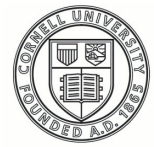

Cornell University
ILR School

Practical Technology for Archives

Volume 1 | Issue 7 Article 1

1-2017

Access and Preservation in Archival Mass Digitization Projects

John Yolkowski
Saint Mary's University

Krista Jamieson
University of Alberta

Access and Preservation in Archival Mass Digitization Projects

Description

[Excerpt] In 2014, the Dalhousie University Archives began its first archival mass digitization project with the Elisabeth Mann Borgese fonds. The successful completion of this project required the project team to address both broad and specific technical and intellectual challenges, from rights management in an online access environment to the durability of the equipment used. To best understand the challenges faced, there will first be a brief introduction to the fonds and project goals of balancing preservation and access before moving on to a discussion of these challenges in further detail, and finally, concluding with a discussion of some considerations, best practices, and lessons learned from this project.

Keywords

mass digitization, access, preservation

Access and Preservation in Archival Mass Digitization Projects

John Yolkowski
Saint Mary's University

Krista Jamieson
University of Alberta

Introduction [i]

In 2014, the Dalhousie University Archives began its first archival mass digitization project with the Elisabeth Mann Borgese fonds [ii]. The successful completion of this project required the project team to address both broad and specific technical and intellectual challenges, from rights management in an online access environment to the durability of the equipment used. To best understand the challenges faced, there will first be a brief introduction to the fonds and project goals of balancing preservation and access before moving on to a discussion of these challenges in further detail, and finally, concluding with a discussion of some considerations, best practices, and lessons learned from this project.

Project Background

Dalhousie University received a personal donation of $100,000 CDN to embark on a mass digitization project for the Elisabeth Mann Borgese fonds—an amount of money sufficient to fund the project for one year (2014–2015) and purchase necessary software [iii] and hardware[iv]. The funding was tied to this fonds and was marked specifically for a digitization project by the donor, a personal friend of Elisabeth Mann

Borgese. One of the key requirements laid out by the donor was for the digitized materials to be made accessible online due to the history of overseas use of this fonds and the international nature of much of its content. However, accessibility is about more than simple availability, so where and how the digitized materials would be displayed, having full text search made possible through optical character recognition (OCR), and access points for the materials needed to be incorporated into project planning. An Access to Memory (AtoM) platform was chosen for online access because of its ability to attach digital materials directly to standards based finding aid descriptions. Online publishing also required that Dalhousie grapple with the issue of copyright in the digital world, an issue not considered when most of the material was created. The Dalhousie University Archives, having never undertaken a project of this nature, decided that this project would act as a pilot for in-house mass digitization.

The Elisabeth Mann Borgese fonds is comprised of 55.5 meters of textual records, plus audio-visual material, housed in 372 archival boxes. The fonds contains materials dating from the 1930s to 2002 and spans the personal and professional life of Elisabeth Mann Borgese. Elisabeth Mann Borgese was a professor of international law, with a focus on maritime law, and a key contributor to the Third United Nations Conference on the Law of the Sea (UNCLOS III). This fonds is arranged into 13 series and numerous sub-series reflecting a wide variety of disparate topics, including administrative records from various organizations in which Borgese was involved. The fonds included everything from personal Christmas cards, photographs, correspondence, audited financial statements, manuscripts (fiction and nonfiction written by Borgese and others), and audio recordings to meeting minutes, reports, and press releases. The fonds was donated after Borgese's death. Most materials were in good condition and a RAD [v] finding aid previously existed for the fonds, though as part of the digitization project the finding aid file level descriptions were updated. As part of the donation, Borgese's family signed a donation agreement that included the copyright to materials Borgese herself created, but this could not cover materials in the fonds created by colleagues, friends, or even NGOs Borgese worked for; and, as was quickly

discovered, materials under the copyright of Borgese account for only part of the material donated.

