Assignment 2

Archival Workflow of Born Digital Content

INF1341 | Prof. Arik Senderovich | MI - Faculty of Information

Word Count 3957

Nov 13th, 2019

Michael Andreae | 991426042
Thais Bittencourt | 1003575115
Emily Hunt | 1006372319
Ronnie (Yanzhe) Feng | 1006373069
Nadir Khan | 911096730
Yuanyuan Zhang | 1006580647

Executive Summary

The following assessment includes an overview and detailed analysis of the archival process used by one of the University of Toronto's libraries, EJ Pratt Library & Archives. The members of EJ Pratt, mainly the archivists, are tasked with analysing donated material and determining its value to the library's collection. As the digital world expands, digital content grows at an exponential rate. Although the current process may perform appropriately for physical content, it has not proven effective for digital-born material. This is due to the volume of the material as it can be significantly more abundant than traditional content. Currently, the process is manual and cumbersome, oftentimes consuming the majority of archivists' time. The proposed solutions target the most prominent issues identified by the analyses: accessibility to digital material, time effectiveness, redundancies and data privacy concerns. The scenarios created by the DFD group centered around improving the flow of digital material through the process whereas the BPMN group proposed scenarios that target the elimination of redundancies and improving the overall time devoted to tasks.

Table of Contents

1.0 Context for the Study

For years organizations have tracked their records with the help of archivists who assess and organise records for long term storage. Various technologies have come along to help this process, such as microfiche and cameras for digitization. All these processes involve the same intake and assessment workflows. However, in the last few years archives around the world have been facing a huge new source of materials, born digital content (Whyle, 2017). Digitally created material has to be processed in a different manner, partially due to the sheer volume of it's content.

The EJ Pratt Library and Archives, is one of the 42 libraries in the University of Toronto library systems. It focuses on acquisitions based on the teaching areas of Victoria University, a college affiliated with the University of Toronto. The library employs approximately 25 staff members whose overall goal is to develop the library's collection and provide support on accessing the materials (books, archives, and ephemera). EJ Pratt Library faces an increased volume of digital artifacts and they are working to improve their digital intake workflow.

The overall focus of the archival process is to determine whether a collection is of value and if deemed relevant to process the materials by determining what will be saved. By applying storage standards to and recording the metadata of those records they can be preserved in a constant state for future researchers (University of California Digital Libraries, 2019). The archivists are responsible for applying the archival process to the born digital documents and anything that can improve the efficacy of their workflow is of tremendous help (Whyte, 2016; Whyte, 2019).

2.0 Analysis using DFDs.

2.1 Detailed Presentation of the As-Is situation.

In the following analysis we explore the Data Flow Diagram we developed and use it to illustrate the current archival paradigm.

The archival team is taking in data from outside entities, including retired professors, staff, related institutions, and notable graduates, and processing that information with two pronged goal. The first is the long term storage of digital materials (Figure 1). The second is to create a display, or public access copy. Currently, the Discover Archives displays only the metadata about the archival records. Eventually the goal is to include the Dissemination Information Package as well (Islandora, 2019).

In the current data flow the archivists focus on ingesting the digital files and packaging the digital material. This has meant that analysis of the data and tracking its process have not received sufficient attention.

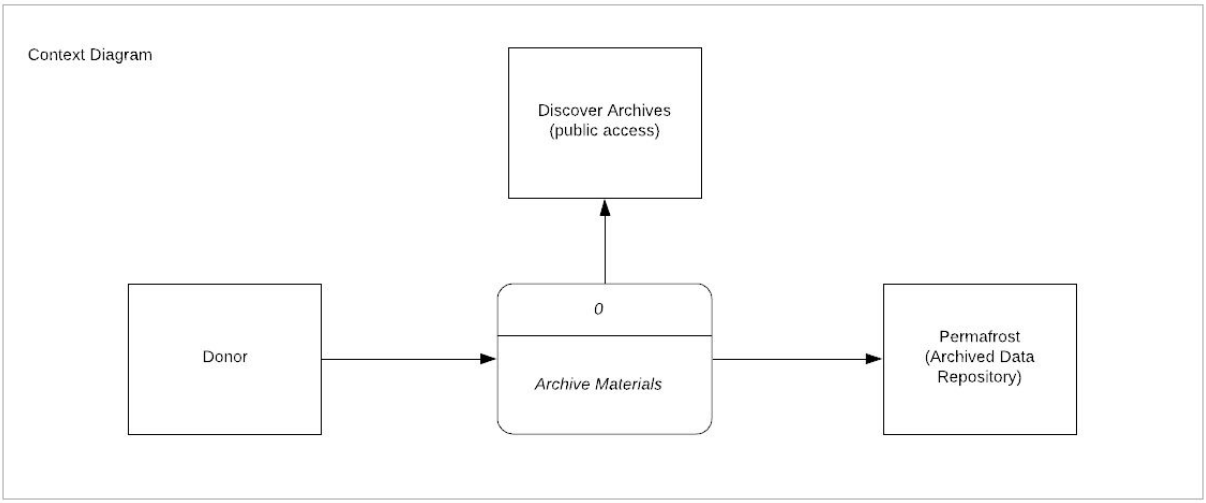

Figure 1. Context diagram

At level 0 of the DFD (Figure 2) the overall process becomes very apparent. The materials are brought into the institution where they are stored in three different places while the archival team migrates the data from raw, unstructured input into well formatted document collections.

