A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value="true";

COURTNEY C. MUMMA, GLENN DINGWALL, and SUE BIGELOW

RÉSUMÉ Les Archives de la ville de Vancouver travaillent depuis 2008 à développer des archives numériques. La plus grande partie du financement et le sentiment d'urgence face à ce projet furent liés à l'obligation de la ville de préserver les documents du Comité d'organisation des Jeux olympiques d'hiver de 2010 (COVAN). Ce texte décrit quelques-unes des difficultés rencontrées à l'acquisition des documents numériques du COVAN. Il documente aussi les efforts des Archives pour faire correspondre les différentes tâches liées à l'acquisition et à l'évaluation du processus de travail relié au traitement des documents numériques, qui est limité par les technologies utilisées et donc moins flexible que pour les documents analogues. Les complexités sont dues à la fois à l'environnement technique utilisé par les créateurs des documents et aux limites imposées par la solution de préservation numérique des Archives. Ce texte examine aussi certaines façons dont les activités archivistiques de base peuvent différer des pratiques bien établies quand on les transpose dans un environnement numérique.

ABSTRACT Since 2008, the City of Vancouver Archives has been working to develop a digital archives. Much of the funding and accompanying urgency were linked to the City's obligation to preserve the records of the Organizing Committee for the 2010 Olympic and Paralympic Winter Games (VANOC). This paper describes some of the difficulties encountered in acquiring VANOC's digital records and the Archives' struggles to map the different tasks involved in acquisition and appraisal against the digital processing workflow, which is constrained by the technology used, and therefore less flexible than for analogue records. Complexities are due both to the records creator's technical environment as well as limitations imposed by the Archives' digital preservation solution. The paper also considers some of the ways in which these core archival activities may diverge from long-established practices when moving into the digital environment.

ARCHIVARIA 72 (Fall 2011): 93–122

Introduction

That archivists have been slow to embrace digital technology has been both a blessing and a curse. Over the past three decades, as digital systems supplanted analogue ones in most organizations, archivists eschewed technological solutions to operational problems, avoiding complications faced by other types of organizations. Many were correct in their intuition that digital recordkeeping was unstable; ultimately, however, its advantages outweighed its limitations and inspired widespread adoption by records creators. Like many archives, the initial response by the City of Vancouver Archives (CVA) to the public's demand for digital services was to begin digitizing select portions of its holdings. But archivists' slow entry into the digital records realm created a steep learning curve when building strategies for custodianship of *born-digital* archives became unavoidable. While digital preservation research in the archival community has substantially increased over the past several years, very little has been said about the acquisition and appraisal of digital records. The CVA's digital archives team[1] has wrestled with this problem for the past three years due to the imminent acquisition of a private-sector fonds with a large digital component, the Archives of the Vancouver Organizing Committee for the 2010 Olympic and Paralympic Winter Games (VANOC).

This paper discusses the Archives' experience to date in acquiring the records of VANOC, and the unique situations that the predominantly digital nature of the records has presented in our efforts to appraise them. While the anomalous nature of the VANOC acquisition may at first seem to render this article less meaningful to repositories responsible for anything other than Olympic or large-event archives, we believe that the scale and nature of this acquisition is representative of the impending deluge of digital records that looms over archivists ill-equipped to deal with it. This is particularly the case for private records, where consistency in recordkeeping practices among records creators is the exception rather than the rule. Furthermore, just as VANOC existed for a single purpose and then dissolved, some grassroots organizations have a similarly finite existence, as do human records creators. Moreover, the acquisition of the VANOC records was legally mandated, rendering them similar in character to public records and therefore relevant to public records repositories. Finally, all repositories faced with digital acquisitions will confront diversity of format and medium. Since the CVA regularly deals with both public- and private-sector acquisitions, we hope that sharing our experiences so far, including our successes, failures, and reasoned speculation, will be helpful to archivists facing similar situations.

1 The digital archives team is currently composed of the CVA manager, two digital archivists, and a digital conservator.

Background

The City of Vancouver Archives is a medium-sized[2] municipal archives that has been acquiring and preserving records documenting the history and culture of Vancouver since 1933. Staff have actively provided access to digital content since the late 1990s, mainly through a photograph digitization program and more recently through audio, film, video, and text digitization projects. Our digitization activities were oriented toward enhancing access to records rather than facilitating preservation. Over the past four years, we have increased the scope of efforts to include digital preservation and curation – specifically, the preservation of born-digital records. Two recent projects intensified the City's commitment to digital curation: one driven by the need to preserve public records being generated by VanDocs, the City's newly procured electronic records and document management system; the other driven by the City's obligation to preserve the records of VANOC. Over the past three years, the digital archives team has been leading the development of a program to acquire, manage, preserve, and provide access to digital records, descriptions, and derivatives.

To effectively carry out our responsibility for preserving City records – including electronic archives – the digital archives team developed requirements and a proof-of-concept prototype for building a reliable, long-term digital preservation environment for transfers from VanDocs. Staff were committed to creating a system based on international open standards and best practices, using open-source software wherever possible. Though the first transfer of digital records from VanDocs is not expected to occur until 2012 at the earliest, the first prototype was completed in consultation with City records management staff and consultants Artefactual Systems, Inc. in October 2009. In November 2009 the team aggressively continued development work, taking the prototype to the pilot stage in anticipation of the acquisition of the digital archival records of VANOC.

Consistent with the Canadian total archives tradition, the CVA's mandate is to acquire not only the records of the City government and its various boards and agencies, but also the records of private businesses, organizations, and individuals. VANOC is the largest and most noteworthy example to date of a private-sector acquisition, but many other records being generated and offered for donation by private citizens and organizations are increasingly digital in nature. Unlike other private-sector donations, the City of Vancouver, under the agreements for hosting the Games, is legally obligated to collect, organize, preserve, and maintain the records and other materials created or received by VANOC, and to provide continuing access to the parties to the agreements and

2 The CVA has ~7,000 linear metres of holdings, nine full-time staff, supplemented by a mix of auxiliary staff and volunteers, and an annual budget of approximately $1 million.

to the public.[3] The 2010 Winter Games Fonds is our first major born-digital acquisition, and represents a pilot project for the implementation of a more complex long-term digital archives system. Developing such a system would not have been possible without the foundation prototype built using funds from the VanDocs project, and that prototype could not have come to fruition as a pilot system without the funds allocated from Vancouver's Olympic Legacy Reserve Fund. This combination of fortunate circumstances has allowed the Archives to dedicate resources to the project that may not have been available to other institutions of similar size. In light of our fortuitous situation, we are making efforts to disseminate our work as much as possible so that it can benefit the profession at-large.