With such a large fonds containing a variety of material types, scalability was a major issue for the Dalhousie University Archives. At the outset it was estimated that approximately 35% of the collection could be digitized within the one year time frame[vi]. Scanning documents is time consuming and doing an in-house digitization project on this scale as a pilot meant making decisions about how to approach the project and also how to balance the major competing archival goals of preservation and access. Our project mandate prioritized access over preservation with online access being a project requirement from the donor while preservation scans were not; favouring access over preservation could have avoided many of the technical issues this project faced[vii]. However, focusing on access to the detriment of preservation leads to issues of time consuming rescanning in the event a preservation scan is needed for publication and causes a significant issue if there is ever damage to the original. Increasing intervention with archival materials, such as re-scanning, requires more staff time and more decision making leading to increased time, and therefore cost, per item. This knowledge lead to a project mantra of 'more product, less process' wherein the project team tried to balance access and preservation and their associated practical and theoretical concerns to satisfy needs without significantly increasing effort and cost [viii]. The practical and theoretical challenges then divide into categories of technical and intellectual concerns involving scalable digital storage and preservation, scanning processes, selection, rights, and availability.

Technical Challenges

When conducting a digitization project, the technology involved cannot be discounted in decision making processes because of its significant impact on workflows and technical requirements. Technical challenges faced by this project touched on issues of scale, image quality and resulting file size, digital storage, the act of digitization, and

the creation of access copies, including OCR. These issues also interact with one another, compounding or creating issues in other technical arenas. For example, the digitization specifications relating to image quality resulted in large files that created a massive amount of data requiring digital storage. It is the preservation scans then that were the root of many of the technical concerns, though not all of them.

Scanning itself became a technical issue the project team had to grapple with. At the start of the project, it was believed that the scanner's Automatic Document Feeders (ADF) would speed up the scanning process. Pages of uniform size that were in good condition and not fragile could, theoretically, be scanned faster through a feeder. It was determined that groups of up to 20 pages could be scanned through the feeder at one time before the pages started to skew[ix]. Feeders, however, use a very small area to capture the image as the page is fed through. As a result, at this quality, even a small speck of dust would create brightly coloured lines running the length of the page, requiring that the glass be cleaned up to half a dozen times before a completely spotless scan was produced. With use, dust became a more and more common issue. In the end, the feeders were almost never used, as the time it took to check scans, clean the window and rescan pages far outweighed the time saved. Thus, the flatbed ended up as the preferred scanning method. In general, however, the scanners started to wear as commercial flatbed scanners are not meant for scanning high quality scans, hundreds of pages a day for months on end. Halfway through the project, the scanner started to freeze or have trouble connecting to the computer and processing the scans. Despite efforts made to make sure there were good connections between the scanner and the computer, the troubleshooting that worked was to reboot the scanner by turning it off, letting it "cool" for 30 seconds or so, and then turning it back on to restart the scanner. It is unclear whether this was the result of standard wear and tear on the equipment (the project used the scanner more intensely than the scanner specs advised), or if there was a software or hardware issue beyond wear and tear that was affecting the scanners.

The technical specifications for this project, such as the resolution and file format, were based on the Dalhousie Library scanning standards[x], optical character recognition (OCR) requirements, and our desire to attempt to make the scanned images resemble the original pages as closely as possible. Based on these requirements, pages were scanned using an Epson DS-60,000 scanner at 600 dots per inch (DPI) in 24-bit colour and saved to uncompressed TIFF files. Using EpsonScan software, the saturation was increased and the brightness was decreased at the time of capture, rather than during post-production, in an attempt to capture the colour of the paper and ink without washing out the pages[xi]. These standards provided us with publishable, high quality images that faithfully represent the material in the collection. This digital duplicate ensures that if ink or paper fades, or a page is torn, stolen, replaced out of order, or otherwise destroyed, there is a snapshot of the fonds a decade after it was accessioned.

The specs used for scanning for this project raise potential points of criticism. In a mainly textual fonds, there is a limited amount of colour. It could be suggested that scanning in full colour is not justified considering the increase in file sizes and the space that could be saved by scanning in greyscale. However, few real-life documents are strictly greyscale—paper might be coloured cream or some other shade, letterhead may include significant colour designs, ink might be blue or red, and so on. It was decided that scanning everything in colour to capture those elements without debating which pages merited colour scanning and which didn't, was more in keeping with the "more product, less process" mantra the project had adopted.