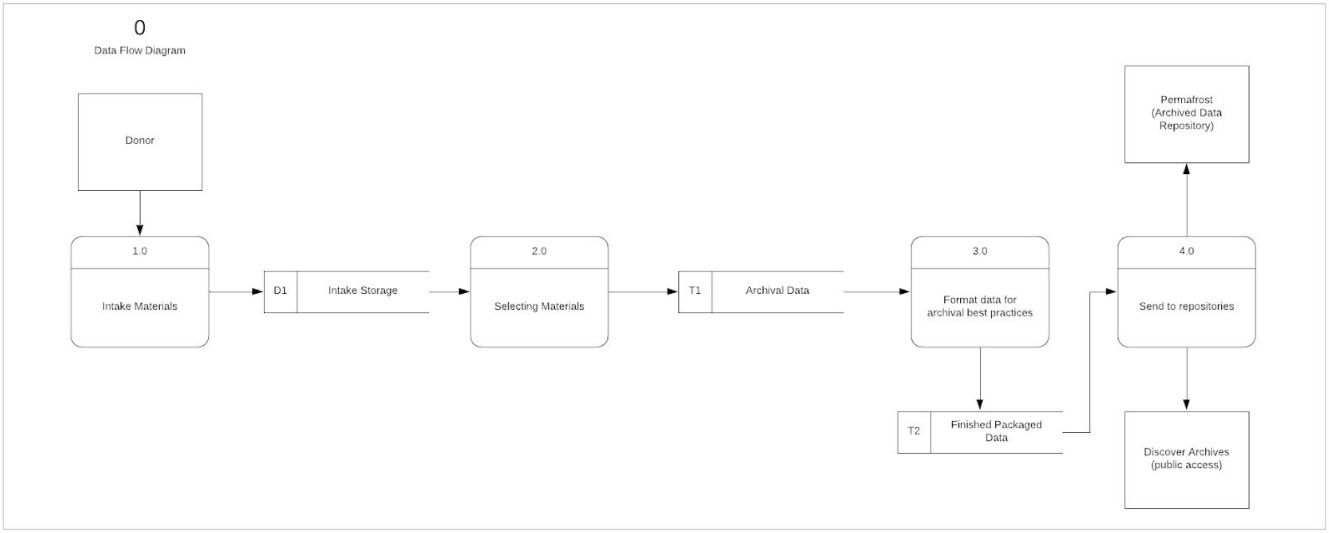

Figure 2. Data flow diagram 0

The first major process applied to incoming data is the processing of intake data highlighted in Figure 3. This is also the first touchpoint for metadata which gets written down in an Excel spreadsheet, or sometimes just in a notebook. Even at this early stage it was apparent that metadata would be accumulated at all stages of the process.

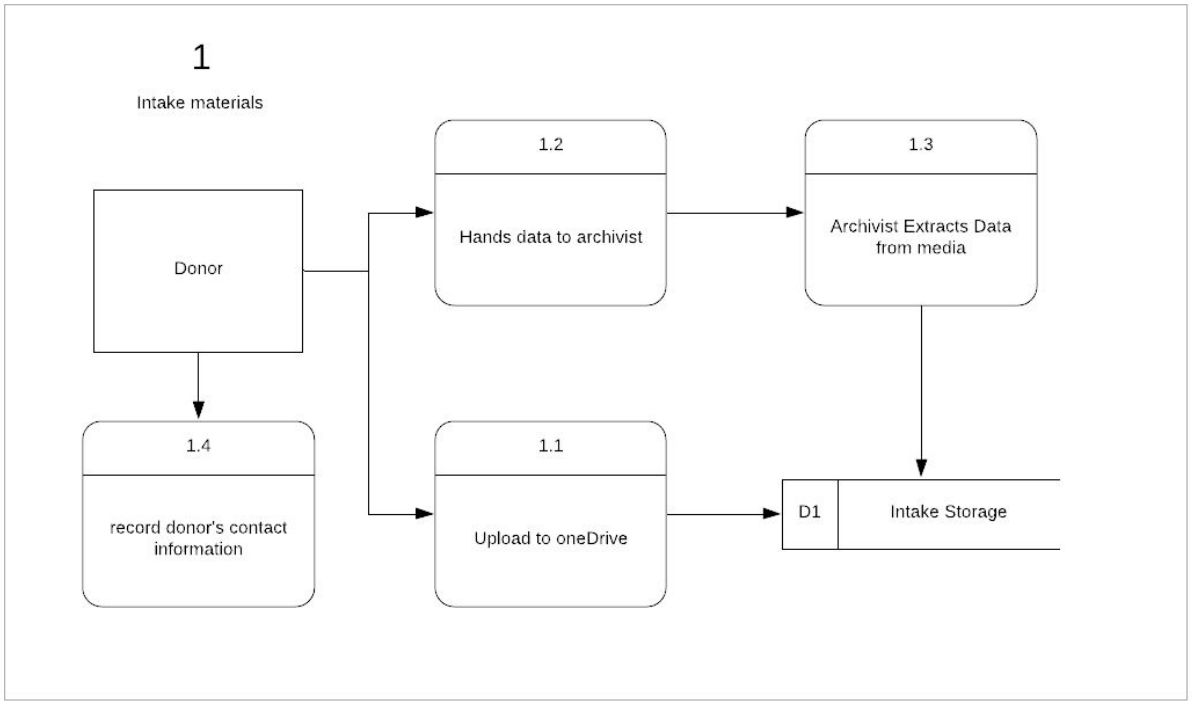

Figure 3. Data flow diagram 1

Once the archival materials have been stored in the selection process (Figure 4), the majority of the time consuming work is done in preparation for the best practices stage (Figure 5). This is an iterative process of analysis and deletion or retention applied to the records. Drawing on the archivists experience they are best equipped to examine each record, determine its relevance to the collection. Once complete the archivist is left with the pruned set of archival data which they wish to store.

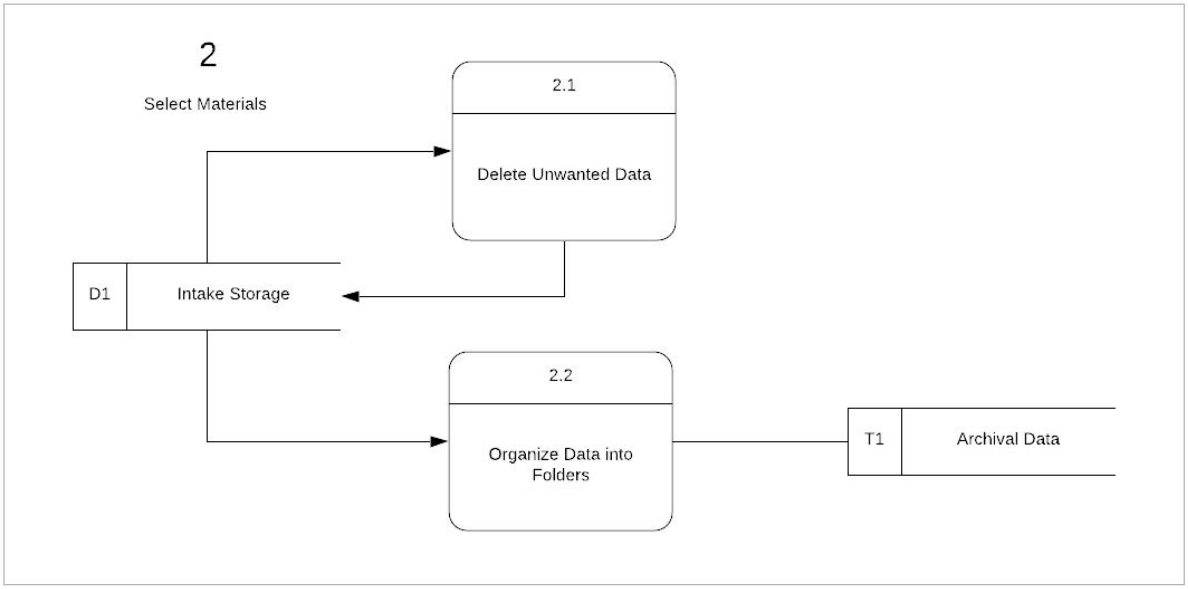

Figure 4. Data flow diagram 2

Now that a cleaned up collection of records has been saved the data has to be transformed into a format for long term storage (Figure 5). At this point the accession data is added to inmagic along with a description. This may involve adding various notes about the intake process as the accession record describes the process by which the data was prepared for archiving as well as the acquisition path. Currently to create these well formatted packages the library uses a tool called Archivematica which was arranged by a consortia of Ontario Universities (OCUL) to standardize digital archival packages (Islandora, 2019).