VANOC Acquisition Project Overview

VANOC was established on 30 September 2003, shortly after Vancouver won its bid for the 2010 Winter Games. The Archives contacted the VANOC librarian in charge of records management in 2004, but despite best efforts on both sides, there was no major movement on the acquisition until the City had entered into a consultative agreement with VANOC, and funding for the project was approved. In the summer of 2009, the Archives hired a temporary digital archivist, Courtney C. Mumma, to manage the project. The VANOC Acquisition Project began in earnest with Mumma conducting a functional and recordkeeping analysis in coordination with VANOC records management and library staff. This information gathering and analysis phase took a little over three months, but the donation agreement negotiations went on for nearly a year. All parties signed the Archival Materials Agreement[4] in November 2010. By then VANOC (which at its apex had a staff of more than 2,500 full-time employees and 25,000 volunteers) had all but disintegrated and the first accrual of nearly 200 boxes of analogue and over 25 terabytes of digital materials was already in CVA's custody, sealed and awaiting processing.

Anticipating the VANOC acquisition, the digital archives team and consultants from Artefactual Systems began work in 2009 toward developing the prototype into a pilot system to bring digital records donations into archival custody and control while at the same time maintaining their authenticity. The software development aspect of the project includes, but is not limited to,

3 There are three such agreements: the *Bid City Agreement*, the *Multiparty Agreement*, and the *Host City Contract*. The parties to the agreements are the City of Vancouver, the Canadian Olympic Committee (COC), the International Olympic Committee (IOC), and VANOC.

4 "Archival Materials Agreement" is the title of the legal agreement between the City of Vancouver, the IOC, the COC, and VANOC.

Archivematica,[5] ICA-AtoM,[6] and various digital forensics tools. The hardware set-up is constantly evolving, but is predominantly managed independently from the City's IT network. Policies and procedures are being drafted for public- and private-sector records, and recommendations are being made to other municipal projects that might have a stake in record dispositions. Additionally, digital archives team members have cultivated a wealth of digital preservation knowledge. By the end of the VANOC Acquisition Project in late 2011, the digital archives pilot system[7] will have acquired, stored, and prepared for access over 20TB of digital materials representative of the 2010 Winter Games.

The next (and ongoing) challenge is to conduct further appraisal during processing before these records can be preserved and made available. The first phase of the Digital Archives Project, funded within the larger VanDocs Project, allowed us to establish the feasibility of the chosen open-source approach; the VANOC Acquisition Project, which is the Olympic Legacy-funded portion of the Digital Archives Project, will result in a pilot digital archives system that processes and preserves records acquired from VanDocs and VANOC. These projects will yield systems and procedures that allow us to acquire records from various City systems and private-sector donors, and transform them into a preservable state. In the remaining year (2011) of the VANOC Acquisition Project, the team aims to develop a hybrid access system based on ICA-AtoM that will facilitate integrated access to the VANOC materials as well as to analogue and digital materials already in the CVA's holdings.

Information Gathering and Analysis

It was clear from the start that selecting records for acquisition from a large and complex organization like VANOC would be challenging. VANOC was established with a single mandate: to support and promote the development of sport in Canada by planning, organizing, financing, and staging the 2010 Olympic and Paralympic Winter Games [hereinafter the Games]. This was its highest priority, with recordkeeping being of importance only in as far as it supported this priority. VANOC's recordkeeping priorities were in some ways similar to those of grassroots organizations that work toward a particular outcome with little expectation of a need for institutional memory after that out-

5 Archivematica is a comprehensive digital preservation system that uses a micro-services design pattern to provide an integrated suite of free and open-source tools that allow users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. See Archivematica, http://archivematica.org/ (accessed on 28 July 2011).

6 ICA-AtoM is free, open-source, web-based archival description software that is based on International Council on Archives (ICA) standards. "AtoM" is the acronym for "Access to Memory." See ICA-AtoM, http://ica-atom.org/ (accessed on 7 March 2011).

7 In the CVA's view, "digital archives system" refers to the people, processes, and technology that make the management of digital archives possible.

come is achieved and the organization dissolves. VANOC, however, was keenly interested in leaving a documentary legacy and recognized that co-operation with the Archives would contribute to a richer cultural memory. Within the first three years of their founding, in fact, they hired records managers and librarians to establish some control over their already somewhat disorderly and rapidly accumulating files.

The first phase of the acquisition project was to observe and conduct an analysis of VANOC's functions, structure, and recordkeeping systems, working with records managers at VANOC to identify records and record groups for eventual transfer to the Archives. The digital archives team conducted a functional appraisal based on this observation and analysis in order to provide VANOC with a list of archival materials identified for transfer (identified as the Schedule of Materials[8] in the Archival Materials Agreement). The overall methodology was typical of a functions-based (rather than a content-based) approach to appraisal, and analyzed VANOC's business functions and their relative value with respect to institutional administrative structures, policies, programs and services, and the records created and managed to support them. Functional and recordkeeping analysis were fundamentally critical elements in this appraisal decision making; it was clear from the start, however, that further appraisal and selection would be necessary once the records were transferred to the Archives.

Ultimately, the team assessed the importance or significance of the context in which VANOC's records were created or maintained and their likelihood of containing content that might be of informational value to researchers in the future. Notwithstanding limited access to VANOC functional groups, the team sought to identify only those record groups that were essential to carrying out VANOC's mandate. To pinpoint those essential groups, Mumma intended to arrange interviews with individuals responsible for recordkeeping within select functions to find out which records they identified as essential to their activities. She worked closely with VANOC records managers, who arranged meetings with representatives from some of the creative departments, such as Brand and Creative Services[9] (BCS) staff, Editorial Services administrators, the iPhoto librarian, and Apple server (MacShare) administrator. Additionally, Mumma spoke with the records managers about Microsoft SharePoint use and with IT

8 The Schedule is a list of specific directories, record groups, and classifications broken down into sections according to the VANOC recordkeeping location; it provides a brief paragraph that introduces the methodology used to determine the records selected and parsed with appraisal methodology summaries in the tables used.