The other major issues created by the technical specs relates to the sheer size of the digital material created. Each letter-sized page resulted in a file 97 MB in size. Scanning between 450-900 pages a day, roughly 40-85 GB of data were created each day for the duration of the year long project, resulting in an estimated 13 TB of data for 125,000 scans[xii]. Scans were then transferred to the library server nightly. With each scanned page image being nearly 100 MB in size, the files were simply too large for

practical online use and access. Since the project goal from the donor was accessibility focused, making access copies was integral to this project. It was decided that the pages from each file folder would be compiled and compressed into a single multi-page PDF representing the contents of each physical file folder from the fonds (as opposed to treating each item in a file as a discrete digital item) [xiii]. This could be done by the Dalhousie University Archives using Adobe Acrobat software, but processing time would be significant. The notion of processing access copies at a later date was considered, however the time to re-ingest the preservation scans from long-term storage added unnecessary process to the workflow and was decided against. LITS (Dalhousie's Library ITS) solved the conversion problem by creating a command line script on the server that could automatically compress and compile TIFF files stored on the server, turning them into multi-page 300 DPI PDFs. This was accomplished by first turning the TIFFs into 300 DPI JPEGs as an intermediate step to compress them and then turning the JPEGs into a compiled PDF. This intermediary step allowed us to achieve a more manageable PDF file size. However, as IT staff are not archivists, maintaining an archival standard for access copies required negotiation. IT staff wanted to improve the appearance of the files by digitally removing pencil marks and dirt, and save space by rendering the images in greyscale. In order that the PDF files remain as true to the original document as is reasonably possible, it was necessary to convince IT to maintain the files in full colour and not process to remove marks. Automating this step of the workflow allowed computer processing time to be run overnight rather than being run one by one by the project team. This saved processing time as the computer network was not as heavily used overnight as it was during the day and also meant that staff did not need to either wait for each file to process before starting another, or lose the use of one of the computers as a team member went back and forth between different tasks while processing the scans into PDFs. Automation does require extra effort in quality assurance and possible losses in a one size fits all approach to processing; the idea of 'more product, less process' was applied to our decision making on what steps to

automate and saving time and effort through automation was determined to be worthwhile for this step.

One of the processing steps that our team did not automate, however, was the OCRing of the documents. The PDFs created using the script developed by the IT department were manually run through OCR software (ABBYY FineReader) by the project team in order to make them text searchable. The possibility of automating the OCR process on the server using a different software was considered, however, a server based OCR software would require a significant investment of time for the software to "learn" to recognize characters and words correctly, whereas high quality results were available using a desktop based software. Given the technical challenges faced by this project already, it was decided that this issue could wait for the next project to be solved.

The final major technical issue faced by this project was the active storage of preservation scans for the duration of the project. The Dalhousie University Archives had been storing all reference and project scans on the library server, Digi. At the start of the project we estimated data of 13 TB being produced over the year. While the project team was also creating access scans during this time, the file sizes of the multi-page PDFs were so much smaller than the preservation scans that the storage needs for access copies paled in comparison to the needs of the preservation scans. LITS was confident that adding 4 TB drives to the existing 4 TB library server would be sufficient to hold the influx of 13 TB data. After 4 TB of data were created a few months into the project, however, the weekly backups of the server could not handle the extra data and the server crashed. LITS spent nearly three months rebuilding the server to have independent 4 TB servers which would show seamlessly on the user interface and were able to be individually backed up on a regular basis. This unexpected crash and delay meant an adjustment to storing material on external hard drives to be uploaded to the server at a later date. As well, the crash resulted in a loss of approximately 300 scans that needed to be redone. As server storage was only ever a temporary solution during

production, the Dalhousie University Archives has approached a third party about cloud storage for long term data storage.