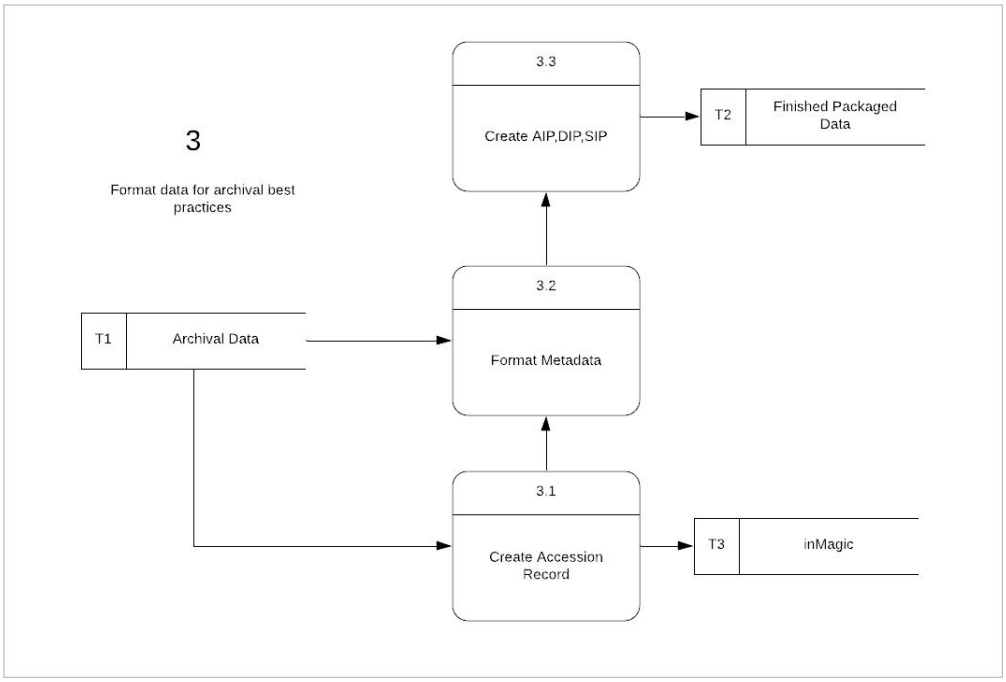

Figure 5. Data flow diagram 3

Finally the archival packages are saved in long term storage (Figure 6). Currently metadata is manually copied to Discover Archives, a University of Toronto based finding aid for archives and the data is stored in long term storage at Permafrost, also provided by OCUL.

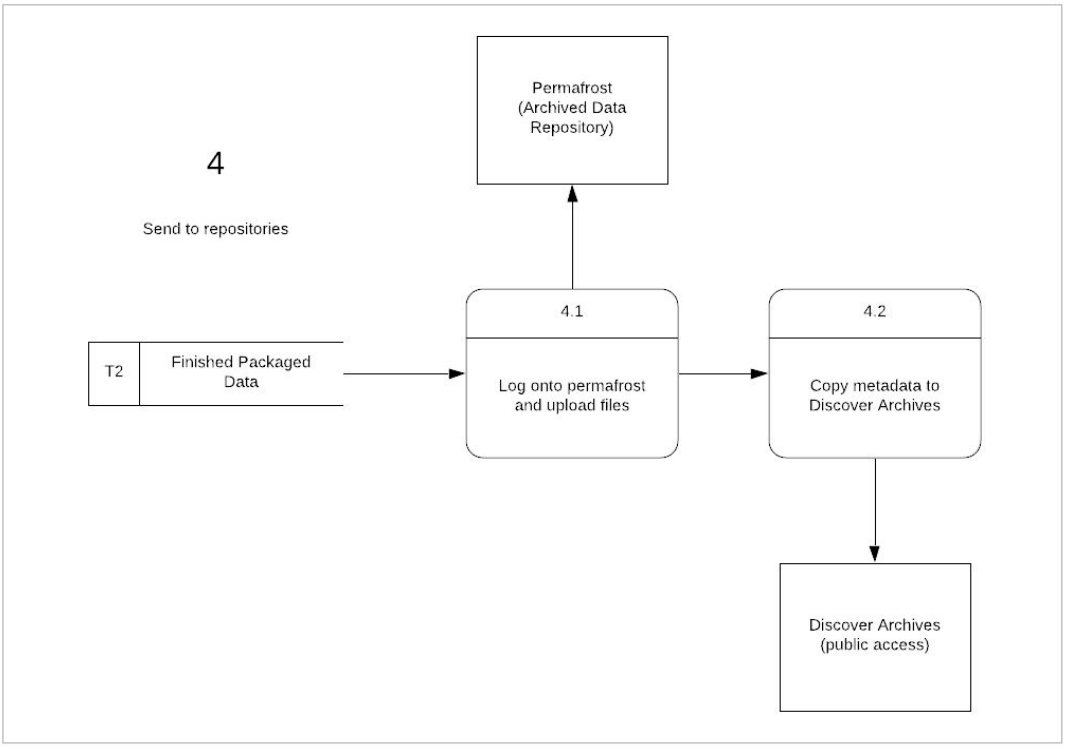

Figure 6. Data flow diagram 4

2.2 A Summary of To-Be Alternatives Considered.

Presented below are three "to-be" suggestions for automation and three for innovation using the DFD model to inform and inspire the ideas. They mainly focus on the intake and material selection process.

Automation

Currently, as shown in the DFD level 4 (Figure 6 from section 2.1) after the materials are donated and analysed, the archivist manually completes the packages and sends them to the storage areas. The first automation suggestion is to auto upload finished archival packages to the display and archival storage areas. Implementing this change would result in significant time savings for the archivists. This suggestion could be classified as a challenge with a big payoff in time, but moderately difficult to implement.

While studying the data flow another issue that became obvious was the metadata that is created at various processes but is not recorded anywhere. When it is recorded in a database, like inMagic, it is siloed away. The archivists mentioned using Excel spreadsheets and various notes to keep track of the metadata and at the end manually enter it into Permafrost and Discover Archives. The additional work of copy and paste with the errors that can arise pose risks and can add significant time to the overall process.

Another automation suggestion to improve the process is to protect against viruses. Digital media of all kinds exposes the organization to potential threats hidden in the digital files. After media is collected and assigned unique identifiers (IDs) in the intake process (Figure 3 from section 2.1), there should be an additional step of a virus detection screening process that would take place before the archivist extracts the data. Since this step would be time consuming, creating extra work for the archivists, the screening should only be for certain high risk types of media. Overall, this would be a beneficial step to add for the organization to protect against digital threats and easy to implement with existing virus detection software.