9 According to the VANOC business plan, the mandate of Brand and Creative Services (BCS) was to "develop, manage, and promote the Vancouver 2010 brand thereby providing value to sponsors, licensees and partners, and inspiring the public to support and participate in the Games. Develop and produce unique, inspiring Vancouver 2010 Games design and creative (mascots, advertising, videos, uniforms, etc.). Provide creative consulting and services for all VANOC functions."

staff regarding SharePoint and other technical systems used to create, manage, and store records. While observation would have been preferable to interviews, she was not allowed to observe staff recordkeeping processes due to VANOC resource constraints leading up to the Games.

Mumma completed on-site visits to VANOC and staff interviews in July and August 2009. In September 2009 VANOC shared the IOC's Transfer of Knowledge (TOK)[10] list with her, allowing for the drafting of an abbreviated Schedule of Materials since many independent records were accounted for in the TOK and could be excluded from other locations in the schedule. The digital archives team's objectives were in line with recognized appraisal methodologies, attempting to include only those materials that would document VANOC's principal policies and actions, and their impact on both the public and physical environment of the region to provide a research resource for our and future generations. With this central goal in mind, both the Schedule of Materials and the Archival Materials Agreement were carefully edited and amended multiple times with regard to transfer of rights, restrictions, and confidentiality issues during negotiations over the next year.

While the VANOC negotiations are an extreme case, any donation requires vetting prior to acquisition. As discussed later in this article, digital records require an even more thorough vetting than analogue materials because of the technical and administrative workflow that the CVA has implemented. Any repository will make decisions about its own technical environment and its limitations, and those decisions will directly impact the investigative methods prioritized during acquisitions. Further, those investigative methods will inevitably change relative to the donor's own technical environment, administrative processes, and access parameters.

Technological Context

In the digital environment, the mediating role of technology between records and users is inescapable; technology always affects how users are able to interpret and interact with records. To understand the relationship between VANOC and its records, it was necessary to conduct a detailed examination of the technological context in which the VANOC records were created. Although that knowledge is critical in understanding the relationship of any creator to its records, it is largely implied when considering paper-based recordkeeping systems. In the digital environment, this knowledge must be made explicit. More

10 The IOC required VANOC to provide it with two sets of materials, the Transfer of Knowledge (TOK) and the Video Transfer of Knowledge (VTOK); both sets represent what the IOC has identified as essential to understanding the way VANOC orchestrated the 2010 Winter Games. Following the completion of the games, VANOC Project and Information Management (PIM) was responsible for providing the IOC with the TOK and VTOK.

pragmatically, knowledge of the technological environments in which the records were created and maintained was critical to the CVA's ability to bring them under its control and submit them to preservation procedures. VANOC's records were amassed on such a diverse variety of media that the digital archives team found that the time allocated for the transfer of the materials to the CVA's custody was insufficient. Despite their preparations, what team members planned to accomplish in weeks instead took months. The rapidly changing technical capabilities of the Archives further complicated the issue, due in part to the agile software development methodology applied to the CVA's digital archives design.[11]

During the first few years after its founding, VANOC staff worked together using a Microsoft Windows shared drive structure. In 2006, VANOC records managers recognized the need for a more sophisticated collaboration and document management system. Their solution was to implement two SharePoint environments: an internal version for employees and an external version for volunteers. Most of VANOC's functional units began using the internal version to store records and collaborate. The majority of records on the shared drives were migrated to SharePoint; however, since the shared drive environment still existed and staff were not barred from using it, some units continued using it for the duration of VANOC's existence. Also, despite their best efforts, VANOC records management staff were not aware of every recordkeeping system being used since units frequently took it upon themselves to find immediate technological solutions to their unique operational problems. For example, staff within Brand and Creative Services (BCS) worked almost exclusively in an Apple environment. VANOC IT staff maintained a "MacShare" server for the BCS, but some of their projects were still managed and shared among other functional units using the Windows shared drive environment. Moreover, BCS staff filled the IT-supported MacShare storage so quickly with large audiovisual files that space soon ran out, leading them to store project files on detached storage devices and optical media. While this was a resourceful solution that worked for them at the time, BCS's ad hoc storage was not backed up or monitored by IT, or overseen by VANOC records management.

11 Agile software development describes a set of software development methodologies that emphasize collaborative work among self-organizing teams, continuously changing requirements, and short release cycles that allow for the continuous evaluation of project successes. Agile Alliance, "The Twelve Principles of Agile Software," http://www.agilealliance.org/the-alliance/the-agile-manifesto/the-twelve-principles-of-agile-software (accessed on 3 August 2011). The actual transfer of the VANOC materials took place at a time when the technical capabilities of the Archives were rapidly changing (for the better), because of improvements to Archivematica, the acquisition of new hardware, and an increasing body of experience among Archives staff. Paradoxically, improved capabilities on the Archives end sometimes complicated the planning and execution of record transfers, as improved capabilities constantly led to more and better options as to how to conduct the transfers.

In the months immediately following the Games, records from VANOC's diverse recordkeeping environments (as noted on the selections listed in the Schedule of Materials) were transferred to the Archives, long before the donation agreement was signed or any analogue records were transferred.[12] Even though the donation agreement had not been finalized, VANOC's records management staff, in the midst of negotiations, accepted our rationale that the digital records were at risk of degradation and loss. They agreed to transfer the media to CVA under the condition that they could be copied for back-up purposes, but not processed until the agreement was signed. While the initial plan was for CVA staff to travel to VANOC's offices and make our own copies of the records stored on the shared network storage, detached devices, and SharePoint environments, time pressures led to a change in plans. VANOC staff dwindled daily as entire departments shut down operations and cleared out of the headquarters building, which made it quite challenging for the records managers to keep track of the multitude of records and storage devices. In a few cases, the records managers resorted to chasing down digital materials that had been intentionally or unintentionally taken home by staff from departments no longer in existence. Eventually, all the detached hard drives were packaged and transferred to the Archives. The contents of the shared drive were copied to one of the detached storage devices, selected mainly because space allowed for such a transfer.

