used to work with born-digital materials, to provide an added layer of security. As the Penn State Special Collections Library’s born-digital program evolved, planning began for a repository to support preservation and management activities for born-digital archival collections. These earlier discussions turned out to be essential foundational steps in preparing for this project. Having understood the archival risks associated with born-digital materials, the IT staff were able to recommend a technical architecture that would effectively quarantine data within the repository environment without limiting the archivist’s ability to eventually process the material and make it accessible for researchers.

e. Develop institution-specific approaches to providing access and engage researchers on
Archivists should be aware of impact their preservation and access decisions can potentially have on researchers viewing a digital archive. Privacy will likely be a prevalent issue for researchers, even when PII has been adequately protected by the repository. The biographer of the late author, Susan Sontag, demonstrated how the digital research experience can be both exciting and disquieting for a researcher, as he described his experience using the digital files in the Sontag papers at U.C.L.A. Benjamin Moser used forensic images transferred to two external hard drives in the reading room of the library under the supervision of archivist Gloria Gonzalez. Moser said the experience left him “queasy” and that it was unlike looking through an author’s personal papers and manuscripts. “Looking through someone’s e-mail is quite another experience, and the feeling of creepiness and voyeurism that overcame me as I sat with Gonzalez struggled with the unstoppable curiosity that I feel about Sontag’s life. To read someone’s e-mail is to see her thinking and talking in real time.”  

Likewise, Emory archivists found that files redacted from the emulated Rushdie computer due to privacy concerns created a small amount of confusion among some researchers.  

It is notable that neither the emulated Rushdie computers, nor the digital Sontag papers are readily accessible to researchers online. Because of these technical realities, and the desire to manage the copyright and reproducibility issues associated with some collections, institutions have provided few access mechanisms that allow for remote access. But even on-site access requires a certain level of mediation by archivists, as the Sontag example demonstrates. Archivists that deploy forensic solutions should realize that disk images are not widely known or understood outside the technology field, and any forensically-managed material delivered to researchers will likely require some assistance.
This may seem onerous to some, but the interaction required to provide access to born-digital collections also presents an opportunity to educate researchers on the persistence of PII in digital files. If archivists must, as described by the 2013 CLIR report, “be realistic about restrictions, redaction, and the potential for sensitive material to be missed in an initial review,” then it will also be important to spend some time working with researchers to explain their responsibility in protecting PII (and identifying it for the repository) where they inadvertently stumble upon it. Such approaches, which should be familiar to archives addressing similar privacy issues in analog collections, might further be solidified in researcher agreements and policies.

5. Conclusion

In many ways, these recommendations mirror the risk management approaches advocated for by Greene and Meissner when coupled with MPLP processing strategies. Whether because of the complexity of technologies used to create cultural heritage, the evolving strategies for preserving it, or the archival profession’s ongoing negotiation of best practices in a world of diminishing resources, it is likely impossible to collect born-digital materials and be completely free of risk. But we can begin to mitigate and manage those risks by educating the various stakeholders -- donors, information technology colleagues, and researchers -- and by making and documenting the intentional choices about how we manage born-digital materials.

We should start by collectively attempting to define what constitutes acceptable risk when it comes to born-digital materials. As our experience with these materials grows, we will likely better understand the challenges and opportunities, and the risks associated with each. For example, email seems to be emerging as a hot zone for privacy risk, but certainly there are other
areas that could use further examination such as telephone and tablet data. Such a discussion would benefit from a common understanding of these risks, which reinforces the need for archives to move past ad hoc management of born-digital materials and start developing programs with defined policies, strategies, and practices which should be shared with the wider archival and library community as these standards evolve.

As with all aspects of working with born-digital archives, we should develop, in concert with formal policies and practices, a high tolerance for uncertainty. Despite our best efforts at outreach, we may still find ourselves challenged to address privacy risks in a manageable way. The Stanford and Emory experiences examined here are instructive: both repositories directly engaged the records creators in a dialogue about the nature of the digital materials being preserved, including where and what kinds of PII may be present, and both created project teams that included archivists and technologists alike. Yet each institution found their desired outcomes limited by the intensiveness of resources required to process their respective digital collections.

The existence and likely persistence of these ongoing challenges reinforces the need for archives to move forward in building programs with defined policies, strategies, and practices and toolsets for the management of born-digital materials. Policies should be developed with an eye toward how they can be both informed by and used to inform the various stakeholders such as donors, records creators, IT staff, and researchers. Practices employed to manage and preserve born-digital material should be solidified enough to explain the institution’s commitment and responsibilities, while remaining flexible enough to accommodate future innovations and growth. The end result may not be security without any obscurity, but we will have a better understanding about how to identify, manage, and contain risk (if not totally protect) the born-digital archival record.
APPENDIX A

Deed of Gift Addenda for collections with electronic records

The Donor acknowledges that the Library acquires the materials with the intent of making them available for an ongoing or indefinite period of time. In order to accomplish this, the Library may need to transfer some or all of these materials from the original media as supplied by the donor to new forms of media to ensure their ongoing availability and preservation. The donor grants the library rights to make preservation and access copies of materials in the collection and to make those copies available for use.

The Library may contract with university staff or outside contractors to store, evaluate, manage and or analyze materials in the collection. Any such arrangements must abide by the terms of this agreement.

Does the Library, or contractor as outlined above, have permission to crack passwords or encryption systems, if any, to gain access to electronic data received as part of the materials?  ____ Yes  _____ No
If no, such materials will not be retained by the Library.

Does the Library have permission to recover deleted files or file fragments, if any, and provide access to them to researchers?

____ Yes
____ Yes, under the following conditions
____ No

Does the Library have permission to preserve and provide access to log files, system files, and other similar data that document your use of computers or systems, if any are received with the materials?

____ Yes
____ Yes, under the following conditions
____ No

Privacy

The Library will review the materials in the collection in an attempt to identify items that contain private information. Please indicate below your awareness of materials that may sensitive electronic information.

____ To the best of my knowledge, these materials do not contain sensitive electronic information.
___ I believe that the materials are likely to contain private or sensitive information such as
   ___ Social Security numbers
   ___ Passwords or PINs
   ___ Credit Card numbers
   ___ Financial records
   ___ Medical records
   ___ Licensed or pirated software
   ___ Other materials that have specific privacy concerns, please specify________________

**Materials Not Retained by the Library**

In the event that the library locates duplicative materials within the collection or materials that are not of enduring research value the Library will remove, discard and/or destroy said materials. In the case of media carriers for born digital content, the Library will either return them to the Donor or physically destroy them after the content has been migrated to new media and verified.

---

2 Information technology staff interview with Seth Shaw and Tim Pyatt, Duke University Library, March 2010.
5 Danielson, p.319.
6 Danielson, p.329.


17 Lee and Woods, p.301.


26 SPEC Kit 329, p.18.

27 Carroll, Laura, et al., p.68.


29 Redwine, Gabriela, et al., p.10.


38 Redwine, Gabriela, et al., p.10.