Innovation

The process of selecting and classifying materials, as modelled in the DFD level 2 (Figure 4 from section 2.1), can be mitigated by the help of the machine learning concept of topic modeling. Topic modeling is a model for discovering the abstract "topics" that occur in a collection of documents, such as archival material. This innovation would immensely expedite the process of selecting and formatting materials. However, this innovation would be difficult to implement as machine learning is still a relatively new field with much more work to be done for it to be reliable in its classifications.

In the intake process (Figure 3 from section 2.1), to make the process simpler for repeat donors, the organization should use TeamViewer to implement a oneDrive folder that can be accessed on the donor's computer. This would add an additional process of interacting with the donor. It would benefit the donor, save time in the long run and is generally easy to implement. Notably, one disadvantage of this innovation is the initial time that would need to be invested to work with the donors on creating the local oneDrive folders.

After extracting data from files, the archivist manually classifies and assigns metadata to digital files (Figure 6 from section 2.1). An enhancement to aid this process involves integrating a script program to look at files and folders and present the archivist a summary of findings (a processing information system). Benefits of this solution include the time saved by the archivist, while still trusting the archivist to check the work of the script. Limitations of this solution include the limited number of files that the script would be able to execute properly (e.g. image files would not be an option), and the solution could be a challenge to implement.

2.3 Detailed Presentation of Two To-Be Alternatives.

Automation

What emerged about the automation process we explored was the number of times metadata seemed to float in the ether until it is finally recorded in one place, inMagic. During the intake process there are a variety of times archivists create metadata related to the collection, for example the donors' contact information is written down (1.4) and various notes from steps in level 2 may be written down on scraps of paper.

Overall this means the archivist copies and pastes the Accession number multiple times and has to prepare different sets of metadata while they work. For the automation solution we propose replacing Excel spreadsheets and inMagic with an updated modern metadata management tool designed to export Permafrost and Discover Archives formated ingestible metadata. This means the archivist can record all the metadata and Accession information once and then export it to the various downstream display and archival repositories as needed. This creates a single point of data entry and flexibility in case the downstream storage destinations change.

Therefore we propose adding in a single simple to use repository for storing all metadata during the intake process, see Figure 7. In addition to the new datastore for metadata two small processes have to be added, 3.4 to link archivematica into the datastore, and 4.3 to export the data from the datastore to the online repositories. Right now process 3.4 is done by hand and copied over to spreadsheets or temporary notes. These disparate records are then combined in process 4.3 and copied into the final storage locations.

While the current inMagic tool is unable to be a single entry point and export tool, the new automated system would have simple data entry and programmable export formats, including an Electronic Archival Description (EAD) formatted option.

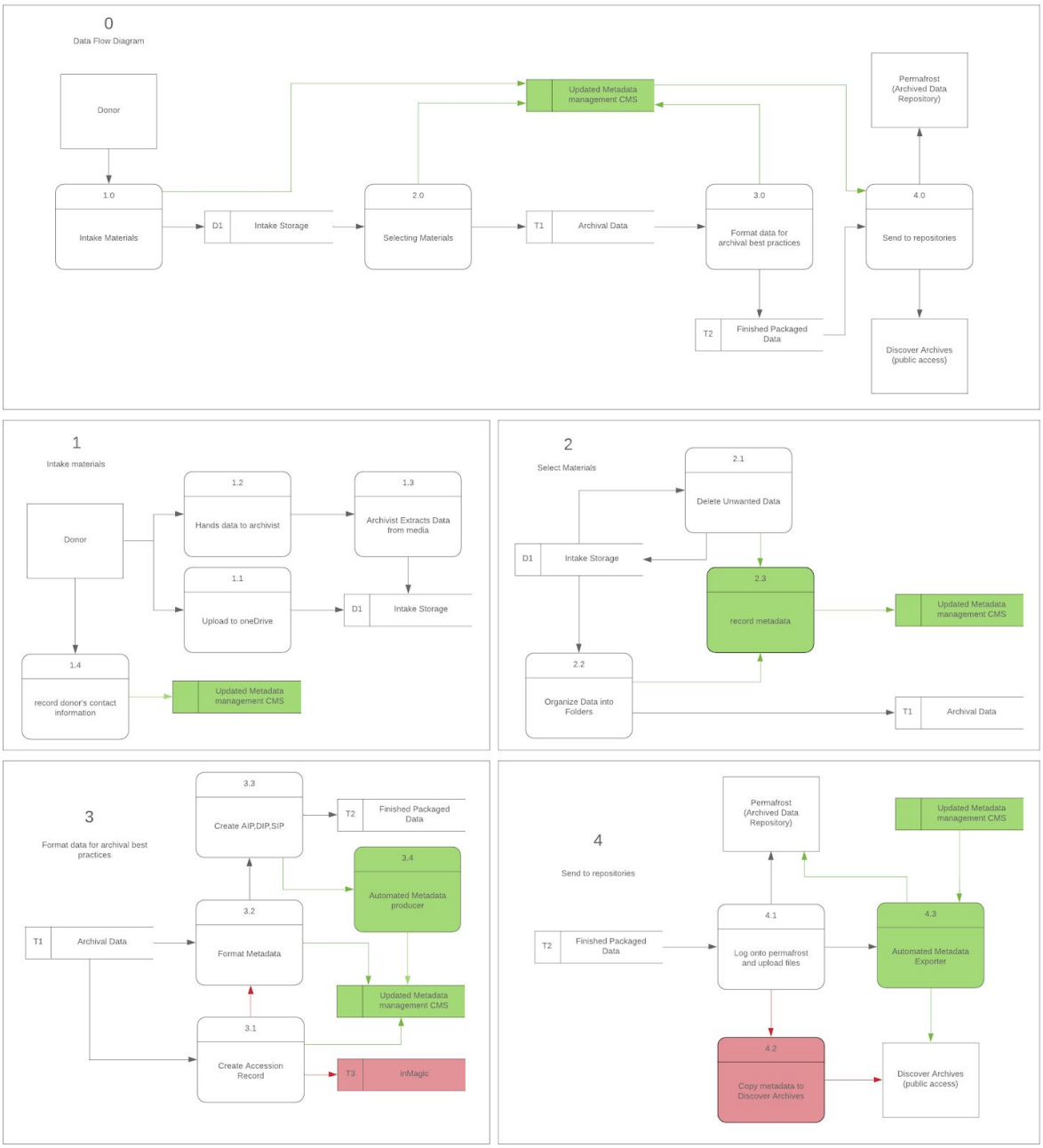

Figure 7. DFD After Automation Change

Innovation

One of the most useful innovation improvements focuses on level 2 where most of the skilled archival work takes place. While this is a relatively targeted process, it has invaluable time savings for the organization, making it a worthwhile idea to implement. We focused on this innovation process because most of the archivists' time is spent on understanding the contents of the data. Innovation in this area will provide high returns in terms of staff efficiency.