Transferring the SharePoint environment proved somewhat problematic. VANOC IT staff copied the selected SharePoint sites, databases, and configuration files to two separate external drives to give to the Archives. Once in the CVA's custody, municipal IT staff worked to reconstruct VANOC's SharePoint environment on their own Microsoft Office SharePoint server using the configuration documentation provided by VANOC. VANOC's IT staff had reconfigured and adjusted their SharePoint implementation countless times, not all of which were documented, so City IT staff could only approximate the emulation of the VANOC SharePoint environment. Further, it was difficult to discern how authentic the City's SharePoint server reconstruction was, since Mumma had very little opportunity at VANOC to examine the native environment. There was no way of knowing for certain what the most important components were, or whether the migration constituted an instance faithful to the original. Determining the most appropriate and effective way of extracting records and accompanying metadata from the reconstructed VANOC SharePoint environment remains an outstanding issue for this project, and could be the basis of an entire paper in itself.

Once in hand, the team set about copying the digital materials. The copying

12 The Vancouver 2010 Winter Games concluded on 21 March 2010, with the official closing of the Paralympic Games. The Archives received digital and analogue transfers intermittently between April and August 2010, while VANOC was shutting down its operations.

process marked the first stage of the VANOC digital materials entering the CVA's physical system. The CVA's physical Digital Archives system includes six computers configured for designated tasks, an array of support hardware, and media storage units. There are four desktop computers connected in a local area network (LAN), isolated from the City's main network, that are loaded with the Archivematica software package. One of these computers is used as a server, two are clients used by digital archivists to manage the ingest workflow, and one is used for duplication tasks. These computers (what the digital archives team calls "processing stations") operate on a Xubuntu/Linux operating system. There is also a Macintosh computer connected to the LAN that is used to view Mac-specific file formats. Additionally, there is a desktop computer loaded with the standard City Windows XP environment that is *not* connected to the LAN. It is connected to the City's main network, and is used to upload preservation copies to the storage environment. The current workflow calls for transferred digital records to be imaged (forensically copied, bit-by-bit) immediately upon receipt and the original media stored in a secure location, preferably off site. The copies must be compatible with the Linux processing environment.

There were issues in mounting and accessing the drives from the various VANOC systems as well as permission concerns created when moving from system to system. These were not insurmountable, but they caused problems. In one case, the presence of undetected hidden files on a journaled HFS+[13] hard drive led to those files being accidentally loaded onto one of the processing machines. The process returned errors because of difficulty recognizing the hidden files and applying the appropriate system code. Being new to Linux and decidedly *not* technicians, our team members discovered, largely through trial and error, that transfers from diverse file systems called for specialized copying procedures so that the files they contained were viable for analysis and processing.

VANOC's creative teams transferred media containing selected records (MacShare and other storage media deployed by self-sufficient staff) to the Archives on a DROBO[14] 4-bay redundancy array device (RAID-like) with 8TB capacity; eleven 1TB and 2TB external drives; an external drive containing SharePoint configuration data along with another containing VANOC's SharePoint sites; and hundreds of DVDs and CDs. The transferred media from creative teams alone contain over 17TB of files and data identified as likely having substantive archival value. The DROBO and external drives formatted for use on Apple systems[15] were copied immediately to external drives formatted to be

13 HFS+ is the primary file system used in Apple Inc.'s Macintosh computers (or other systems running Mac OS). It is also referred to as Mac OS Extended.

14 DROBO is a brand of storage arrays made by Data Robotics Inc.

15 The drives from Apple systems were all formatted HFS+ journaled. Journaling is a feature available for the Mac HFS+ file system; it helps protect against corruption after a system crash. See "Mac OS X: About File System Journaling," 22 July 2008,

more compatible with our Linux processing stations.[16] The optical media and drives formatted for Microsoft machines were copied and indexed on drives formatted for our Linux processing stations. During the Games (after discovering that VANOC's website was not managed directly by VANOC but by contractors), the digital archives team also used HTTrack website-copying software to periodically capture VANOC's website onto drives appropriate for Linux processing. While such a listing of copying procedures may seem tedious, it highlights a sampling of what the team experienced and for which it was ultimately unprepared.

Processing of the analogue transfers commenced upon signing of the donation agreement, months after the initial transfer of digital media. Additional digital media requiring immediate copying were discovered when we began unsealing these boxes and reviewed the box lists. Though most of these discoveries were DVDs, the boxes also contained hundreds of digital videotapes that had not been accounted for in the transfer documentation. We were unprepared for this, and have no players for viewing these tapes; therefore, we are currently investigating a copying strategy using a combination of on- and off-site resources.

Once the transferred records were properly backed-up, processing efforts could begin. The digital archives currently consists of three main components. The basic architecture is established by the Open Archival Information System (OAIS) functional model.[17] Archival Storage consists of 50TB[18] of hard disks residing on the City's network-attached storage (NAS). It was installed in the

http://support.apple.com/kb/ht2355 (accessed on 4 April 2011).

16 Linux has limited support for HFS+ journaled file systems. HFS+ journaled hard drives from VANOC were copied onto HFS+ non-journaled drives so that they could be processed using the Xubuntu/Linux processing stations.

17 Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, "Figure 4-1: OAIS Functional Entities," http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed on 7 March 2011) (also available as ISO 14721:2003). The authors infer that the reader has a basic familiarity with the terminology used to refer to the three information packages (as well as the core functional entities: Ingest, Data Management, Archival Storage, Access, Preservation Planning, and Administration) described in the model: Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP). In brief, the SIP is the records and accompanying metadata that is submitted to the archives. SIPs are transformed by the Ingest function into AIPs, which are records that the archives seeks to preserve. The transformation from SIP to AIP adds preservation metadata to the information package, and may migrate the content into alternative formats deemed more viable for long-term preservation. Descriptive information about the AIPs is managed by the Data Management function. This metadata is used by the archives to manage its holdings, and by researchers for discovery purposes. Interaction with researchers is mediated by the Access function, which allows researchers to search archival descriptions and receive access copies of records in the form of DIPs.

18 The digital archives team asked for 50TB because 25TB of digital material will likely at least double once normalized, since the plan is to store both the original and the normalized copies along with metadata and logs about the processing.

fall of 2010 and is expected to expand to 100TB in 2012. The Data Management and Access functions will be supplied through ICA-AtoM, which will be used to: control descriptive information about stored records; provide Internet access to the public to these descriptions; and distribute access copies of records. The CVA's initial use of ICA-AtoM is currently limited to the control of VANOC's records, as well as some other small fonds with large, born-digital components deemed suitable for initial testing. The CVA plans to migrate its existing archival descriptions into ICA-AtoM in 2012, which will result in a hybrid access system where researchers are able to search for analogue, digitized, and born-digital holdings using a single interface.