Intellectual challenges

Intellectual challenges were encountered from the very onset of the project. The first concern, which any project team with a limited amount of time needs to address, is choosing what should be digitized. As noted above, the fonds is vast (55.5 meters), and could not be digitized in its entirety within the year the project was funded. Examining the materials, we saw a wide variety of topics of potential interest to various scholars and other archival users. Some of the topics included international development, international relations, marine research, environmental studies, German literature and culture, Canadian politics, legal studies, and animal intelligence.

After performing some tests to determine the speed at which the team could be expected to digitize materials, it was estimated that we could reasonably expect to complete digitization for only about one-third of the collection. Based on this approximation, two possible approaches were identified: The first would be to digitize the collection in physical order—start at box 1, and move through each box in order, digitizing every folder in turn. Selecting this strategy would mean that, over the course of the year, we would have digitized the first 120 boxes or so in the fonds. As the intellectual and physical arrangement are separate, this would result in a sample of materials from various parts of the collection being digitized, but there would be no guarantee as to how much of each series would be digitized.

The second approach involved ranking series by their interest to researchers and stakeholders and digitize series-by-series, with the series believed to be most interesting digitized first. The latter approach was the one chosen by the project team. One stakeholder in particular whose interests we considered was the donor, who was a personal friend of Borgese and had a keen interest in the International Ocean Institute,

an NGO founded by Borgese. Other areas of interest included records relating to the United Nations' Convention on the Law of the Sea (UNCLOS) negotiations, and Borgese's writing, personal records, and photographic materials. To begin, we chose to test the workflow by digitizing the smallest series in the fonds (from Borgese's time as Chair of Canadian Crown Corporation, the International Centre for Ocean Development) to test our strategy and workflow and adjust it to the materials before moving on to more time-consuming series. With this approach in place we made a priority list of series for digitization. These series were:

1. The administrative records of the International Centre for Ocean Development (ICOD)
2. The United Nations
3. Publications, drafts, and speeches
4. The administrative records of the International Ocean Institute
5. Elisabeth Mann Borgese's personal records
6. Photographic and audio-visual materials

Once we decided which series to prioritize for digitization, we encountered a second intellectual consideration: copyright. A major part of this project was to make material available online, so copyright was a significant issue we had to grapple with. the Dalhousie University Archives had received the copyright to the materials created by Elisabeth Mann Borgese in the fonds upon donation, however, as is the case with many personal fonds, such material accounted for only a portion of the total material.

Being in Canada, the project team was required to work within the framework provided by Canadian copyright law. This means all works are copyrighted from the moment of their creation with no need to register their copyright, as copyright is automatically granted to the author until 50 years after their death. Thus, individual

pieces of correspondence, for instance, would be considered the copyrighted property of their original authors. As the majority of materials in the fonds date from the late 1960s to the early 2000s, virtually all of the materials in the fonds are still under copyright, meaning that the Dalhousie University Archives would be exposed to risk if the materials were just placed online without addressing copyright first.

As the goal of our mass digitization project was to make materials available online with "more product, less process," we had to address the copyright challenge in a manner which would allow us to continue production in an efficient manner. Digitization couldn't be stopped to clear copyright on items one at a time. Thus, we had to address the "orphan works paradox[xiv]." This is the idea that works of the lowest commercial value take the most in terms of resources, when it comes to clearing copyright[xv]. Tracing the copyright owners for items created by these groups (which given the political nature of the fonds we were digitizing, were numerous in our case) is virtually impossible. Given that the fonds has thousands of individual items of orphan works, clearing each and every item would have greatly reduced the amount of material that could be placed online[xvi]. Addressing the problem was crucial to creating a functioning workflow, since we did not wish to scan large amounts of material that could not be placed online.