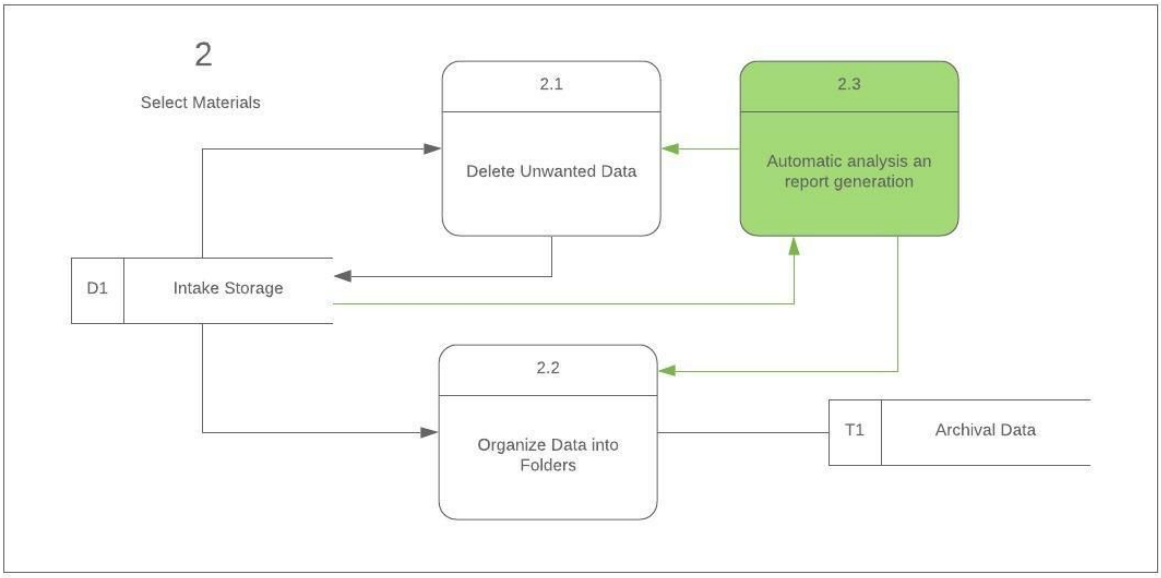

Figure 8. DFD Level 2 After Innovation Change

The innovation is modelled in Figure 8 shows level 2 of the DFD with the addition of a machine learning script, highlighted in green. The DFD model clarifies the innovation idea by simply visualizing the added process that relies on an integrated script program. This integrated script program looks at files and folders and presents the archivist with a summary of findings to increase their efficiency in the manual processes modelled in DFD level 2 (Figure 4 of the as-is situation). This proposal effectively removes the complex array of individual programs that an archivist uses, and applies a similar toolkit as Archivematica during the assessment phase.

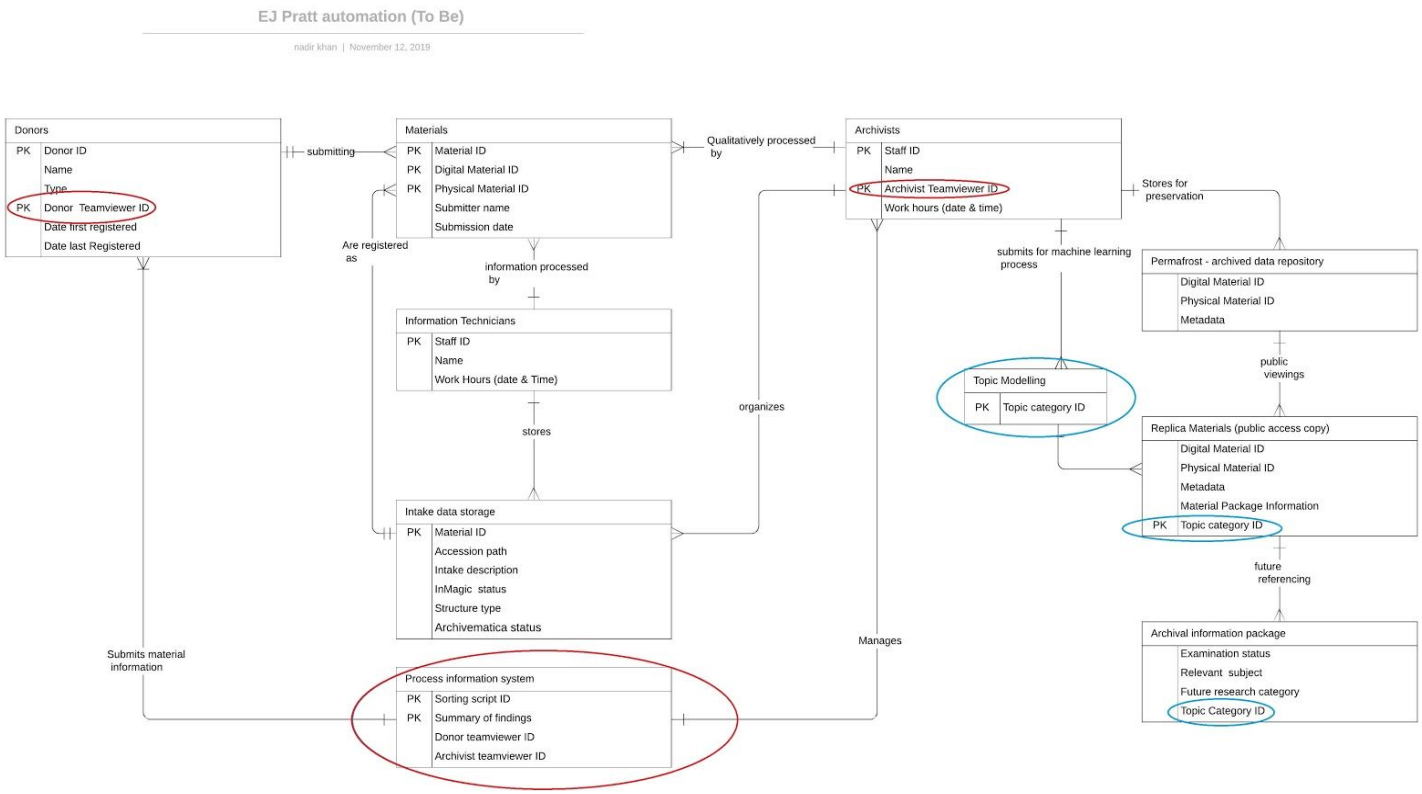

Figure 8.1. ERD Model Innovation Change

This innovation is depicted in our ERD model (*Figure 8.1.* ERD Model Innovation Change) below with blue circles; the addition consists of a new entity, “Topic Modeling”, to situate this topic modelling results into the process as well as amending new attributes to the entities, “Replica Materials (public access copy)” and “Archival Information Package”.

The red highlights in Figure 8.1 models the alternative innovation changes.