Within our digital archives system, the Ingest function (responsible for transforming SIPs into both AIPs and DIPs)[19] is accomplished by Archivematica, a collection of free and open-source software developed by Artefactual Systems in collaboration with the CVA and other organizations.[20] Archivematica accepts a SIP containing records and metadata packaged together, and uses a micro-services[21] approach to produce AIPs and DIPs, calling on features provided by a number of open-source tools that have been incorporated into the Archivematica package to perform tasks[22] such as:

- Validating integrity checksums;
- Detecting malware;
- Characterizing file formats;
- Extracting metadata;
- Migrating files to preferred preservation and access formats, according to configurable normalization policies;
- Compiling extracted metadata and Ingest metadata into METS[23]

19 The CVA's processes deviate somewhat from the OAIS model here. Strictly speaking, the Access function is responsible for transforming AIPs into DIPs in the OAIS model. The approach used in Archivematica is to pre-generate and cache the DIPs at the same time that AIPs are formed, making them available for retrieval immediately, rather than waiting until an access request is made to create the DIP.

20 Archivematica wiki, http://archivematica.org/wiki/ (accessed on 8 March 2011). Testing for the purpose of developing Archivematica software and developing the CVA workflow has involved several versions of the software, up to and including release 0.7 alpha (as of the time of writing). It is expected that by the time this article is published, Archivematica will be in version 0.8 beta or later, and be in use in a production environment at the CVA for preservation of VANOC and other records.

21 Micro-services refers to "an approach to digital curation based on devolving curation function into a set of independent, but interoperable, services that embody curation values and strategies." California Digital Library, "Curation Micro-services," http://www.cdlib.org/services/uc3/curation/ (accessed on 28 July 2011).

22 The tasks are listed in the order they occur during processing. A complete list of all micro-services can be found at http://archivematica.org/wiki/index.php?title=Micro-services (accessed on 8 March 2011).

23 Metadata Encoding and Transmission Standard (METS) is used as a wrapper for the metadata included in the AIP and DIP. The METS schema is available at http://www.loc.gov/standards/mets/ (accessed 18 July 2011).

containers;
- Creating an AIP (original files and submitted metadata, normalized files, extracted metadata, preservation metadata, etc.) for Archival Storage; and
- Creating a DIP (access copies, metadata) and uploading it to the Data Management and Access system (in this case, ICA-AtoM).

The Archivematica tool acts as a processing pipeline. It requires that the files submitted for processing conform to a defined structure that organizes the files and accompanying metadata into specific sub-directories. Each step in the process accepts a standardized input and hands off a standardized output to the next step, the end result being an AIP and a DIP that can be handed off to the Storage and Access systems respectively. Archivematica also includes a Web-based workflow management tool that allows users to follow the progress of SIPs and provides notification whether or not a service has been completed successfully. It also informs users when an intervention or approval is needed, including those interventions that require appraisal decisions (addressed later in this paper).

The final aspect of technological context that the digital archives team considers when conducting appraisal, and one that directly impacts system development, is policy regarding formats that are feasible for transfer to their own digital archives system. The decision to use open-source software, for instance, has significant consequences. Licensing requirements limit the software that can be included in the Archivematica package, in turn affecting how certain processes are performed. A prime example of this is the format normalization process. The preferred preservation format for word processing documents is PDF. Although the Archives will accept MS Word documents (.doc) in a SIP, the Archivematica software does not include the proprietary MS Word software. Migration from Word to PDF therefore involves an intermediary step; the Word file is opened using OpenOffice and converted into a PDF. While this produces acceptable results for Word documents with simple formatting, documents that have complex formatting are not always rendered faithfully by OpenOffice, producing a normalized preservation copy of questionable authenticity. It is likely that CVA archivists will have to go outside of Archivematica in order to achieve authentic migrations for some file formats.

Addressing the vast diversity of file formats that may exist in potential donations has been a major issue during development. For instance, the VANOC records include over thirty different formats from complex files created in Adobe Illustrator and Final Cut Pro to Microsoft Office documents and simple text files. It was decided at a very early stage *not* to require that donated records conform to a particular set of acceptable formats; the CVA's policy is to accept digital records in any format and attempt to preserve them to the best of its capabilities. This is true for both public- and private-sector records. Preservation efforts at the point of records creation are more feasible when dealing

with public records; as part of the City of Vancouver, Archives staff have some ability to influence decisions related to recordkeeping. The reality is, though, that decisions about recordkeeping must above all support the conduct of the business at hand. Because of this, the CVA does not prescribe a limited set of particular formats at the time of acquisition.

The alternatives to this policy would be either to decline records outright because of their file formats, or to ask the donor to normalize their records into preferred file formats prior to acquisition; both were considered unacceptable. In the first case, the Archives would be declining records that would otherwise be sought as part of the acquisition. Not only would those records not be preserved, but contextual links with records that were accepted would be destroyed. In the second case, there is the risk that the resulting file may be of poor quality, or that file metadata would be lost during the migration. The Archives can avoid these problems by accepting the files in their original format, extracting any relevant metadata *before* performing the normalization, and controlling the normalization process so that it meets its standards for quality.

If a donated digital record is in a format that the Archives has designated as an appropriate preservation format, no further action is required. If the identified format is not appropriate for preservation, digital archives team members will migrate the record to a normalized preservation version that is acceptable,[24] while at the same time keeping the original file in case assumptions about the viability of the original format turn out to be wrong. In cases where the file format cannot be identified, or the Archives does not have a preferred preservation format or a tool to migrate the record to a preservation format, team members will preserve the original file and look for opportunities in the future to migrate the file to a more acceptable format. Archivematica uses a set of configurable format normalization polices that determine, based on the source format, what the normalized format should be, the tool used to perform the format conversion, and the parameters for the conversion process.

File format decisions may be made at a repository policy level, as they were at the CVA, or on a case-by-case basis. Arguably, those decisions permit more flexibility than the hardware and software environments of the creator and custodial bodies, as well as technological contexts that affect an archives' willingness and ability to accept born-digital donations.

24 Archivematica, "Media Type Preservation Plans," http://archivematica.org/wiki/index.php?title=Media_type_preservation_plans/ (accessed on 9 March 2011).