Thus, copyright presented the project team with a major intellectual challenge to making digitized materials available online, especially working within the time constraints of the project. Canada's Copyright Act contains several exemptions which were considered, most powerfully Fair Dealing (as outlined in Section 29 of the Act, it is similar, although not congruent with, the American concept of "Fair Use"). However, there is much debate about whether Fair Dealing could be used to justify the dissemination of archival materials on the open web. Fair Dealing, in Canada, is decided on a two-part test. First the "dealing" must be for one of the approved reasons explicitly mentioned in the Act (research, private study, criticism, review, or news reporting. In 2012 three additional categories: education, satire, and parody, were added) [xvii]. Then,

as a 2004 Supreme Court of Canada decision laid out, six additional factors must be weighed to examine the individual use [xviii]. Jean Dryden, in her 2008 PhD dissertation, argues that there are two concerns with applying this concept to the dissemination of archival material on the web. One is identified by legal scholar Wanda Noel who argued, in a brief provided to the Bureau of Canadian Archivists' Copyright Committee, that it would be a stretch to see mass dissemination as fitting into one of the approved reasons. "Research" falls the closest, but Noel argued that it would be too broad of an interpretation of the category[xix]. As well, Dryden argues that even if one could get the posting of materials to fall within one of the approved reasons (education, added four years after Dryden and Noel's analysis, might be worth considering), it would still, as Dryden argues, run afoul of one of the factors set out by the Supreme Court of Canada: the character of the dealing[xx]. According to that ruling, the fewer the copies made, and the more transient they are, the more likely the copying/reproduction will be Fair Dealing. This does not seem to be in keeping with the aim of posting large numbers of archival documents online[xxi].

Given that we likely could not rely on Fair Dealing as a strategy to disseminate orphan works online, we had to find another solution to the orphan works paradox. One was found in the examination of copyright policies created by other institutions. Simon Fraser University developed a unique approach to addressing this problem: risk analysis[xxii]. The idea of applying risk analysis to the problem of orphan works in archival collections is a viable option when the rarity of copyright lawsuits against archival institutions is considered[xxiii]. To that end, we applied this type of framework and thinking as part of our processing of the collection. When preparing the materials for digitization, we assigned the materials in the files as either "low risk" or "high risk". "Low risk" materials could be posted online immediately without seeking permissions for the copyright holders, while "high risk" items could only be disseminated with the explicit permission of the rights holders. This allowed us to address the problem of the "orphan works paradox," and keep our decision making process within the spirit of "more product, less process." To this end, we worked to identify considerations that

would help us decide whether something was “low risk” or “high risk.” In general, we decided the following considerations would guide this assessment: copyrighted works could be disseminated without seeking permission if an evaluation of the following factors led one to conclude there was minimal risk. The three factors are:

1. There is a low-risk of damage to the university’s reputation resulting from dissemination

2. There is a low-risk of damage to relationships with university donors, or university communities

3. There is a low-risk of exposing the university to copyright infringement claims [xxiv]

As well, one of the lessons learned from this project was the codification of this into a “Copyright Assessment Worksheet[xxv].” This allows the individual assessing a file of archival material to identify questions surrounding the above three questions by asking them to look for certain material that may prove high risk (i.e., books or articles from publications with ISBNs/ISSNs, as there is a high risk of publishers claiming infringement) [xxvi].

Our risk assessment approach to copyright comprised only half of the copyright framework we needed to develop. The other half was establishing a robust take-down procedure, and communicating this to rights holders. To this end, we constructed a take-down policy, which runs through the process that is to be taken when an infringement claim is made (the material is immediately removed, an independent assessment is undertaken by the library’s Copyright Office, and a final decision is made and communicated to the parties involved)[xxvii]. As well, a form was placed online which rights holders can fill out to begin the process of filing a claim[xxviii]. This makes it a straightforward and transparent process for all those involved.

By the project’s end in July 2015, a total of 1,508 files were online (a further 300 scanned files were uploaded at a later date). Between the time of the first file being

posted in February 2015 and the time of this article's writing in August 2016, a total of one infringement claim was made. An individual asked that correspondence between him and Borgese, as well as a contract included with the correspondence, be removed. The process worked, and the material was removed from the file[xxix]. Although it is hard to identify the number of third-party copyright owners in the collection, the fact that only one has expressed concern in the past year, indicates we likely struck the right balance in respecting rights, while still being able to provide access to the voluminous materials in the fonds. Thus, we had a copyright framework that followed the spirit of "more product, less process."

Lessons Learned

This brings us to our final point: in conducting mass archival digitization projects what advice can we provide in creating a balanced project that meets both the goals of access and preservation? Three ideas are worth noting: automation, documentation, and risk assessment.