3.0 Analysis using BPMN

3.1 Detailed Presentation of the BPMN As-Is situation.

BPMN of As-Is:

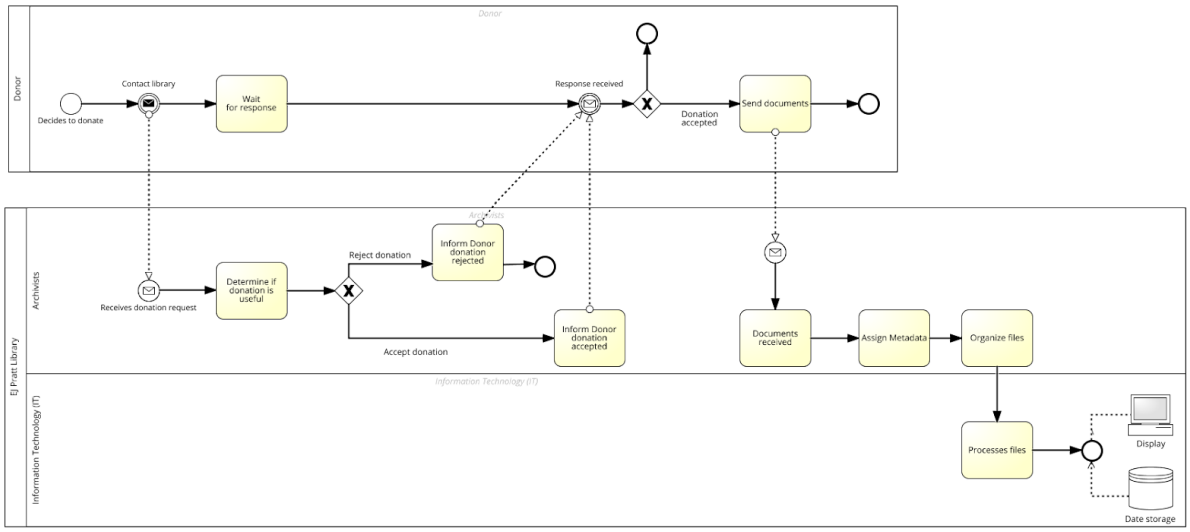

Figure 9. BPMN as is diagram

The BPMN model can be broken-down as follows:

1) Donor (individual and/or corporate) begins the process by contacting EJ Pratt's library archivist. The donor will then donate digital documents via email depending on the response received from the archivist.
2) The library reviews the donor's request to determine whether the donation is of value to the library archives. The process will end for the donor and archivist if the documents are not accepted. However, instructions are sent to the donor should the donation be accepted. Subsequently the archivist reviews and assigns metadata, as well as organises the documents after they have been received.
3) The library's Information Technology (IT) department will then process the files by storing them in a secure database located in the University of Toronto's server.

The objective for improving this process is to increase the efficiency of how donated materials are processed. Delays encountered during this process may cause the donor to lose interest in donating, which has a financial impact on the University. Furthermore, this process is limited by archivist experience and workload, outdated software packages and technology (i.e. slow servers and lack of virus protection). The main issues with improving this process involves the donor's willingness to contribute after delayed responses, as well as obtaining the Institution's investment into newer processes and computer equipment.

3.2 A Summary of To-Be Alternatives Considered.

The current intake process, as shown in the BPMN diagram, is carried out by two main actors: the donor and the archivist. Once the donor contacts the library, the archivist contacts the donor to request more information about the material to be donated. This creates an unnecessary workload for the archivist and the channel of communication breeds uncertainty. Additionally, the archivist has no incentive to expedite this process, as the archivist has other tasks to attend to on a daily basis. We also considered the safety of the library's database since there are many materials uploading to the database and no virus check process in place. Our team was also concerned with the archivist's workload as all materials are sorted manually. Based on those circumstances, the team devised 3 automation alternatives as well as 3 innovation alternatives.

Automation

The first automation alternative is posting the criteria on the EJ Pratt website to help the donor determine whether the materials are relevant. Instead of contacting the archivists, donors check the criteria themselves and then upload relevant materials.

The second automation suggestion is to have the EJ Pratt IT sorting files by categories. When donors upload the materials, they select the categories that best describes the documents. After the materials are uploaded to the drive, the EJ Pratt IT system sorts these materials to different categories by matching keywords and categories chosen by the donors.

The last suggestion is creating processes for archival information package (AIP) and dissemination information packages (DIP) to determine whether the materials are also packaged for displayed or solely stored. After archivists read the materials to determine their destination, all materials are sent to the system to process AIP and DIP. This process automatic sorts materials and determines if materials need to be displayed or stored.

Innovation

The first innovation alternative is to implement an online portal that acts as a preliminary screening to verify the owner's identity, collects owner information and collect the donation material. Rather than contacting the library, the owner would follow the instructions and steps of the online portal and upload material, which transfers the workload from the archivist to the donor.

Following the online portal, the second innovation point comes after the material is donated. As the process of intaking material gets automated, the library should expect large amounts of data to come into the library's data storage bank. The second innovation point is a virus detection procedure that scans viruses that could potentially damage the library's database. This ensures the safety of the library's database, as well as eliminating any malicious attempts to destroy records.

The third innovation point is structured to process large amounts of text material. In the current process, the archivist reads through the donated material, then decides whether the material has value. In order to make the material available for the public to see, the archivist also needs to draft a summary. The third

innovation point is a procedure that goes through large amounts of text material and generates a summary for the archivist to determine the relevancy of the material.