Accommodating Appraisal in the System Development Methodology

When the CVA began work on the development of the digital archives in 2008, there were a number of open-source tools available to archivists; each facilitated an aspect of digital preservation.[25] At the same time, there were a number of standards documents available across a range of conceptual levels that sought to provide guidance for digital preservation activities. We developed a project methodology in collaboration with Artefactual Systems to produce use cases based on the OAIS standard, and then to develop detailed technical requirements that could be used to define a digital preservation system. Once initial requirements had been defined, we tested existing open-source tools to determine what they functionally contributed to the system and where they fit into the overall workflow. Programmers from Artefactual Systems built a framework around those tools to fill in any identified gaps and to manage their interaction. Requirements were developed, and processes tested and evaluated in rapid iteration using an agile development methodology. Lessons learned from each current iteration, as well as knowledge drawn from an expanded base of standards and external technical and professional knowledge, were used to develop more detailed requirements and locate appropriate tools to be incorporated into any subsequent iteration.

An obvious local resource for our project was the InterPARES 3 project at the University of British Columbia. Analysis conducted by graduate research assistants from the project identified early on that there were gaps between the workflows established from our OAIS-based use case scenarios and the corresponding parts of the InterPARES Chain of Preservation (COP) model.[26] The most significant gaps related to appraisal.[27] This analysis prompted us to consider more thoroughly how appraisal fit into the overall workflow, both within and outside Archivematica. Like all archival processes, appraisal is an iterative one that becomes progressively more refined as more information about the records and their context of creation becomes available. The amount of detail that can be applied to appraisal decisions is proportionate not only to the breadth,

25 Examples of such tools that have received considerable attention include: JHOVE, the JSTOR/Harvard Object Validation Environment, used for performing identification, validation, and characterization of digital objects; PRONOM, a file profiling tool built by the UK National Archives; and XENA, XML Electronic Normalising for Archives, developed by the National Archives of Australia for converting file formats to preservation formats.

26 InterPARES 2, "Chain of Preservation Model," http://www.interpares.org/ip2/ip2_model_display.cfm?model=cop (accessed on 9 March 2011).

27 InterPARES 3, "Case Study 16 – City of Vancouver Archives – Requirements Analysis for a Digital Archives System: Workshop 04 Action Item 07 – Draft Gap Analysis Report," August 2009, http://www.interpares.org/rws/display_file.cfm?doc=ip3_canada_cs16_wks04_action_07_v1-1.pdf (restricted access). Information about the CVA InterPARES 3 case study can be found at http://www.interpares.org/ip3/ip3_case_studies.cfm (accessed on 9 March 2011).

but also the depth of contextual knowledge about the records. While general knowledge of provenancial context may be sufficient to make broad appraisal decisions, lower-level information about the detailed structure of the fonds and of the records themselves is required to make more focused decisions. Acquiring this lower-level information can be problematic in the digital environment. In the paper environment, the archivist is able to interact with the records in an unmediated and unstructured fashion. However, in the digital environment, interaction with the records is mediated through technology and there is less flexibility in terms of when and how the archivist can interact with them. Based on the feedback we received from InterPARES, we identified essential appraisal tasks in the COP model and inserted them into our workflows. The workflows were tested using early versions of the Archivematica software and sample SIPs. Months of testing led us to identify three distinct stages of appraisal: Selection for Acquisition, Selection for Submission, and Selection for Preservation. As more information about the records becomes available at each of these stages, the ability to make informed appraisal decisions improves incrementally.

In its current alpha version, Archivematica's role as an Ingest tool limits the archivist's ability to interact with the records after they are submitted for processing. The workflow requirements of the processing pipeline model mean that there are few opportunities for the archivist to intervene in the process. After a SIP is submitted to Archivematica, intervention for the purpose of appraisal is limited to two opportunities. The timing of these opportunities became evident as the team refined requirements and use case scenarios; the team realized that some types of information would not be available until after certain processes had been completed. For example, if file format information does not exist or is incorrect at time of submission, it is necessary to identify the format before selecting software to view it. Information about the records generated during Ingest processing is used to inform appraisal decisions at these two intervention points. It is chiefly presented to the archivist as summary information through dashboard notifications[28] that indicate that a processing task has been completed, and that a log or metadata report is available for review in the corresponding Archivematica directory. The logs and metadata files generated by various services contain very detailed information. However, while the information contained in the reports may be readable by humans (i.e., it uses a text-based encoding scheme such as ASCII or Unicode, rather than a binary encoding scheme), it is not always readily comprehensible. Depending on the particular software that has generated the logs, there may be

28 The Archivematica "dashboard" is a web-based interface that allows users to monitor the status of SIPs as they pass through the Archivematica pipeline, identifying micro-services that are pending and need approval to continue, or that have been completed and have generated an exit code indicating success or failure. See http://archivematica.org/wiki/index.php?title=Dashboard (accessed 2 August 2011).

additional knowledge that the archivist must have in order to reliably interpret the information.

As we considered in increasing detail how our processing workflows should be structured in order to accommodate appraisal tasks in the digital environment, we started to think more carefully not only about *how* appraisal is conducted in the digital environment, but also whether or not assumptions about *why* archivists appraise hold true. Appraisal is recognized as a core archival function. It is through decisions about what records to acquire, and whose records to acquire, that we shape the image of the past and present that we hand off to the future. Barbara Craig provides a succinct description of the purpose of appraisal:

> The objective of appraisal should be to rank records based on the values we assign to them as evidence of the functions and activities whose historical profile the archive seeks to build or create. The aim of appraisal is to highlight or to isolate the *small portion* [emphasis added] that should be acquired as the best evidence for a particular view or views of the past, anticipating uses and needs.[29]

Archivists talk about the "small portion" that should be acquired for several reasons. The first is the assumption that it is not necessary to keep all of the records in order to gain an adequate understanding of the creator; some records are so devoid of value that they contribute little or nothing to research efforts. The assumption is actually quite a strong one: very few of a creator's records contain sufficient evidence of past activity to be of use to future researchers. There is, as well, a more pragmatic component to appraisal. Archivists select the "small portion" from the greater mass of the creator's records because they do not have the resources to keep everything. Appraisal is a means by which they can reduce the records of a creator not only to a selection that provides the best evidence of its activities, but to a portion that is manageable, given the resources of the preserving institution. As Terry Cook points out, "[t]he central dilemma for archivists is simply this: not all records having archival value can be kept."[30] Archives must allocate resources to: storing records; establishing physical and intellectual control over them before they enter storage; and providing access to them for researchers. These resources are finite and are frequently insufficient.