First, the concept of automation. Obviously, the number of steps one can cut out of a project the easier it becomes to complete multi-step tasks. A major contribution was the script that compiles preservation TIFFs into access PDFs. This allowed us to complete two steps at once, and allowed us to use time when we were not working (overnight) in an effective manner. In mass projects, it is not scalable to treat each file and item individually. If you are applying standards to a step in the workflow, automation can make that step significantly more efficient. Quality assurance is important however, so it is necessary to account for time needed to check automated processes to make sure they are working properly and results are up to standard.

Second, document as much as possible. It is important to record what was done and, even more so, when. As the project unfolded it became clear that time spent on certain tasks could end up taking either more or less time than originally thought.

Technical and intellectual challenges affected the workflow and so it became essential to try to track what steps had been done both with respect to access and preservation. For instance, whether the copyright issues of a file scanned months ago for preservation been resolved. In dealing with a small fonds with a dozen files it is easy to keep track of all of the competing claims, but when dealing with thousands of contacts, it is important to track the material. Our tracking sheet (an Excel file with the tasks that had to be completed for each step[xxx]) was essential in helping with this. It also has the added benefit of assisting in long-term preservation goals as it allows the Archives to know the date of each step and provides a basis of preservation metadata for the scans. Thus, it can be helpful in answering a question such as "is it time to migrate forward?" No format will last indefinitely, so that cannot be our aim when digitizing; all we can do is make informed choices about formats we will still be able to work within a few years' time, making that date stamp on each step invaluable. If the material was scanned two years ago or seven years ago, it becomes an important consideration for migration and maintenance workflows. Knowing the timeframe, quality, software, and file types, and not having to spend the time to figure it out later, can save a lot of work in the future. Sharing documentation is also important for the archival field. Being able to find out how other projects dealt with challenges can save time and effort by not repeating work others have completed. Making automation codes available, presenting project details at conferences, and publishing at least project summaries on institutional websites are all valuable resources for others in the field. Our project produced significant numbers of institutional standards and practices surrounding specifications and copyright, all of which are available through the institution's website. Challenges and lessons learned detailing failures are also important; glossing over problems doesn't allow for anyone else to learn from your mistakes and leads to institutions repeating each other's mistakes.

Third, we learned about striking a balance when it came to copyright. We wanted to avoid an overly conservative approach, which would involve either contacting each and every rights holder for permission and restricting online dissemination to

those works for which we obtained explicit permission. This approach would have the result of privileging preservation over access, since much of the material could not be placed online under such a scheme. Our attempt at balancing the dissemination and creator's rights through risk assessment led to a more efficient and, as our one complaint reveals, relatively safe, system for allowing access. Thus, we can conclude that applying a risk management framework to copyright in digital collections is a useful and functional model moving forward.

This concludes our outline of the Elisabeth Mann Borgese fonds digitization project. It is our hope that by sharing the lessons outlined here, this project will be of use to archives interested in mass digitization as one of their major outputs in the future, providing insight into how we balanced various intellectual and technical concerns, with an eye to both access and preservation. Our specific decisions were related to project goals and were informed by the resources, budget, support, and technology available to us at the Dalhousie University Archives. Though we consider this project to be one of mass digitization, really this project was about a standards based approach to digitization that tried to walk the line between boutique digitization and large scale projects that do not address issues like selection or rights management on a more granular scale. Online access has created shifts in approaches to copyright that are still developing and changing on a regular basis. However, our hope is that the work we have done on this project points a way forward for a shared culture of practice regarding archival dissemination, at least until legislation catches up to digital access.

About the authors

John Yolkowski holds a MLIS from The University of Western Ontario. Between 2014-2015 he worked as the Project Manager for the Elisabeth Mann Borgese Digitization Project, and is currently the Acting Librarian–Archives, Special Collections, and Records at Saint Mary's University in Halifax. He is especially interested in copyright and archives.