3.3 Detailed Presentation of Two To-Be Alternatives.

Automation

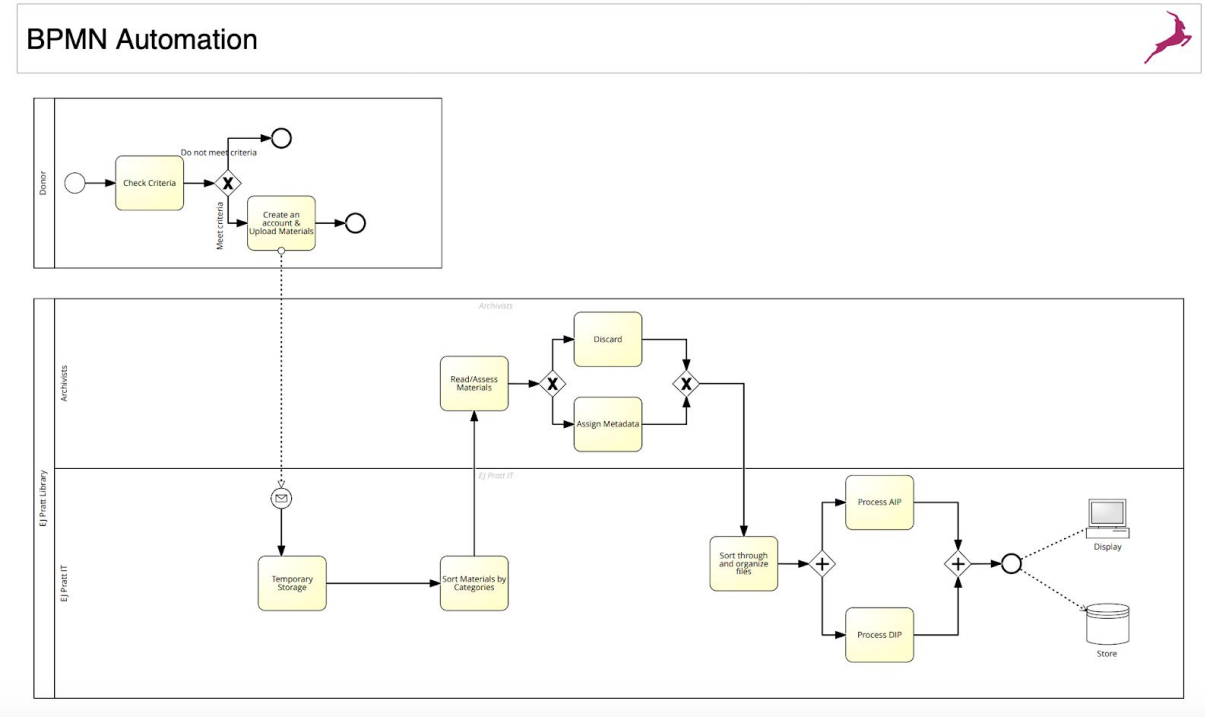

Figure 10. BPMN Automation diagram

In the current process, the donor manually contacts the library to inform the archivist about the donation. This requires the archivist to go back-and-forth with the donor in several emails to establish meaningful communication and determine whether the material has value to the library's digital archive. The first point of automation is structured at the beginning when collecting material from the donor. The suggestion is to build a sub-page on the library's website to inform donors about a clear set of criteria on what the library considers relevant and valuable.

This automation point would help potential donors to check against their material, so they can gauge whether to contact the library for the next steps. Additionally, this step would eliminate meaningless inquiries and save time and workload for the archivist. In this way, donors will only contact the library after checking through the donation criteria and make certain that the material is valuable to the library's digital archive. At the same time, this would give potential donors a better perspective of the archival process.

Innovation

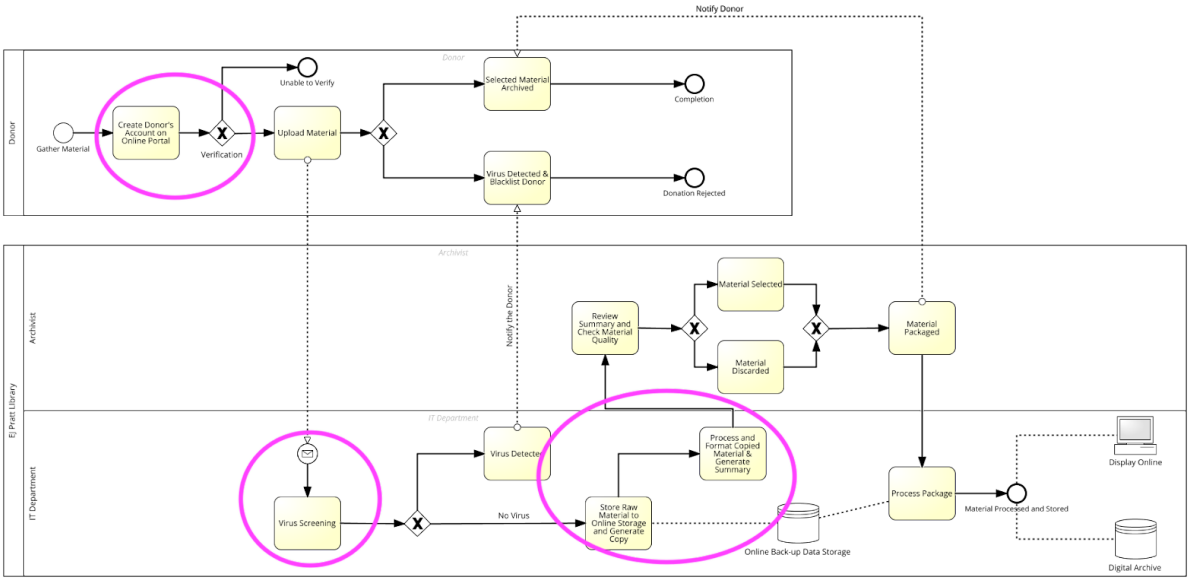

Figure 11. BPMN Innovation diagram

The first innovation point is implemented at the beginning of the archival process with the goal of eliminating this cumbersome procedure and allowing the donor to take the initiative to share the workload from the archivist. An online portal that acts as a preliminary screening check can solve this issue effectively. The online portal consists of several functions: register the donor, categorise the material, and verify the donor's identity. The library server will host the online portal where donors can register their information. This gives the library information about the donors such as their contact information and helps to expedite the intake process.

Another function of the online portal is verification. In the current process, the verification process is built on mutual trust. The archivist contacts the potential donor to verify their identification. This method is functional, but unreliable at a larger scale. When the digital archiving process becomes automated, large volumes of data is expected to be stored in the server, and a verbal confirmation of the donor's identity is not sufficiently secure to ensure the safety of the library's data. Therefore, when using the online portal, the donor is required to verify their identity before continuing. The verification will consist of a two-factor verification process requiring an email and a code sent via text message. The archivists only need to contact the donors who cannot fulfill the automated verification process yet still want to donate material to the library. The volume of people that the archivist contacts can be significantly reduced.

EJ Pratt Archives (University of Toronto)

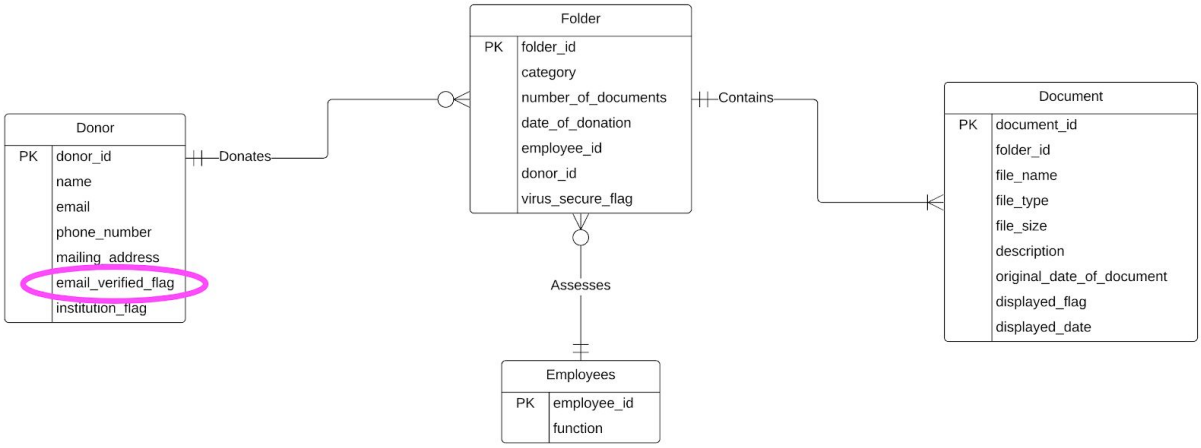

Figure 11.1. ERD Model Innovation Change

The innovation change is highlighted in Figure 11.1, which displays the attribute ‘email_verified_flag’.