The tremendous depth and breadth of discussion about appraisal within archival literature is no doubt a reflection of the perceived importance of the activity within the profession. Despite the diversity of viewpoints within the literature, there is a common goal among them: to provide for professional grounding within a body of knowledge from which consistent appraisal

29 Barbara Craig, *Archival Appraisal: Theory and Practice* (Munich, 2004) p. 51.

30 Terry Cook, "'Many Are Called, but Few Are Chosen': Appraisal Guidelines for Sampling and Selecting Case Files," *Archivaria* 32 (Summer 1991) p. 26.

practices may be developed and applied, rather than selecting records on a mere whim. Like many other archival activities, appraisal is highly subjective. Decisions made by archivists about what should or should not be preserved are shaped by the biases and perspectives of those conducting the appraisal; being able to establish links in reverse, from the set of preserved records in the archives to the methodology used to construct that set and the theory that shaped the methodology, does not eliminate those biases. If properly documented, those links add transparency by exposing the process and the archivist's reasoning in order to permit a third party to understand how institutional, professional, and personal biases contributed to the formation of what remains of the records.

Regardless of the theory that guides it, the act of appraisal is essentially a pragmatic one that evaluates the records of a creator against a set of rules devised by theory and implemented by methodology in order to make a value judgement. The end result is a ranking of the records from least worthy to most worthy of preservation, with a divider that separates the records that will be preserved from those that will not. Where the divider is placed is independent of the theory used to devise the value ranking. Different approaches to appraisal may result in different value ranking within the same set of records, but regardless of the theory and methodology applied, the result is a qualitative assessment of the relative worthiness of preservation of the records appraised.

It is in the decision about where to place the divider that appraisal becomes predominantly a pragmatic exercise about "How much can archivists afford to keep?" rather than solely a theoretical exercise about "What documentary legacy do archivists want to pass on to the future?" The availability of institutional resources will be a significant consideration in deciding how much of the corpus of records from a creator can be preserved. While differences between the analogue and the digital environments may necessitate different methodological approaches to appraisal, the values that the theoretical basis of appraisal seeks to expose remain the same, independent of physical form. The fact that records exist in digital rather than traditional formats does not alter the values that archivists seek to identify. However, it does fundamentally change the level of resources available for preservation. The way that archivists interact with digital records, the activities and processes needed to establish control over them, and the environment in which they keep them are all very different than in the analogue world; the resource allocation model in the digital world is also different. Some of these things are more resource intensive than they were before, many are less. They will be different for each archival institution depending on their size, their ability to take advantage of economies of scale, their ability to leverage resources shared with a parent body, and their ability to overcome initial barriers to participation.

Stages of Appraisal in the Digital Archives System

Appraisal is an iterative process, progressing from general to specific. Using a micro-services model that follows from requirements-based workflows necessitates appraisal iterations that are more controlled than those for analogue records. The digital archives team had to determine and create models for which appraisal decisions needed to be made and when enough information about the records would be available so that they could be made. The OAIS model offers little in support of this goal.

All of the records identified on the Schedule of Materials for transfer at the end of the Games[31] are now in the custody of CVA; the Selection for Acquisition stage is complete. Requirements for the next two stages are being tested, and are still being planned and revised. The team hopes that its experiences during testing and its expectations based on the research and results thus far will be useful to other archivists building digital archives systems and/or beginning to acquire to digital records. The following sections describe more comprehensively the three stages we have identified for the appraisal of digital records using the CVA's system: 1) Selection for Acquisition; 2) Selection for Submission; and 3) Selection for Preservation. Wherever possible, examples from the VANOC acquisition illustrate the stages.

Stage One: Selection for Acquisition

Selection for Acquisition is akin to the kind of appraisal that an archivist would conduct on any donation offered to their archives: it sets the parameters for a subset of the whole of the donor's material to be set aside and transferred to the archives. Selection for Acquisition occurs outside of the processing environment, before the records are accessioned. In general, it is the common archival practice of gathering information about the records creator, the recordkeeping systems, and the records. For born-digital records, an understanding of the technological context is also necessary because it is connected with the ability to understand all three of these entities. While the archivist may seek to understand the technology at this point, the technology is not always easy to understand. In theory, everything about the technological context is knowable; in reality, the time and resources that can be devoted to acquiring this knowledge are limited. The complexity of the technological context has a bearing on how easy it is to discover information about it. Further, remote access or barriers to access, such as permissions and encryption, can stand in the way of the appraisal. How the

31 VANOC, which continues to exist as a legal entity, retained some records necessary to fulfill its continuing legal obligations. The Archives will continue to receive records from the rump corporation until it completely winds down in 2017.

technology has been implemented affects the archivist's ability to gather information about the organization and the recordkeeping system.

Analogue donations, depending on processing priorities and preservation issues, may linger in an archives' backlog for some time. While this may be less than desirable, the passive nature of analogue preservation means that records that were in a reasonable physical state when they were received will likely remain in that state for some time to come. The same cannot be said for digital acquisitions, which rely on constantly changing external technologies to remain accessible. Consequently, processing and preservation actions become more urgent for digital materials. In the case of the VANOC materials, known digital content (unlike digital materials buried within analogue boxes) was immediately imaged or copied because the digital archives team knew that there were no backups, and that some of the materials had been unused and untested for five years or more.[32] In our ideal scenario, digital media is imaged using a physical write blocker as soon as possible upon transfer of custody. The original media, or media transferred from the donor, is to be stored and the image or copy is to be used for further processing. Archivists can return to the original if necessary.

The known digital materials within the VANOC records took nearly three months to copy. The work was so laborious that the digital archivist performing the copying developed a repetitive stress injury! Imagine the situation if the digital portion of the acquisition had not been pared down prior to transfer. At one point during negotiations, in fact, VANOC staff had suggested that it might be easier to bundle all of the VANOC servers in plastic wrap and ship them to us. In that case, copying and/or imaging the server environment of an organization would have been difficult. Clearly, planning is required so that pre-Ingest procedures can be kept under control and within the bounds of the archives' resources.