Krista Jamieson has an MLIS from McGill University and an MA from the University of Amsterdam in the Preservation and Presentation of the Moving Image. She was the Digitization Specialist for the Elisabeth Mann Borgese project at Dalhousie University and is now the Digital Archivist at the University of Alberta.

Notes:

[i] A version of this paper was presented at the Association of Canadian Archivists annual conference in Regina, Saskatchewan, in June 2015.

[ii] The Access to Memory (AtoM) based finding aid including attached access copies resulting from this digitization effort can be found at: Access and Preservation in Archival Mass Digitization Projects.docx

[iii] ABBYY FineReader had been used by the Dalhousie Library previously with good results and the additional licences needed for a project of this scale were within the budget.

[iv] The Dalhousie University Archives purchased two Epson DS-60,000 scanners before the project team was hired. These commercial grade machines were of sufficient quality to match the library's digitization standards (discussed below) and were well reviewed scanners at the price point available for this project.

[v] Rules for Archival Description, or RAD, is the Canadian descriptive standard for archival fonds and is maintained and updated by the Canadian Council of Archives.

[vi] This estimate was based off of timing how long it took to scan one box of textual materials and then extrapolating based on the duration of scanning one box and the project timeline. This estimate proved to be accurate as about one-third of the collection was digitized by the end of June 2015.

[vii] Note that preservation scans tend to be scanned uncompressed and at a higher DPI, making their file sizes significantly larger. This means that they are not ideal for hosting on the Internet, or for providing access to remote users.

[viii] To get an estimate of how long scanning would take, the project manager scanned one box of textual records and timed how many hours it took. Once the digitization settings were finalised, that box was rescanned with appropriate specifications. The extra day and a half this took drove home the point that rescanning was exceptionally time consuming and something to be avoided if at all possible.

[ix] The ADF scanners also added a layer of pre-processing to the material, as the condition of the archival documents varied widely, we had to flag anything that could not go in the sheet feeder.

[x] We were fortunate that a Working Group of the Dalhousie University's Library Council had already conducted a survey of digitization standards used at other universities (through institutions such as the University of Maryland, see http://ourdigitalworld.org/wp-content/uploads/2012/04/DigitizationBestPractices_Schreibman.pdf) and compiled institutional standards for the Dalhousie Library System. These standards provided a baseline and guidance for most of our materials (i.e., suggesting that copies not be made below 300 dpi). The standards had only been used for small scale projects and one-off reference digitization up until that point however, and so the Borgese project allowed the Dalhousie University Archives to test the standards more fully and expand them where need be, such as the DPI for slides and file naming conventions for scans. The updates to the standards the project team made were then incorporated into all digitization conducted at the Dalhousie University Archives.

[xi] The saturation was increased to 50 and the brightness decreased to -20 for textual documents.

[xii] Despite the size of the data created, scans were relatively quick, averaging under 30 seconds per page including processing time. Resizing pages or scanning double sided or bound pages added time, hence the fluctuation between number of pages scanned each day. With multiple team members scanning, a linear metre of textual records was scanned every eight to nine days.

[xiii] This was a conscious decision. Alexandra Chassanoff highlights in her article "Historians and the Use of Primary Source Materials in the Digital Age" (p. 470) that researchers are concerned about what is being included in digitization and the completeness of records. Compiling scans of entire files gave researchers an experience akin to that of a traditional reading room where they could look through the contents of a file page by page. Of course, as will be discussed below, given that files are not

homogenous, in some cases it was necessary to remove pages from the access copy (for copyright or privacy reasons). In this case a Separation Sheet was added to the file for the access copy. Using a one to one ratio of digital objects to file folder also allowed us to make use of the existing file level descriptions rather than having to create item-level descriptions for digitized documents and complex number systems to indicate page order within files. An example can be found below, see the link in footnote 29.

[xiv] This problem, and how it plagues archives, has been examined in detail by Jean Dryden. For instance, in a recent conference presentation she discussed this as one of the major barriers in digitizing archival collections. See Jean Dryden. "Releasing the Orphans" (presentation, Copyright in Canada Conference, Toronto, ON, October 2, 2015).