4.0 Comparison of the Two Modeling Techniques.

4.1 Overview

While both DFD and BPMN focus primarily on processes, BPMN models a more accurate display of the process flow and sequence of tasks and as such our BPMN scenarios center around tasks such as eliminating redundant steps between the archivist and the donor. In contrast, the DFD facilitates tracking the flow of data and to isolate certain processes thus the DFD scenarios focus on the flow of data, such as formatting data with topic modeling and extracting data into a script program.

4.2 DFD Strengths

The DFD model is easy to use and understand by technical and non technical audiences which is useful when showing potential redesigns to stakeholders. It can go in depth with unlimited levels of detail. For instance *Figure 4* in section *2.1 Detailed Presentation of the As-Is situation*, displaying the process of selecting materials, is a detailed version of *Figure 2*. It easily outlines the data flow and boundaries of the system. In the context the Archival Workflow, outlining the data flow is essential to understanding the underlying processes that are in place and to inform "to-be" scenarios. Therefore, due to the nature of archival activities revolving around the flow of data, the DFD modeling technique provides many advantages over other modeling techniques.

4.3 DFD Weaknesses

DFD models may not be superior to BPMN models since it is difficult to check the accuracy of the logic behind them. For example, BPMN models are able to be computed by a computer program that can easily identify whether the model's logic is correct given a specific set of rules, whereas the DFD model can't be computed. Another weakness, is the time it might take to make the DFD models. Since the potential for unlimited levels of DFD models exist, the model could get extremely detailed and time-consuming.

4.4 BPMN Strengths

One aspect of the BPMN model that was lacking in the DFD was the idea that finding aids are created. By tracking the flow of archival data, the DFD team missed that finding aids are generated as a separate item. This was likely due to the BPMN's focus on the process, and facilitating the search and retrieval of information from the archives was much more apparent as a process rather than a data flow compilation and check for errors, which can be run through BPMN's model simulation.

BPMN results in better collaboration between the donor, archivist and IT personal, as the roles between each division are clearly defined.

4.5 BPMN Weaknesses

The weakness of the BPMN processes can occur when the processes expand and become more complicated. It was difficult to display how AIP and DIP sort and store materials in our automation suggestions. There was no data flow, so people lacking knowledge of library systems may not understand the process.

Additionally, the BPMN model does not showcase the details of innovation point. The entirety of the activity can be described as “online portal,” however, the dataflow, data storage, and the specific steps the donor takes in material intake is largely hidden. While BPMN gives a clear picture of the series of activities and presents the entire process, it does not provide detailed information of specific activities within the business process.

5.0 Methods, Activities, and Tools Used.

A team member interviewed the archivist and the systems supervisor at the EJ Pratt library. Materials such as articles relating to archival workflows, EJ Pratt's current direction documents, and digital intake, planning, and tracking documents were examined. We also researched other similar archival processes at different libraries. Based on our findings, we generated our as-is model.

The directory structure used by the library offered insight into the requirements and goals. The systems supervisor provided us with the current technical structure and the archivist provided key information about the current digital archive, and pointed out some of the pain points of the current process. Additionally, she offered her opinion on potential changes that would benefit the current system, which was quite helpful to our system analysis and assessment.

BPMN

We used signavio.com to create our BPMN models. After the interview with the archivist, we created the as-is diagram, which gave us a clear picture of the library's digital archive process. Given the information obtained from the interview and the as-is diagram, it was a straightforward process to discover the current process' limitations and comprise the to-be alternatives. Those diagrams disclose all activities in the processes but exclude detailed data flows. Without the explanations of the diagrams, there may be some confusion for those who do not fully understand all processes.

DFD

The 'as-is' DFD context diagram provides a simple overview of information being received from the donor, archived, and sent to the library website and the archived data repository. Furthermore, the hierarchical nature of Data Flow Diagrams, which were created with Lucidchart, also allows an infinite level of detail to be modelled while still being a readable format. The simplicity of the 'as-is' diagrams helped to show where processes could be automated for the 'to-be' models, resulting in a more efficient flow of information. However, the analysis using DFD models doesn't show the order in which simultaneous processes should be carried out and it was hard to check for the correctness of these models.

ERD

The ERDs were modeled using Lucidchart software. They were constructed once we understood the process and the data being captured at each step. The ERD helped to conceptualise the associations and the structure of the data being stored. The entities reflect the relationships between the actors and the digital materials involved whereas the attributes describe the stored content such as the metadata gathered for each donation material.

6.0 References

Guide To Archiving. (November 2019). Retrieved from https://learn.scholarsportal.info/all-guides/handling-digital-archives/

Islandora at the University of Toronto. (October 2019). Retrieved from https://open-shelf.ca/islandora-at-uoft/

University of California Digital Libraries. Retrieved November 8, 2019, from https://github.com/uc-borndigital-ckg/uc-guidelines/tree/master/INTRODUCTION

Whyle, J (September 2019). Fischer library digital workflow. Retrieved from https://connect.library.utoronto.ca/display/DPG/Fisher+Digital+Preservation+Pilot

Whyle, J (June 2017). *Preservation planning and workflows for digital holdings at the Thomas Fisher Rare Book Library*. JCDL 2017 June 19-23, 2017, Toronto, ON, Canada

Whyte, J (April 2016). Clearing the digital backlog at the Thomas Fisher Rare Book Library. Retrieved from https://saaers.wordpress.com/2016/04/12/clearing-the-digital-backlog-at-the-thomas-fisher-rare-book-library/

Statement of Individual Contributions.

DFD Group

Emily - a part of the DFD group. Significant contributions to the: summary of to-be alternatives, detailed presentation of innovation idea, comparison of models, and methods/activities/tool used.

Michael - Major focus was the introduction to the DFD section as well as working with Emily writing up the descriptions. Also contributed to some of the sections comparing the two techniques.

BPMN Group

Ronnie - primary focus is the innovation, parts of automation and comparison of BPMN. Analysis of innovation points and BPMN. Conducted on-site interviews with the archivist, secured meeting rooms and attended all meetings.

Yuanyuan Zhang - Primary focus is the summary of To-Be alternatives, detailed presentation of automation alternatives of BPMN. Major focus was the comparison of BPMN and methods, activities and tools used. Attend regular meetings.

Nadir - Prepared ERD based on DFD's, BPMN As-is model, BPMN strengths and weaknesses, group discussions, attended all meetings.

Thais - BPMN ERD, overall group discussions (focused on the BPMN), BPMN strengths/weakness, editing final paper. Attended all meetings.