[xv] For example, take a letter from a former colleague of Borgese's written in 1981 where the colleague passed away in the 1990s; identifying a next of kin for the colleague can prove near impossible in some cases. Another example of orphan works are NGOs or political lobby groups that may exist either on an ad hoc basis or for a very limited time with no formal structure or governance.

[xvi] One other uniquely Canadian option that exists is that the Copyright Board of Canada does provide tariffs for orphan works. However, they can only do so in the case of published materials, not unpublished materials. Given that the bulk of archival collections are unpublished, this is of limited value in our context. See Katz, Ariel. "The Orphans, The Market, and the Copyright Dogma: A Modest Solution to a Grand Problem." Berkeley Technological Law Journal, vol. 27 (2012): 1322-1331 for a discussion of this process and the limitations.

[xvii] Unlike American Fair Use, which lists "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research", Fair Dealing is more conservative in that the list is exhaustive, there is no "such as" written in the Canadian Copyright Act.

[xviii] See http://www.canlii.org/en/ca/scc/doc/2004/2004scc13/2004scc13.html for the full text of the decision, and a discussion of the factors.

[xix] Dryden cites Noel's reasons in full: "There is a difference between copying for research and copying for other
purposes. Putting a photograph on a website or including a literary work in a publication like a book can NOT [emphasis in the original] be considered to be research copying. These, and other activities that are not research related, remain governed by normal copyright rules requiring permission from the copyright owner." From Dryden, Jean Elizabeth. "Copyright in the real world: Making archival material available on the Internet." PhD diss., University of Toronto, 2008: pg. 43-44.

[xx] Ibid. Pg. 43.

[xxi] A potential solution might involve adding a technological protection measure to the digital archival object, such as an IP filter for on-campus only access to documents or not allowing materials to be downloaded to personal computers, but this was not feasible to develop given the project's time limitations.

[xxii] See https://www.sfu.ca/archives/digital-repository/CopyrightPolFrame.html

[xxiii] Amanda Wakaruk of the University of Alberta, in a recent presentation, noted that her search of legal databases led to a total of zero relevant cases. From Don Taylor, Jennifer Zerkee, and Amanda Wakaruk, "Assessing Copyright Risk Tolerance for Large Scale Digitization Projects" (presentation, ABC Copyright Conference, Halifax, NS, May 27, 2016).

[xxiv] The full policy document is available here: Access and Preservation in Archival Mass Digitization Projects.docx. These tools were discussed in more detail by Creighton Barrett and Roger Gillis in a recent presentation: "Dalhousie Libraries' Copyright Tools for Online Collections" (presentation, ABC Copyright Conference, Halifax, NS, May 27, 2016).

[xxv] A PDF copy can be downloaded
here: https://libraries.dal.ca/content/dam/dalhousie/pdf/copyrightoffice/Copyright%20Assessment%20Worksheet_v1.pdf

[xxvi] An example of a "high-risk" file can be presented here: Elisabeth Mann Borgese met American novelist John Irving in the 1990s, and they shared a correspondence over a

five year period. Given Irving's status, it is likely that his correspondence could be commercialized and published. Placing his correspondence online without explicit permission would run afoul of the third factor discussed above. The Copyright Assessment Worksheet captures this in its statement that correspondence from "prominent literary, artistic, scholarly, or public figures" can be high risk.

[xxvii] The policy can be read here: https://libraries.dal.ca/content/dam/dalhousie/pdf/copyrightoffice/Takedown%20request_v1.pdf

[xxviii] The form can be downloaded here: https://libraries.dal.ca/content/dam/dalhousie/pdf/copyrightoffice/Takedown%20Request%20Form_v1.pdf

[xxix] We used separation sheets to remove pages from a file that have been deemed high-risk. See http://findingaids.library.dal.ca/sea-and-dreams-of-man-by-elisabeth-mann-borgese-published-version-and-handwritten-draft for an example.

[xxx] This was inspired by Petersohn, Barbara, Traci Drummond, Melanie Maxwell, and Kelly Pepper. "Resource Leveling for a Mass Digitization Project." Library Management 34, no. 6/7 (2013): 486-497.