Traditional functional appraisal assumes a high level of access to the creating organization. In the case of the VANOC analysis, Mumma was allowed only minimal access. During the summer of 2009, VANOC had little more than six months to finalize preparations for the Games. It took years of negotiations between VANOC and the Archives before an archivist was finally allowed to begin a functional and recordkeeping analysis from within the VANOC campus. Mumma was welcomed by records management staff eager for direction assessing which of the records they had amassed so far belonged in the Archives; however, like most of the functional groups within VANOC, the records managers were, by that time, entirely preoccupied with their final push toward

32 Ideally, all of the digital materials would have been forensically imaged (bit-by-bit copies); however, the digital archives team did not have the expertise or resources to do so at the time of transfer. Throughout the rest of the paper, images are referred to as copies.

the Games. Despite their best efforts, they were unable to arrange for her to meet with a representative from every functional group.

The assumption that the archivist has full access is not a problem unique to the functional appraisal of digital records: it can affect analogue appraisal as well. Since the majority of VANOC's records were digital, Mumma's ability to complete a comprehensive analysis was further complicated. More than just staff preoccupation got in the way of a comprehensive analysis: understanding one's own recordkeeping practices is very much a mystery to all but a few recordkeepers in digital recordkeeping environments. Indeed, to many employees at VANOC, the digital systems in which they stored their daily work were black boxes they did not need to understand beyond their ability to facilitate the immediate task at hand. Whatever magic happened behind the scenes to connect their workspace to a project manager in another functional group was left to the IT staff. IT staff, for their part, were concerned primarily about security, keeping the systems up and running, and maintenance of the servers. Though short interviews were conducted with IT staff, they could provide little beyond technological details about the hardware. In the end, we could only make educated guesses about the likelihood of the presence of records with archival value in known VANOC recordkeeping environments based on partial interviews, draft documentation, and the experiential knowledge of the records management staff at VANOC; when it came to preparing a Schedule of Materials for transfer, trusting the knowledge and recommendations of the records management staff was based on mutual respect and an extensive collaborative relationship.

Stage Two: Selection for Submission

The next appraisal stage, Selection for Submission, is the process of forming Submission Information Packages (SIPs). In its current configuration, the pilot system divides Selection for Submission between pre-Ingest (external to Archivematica) and Ingest (conducted within Archivematica) processes. Eventually, tools for pre-Ingest processes will be packaged within Archivematica. Presently, the digital archives team has completed only draft workflows and requirements. Regardless of system changes over time, the intended outcome of this stage of appraisal is the subset of records known as the SIP. All SIPs begin as transferred digital files on assorted media that must be transformed and supplemented to be compliant with the Ingest mechanism, Archivematica. This stage can involve some arrangement, minimal description, and further appraisal.

The proposed workflow for Selection for Submission begins when the transfer arrives at the Archives. Our goal is to begin with a forensic copy of the transferred media, then identify, extract, or crosswalk all metadata about the files on the copy; identify and segregate or "tag" password-protected files and confidential information (e.g., credit card information, names, phone numbers,

social insurance numbers, email addresses, etc.); create an accession record and/or an XML submission agreement (e.g., TAPER's schema)[33]; intellectually arrange files into series while retaining a record of their original order; and log all pre-Ingest processes and arrangement decisions.

In Archivematica, a micro-service called *appraiseForSubmission* allows the user to review the SIP to confirm that it complies with any submission agreements. The user can delete unwanted files at this point and a log of the deleted files is added to the information package.[34] In the current system, any SIP must consist of three folders: objects, metadata, and logs. Within Archivematica, the SIP is backed up, assigned a checksum, and its basic Dublin Core metadata is recorded before Ingest begins. In the current system, automated processing pauses here; the archivist manually reviews the SIP to verify that it contains the expected contents (this step could eventually be automated). The SIP container, aggregate metadata and logs, and records are examined at this point in order to confirm that they are what we think they are and what we expected based on any transfer documentation. This is also a decision point for reviewing what should be submitted for Ingest from what we have received. Anything the archivist decides is not worth establishing control over will be eliminated.

This stage is the first opportunity for the archivist to validate the appraisal analysis that guided the overall acquisition decisions. The archivist can confirm that the records are what he/she thought they were, and make decisions about what part of the received records the repository should commit to preserving (e.g., given the size of the transfer and the resources allocated, is selection necessary and worth the effort?). In principle, SIPs are submitted with the intent that they will be transformed into Archival Information Packages (AIPs) and preserved. While there is the recognition that information generated during Ingest may result in a decision to reject some or all of the SIP content, such cases should be viewed as exceptions.

This is where the pragmatic question about selection – how much of the materials transferred should be kept, processed, and stored – becomes relevant. A persistent mantra from the IT community is *storage is cheap, so why not keep everything?* Google's Gmail service is a highly visible manifestation of this attitude. When it first launched, the free Gmail service offered users 1GB of storage space, over one hundred times the storage provided at the time by rivals such as Hotmail and Yahoo.[35] In the seven years since its launch, the storage

33 TAPER (Tufts Accessioning Program for Electronic Records) at Tufts University, http://dca.tufts.edu/?pid=49&c=70 (accessed on 7 March 2011).

34 Archivematica Micro-services, http://archivematica.org/wiki/index.php?title=Micro-services#Archivematica_Micro-services (accessed on 9 March 2011).

35 Google press release, "Google Gets the Message, Launches Gmail," 4 April 2004, http://www.google.com/press/pressrel/gmail.html (accessed on 7 March 2011).

provided to each Gmail account has increased to over 7.5GB and is growing.[36] Gmail's free storage is so plentiful that it encourages its users never to delete anything. Anyone can buy a 1TB hard drive for $100 or less, contributing to the perception that digital storage is so plentiful and so cheap that it makes sense to keep everything. Of course, digital storage is not the same as digital preservation, but it is a substantial component of it. In December 2010, the CVA procured 50TB of storage for its digital archives system. The total cost was approximately $2,500/TB for a tier 2 storage solution, backed up at a second site, with appropriate levels of redundancy and network connectivity. While this was considerably more than the cost of the equivalent volume of storage for their home computer, it was well within the expected range, considering the additional requirements of the team. To put this in perspective, 1TB is equivalent to 48,000 8"x10" colour photo prints, 145,000 8"x10" black and white prints, or 100,000,000 pages of textual documents.[37] These would occupy 30, 90, and 10,000 linear metres of shelf space, respectively.[38] For that same $2,500, one could purchase 50 metres of shelving (just the shelving), 50 hours of labour (wages and benefits), or about one square metre of real estate in Vancouver. Storage is perhaps not as cheap as commonly believed, but it is still significantly cheaper than the physical infrastructure required to store equivalent volumes of paper records, even after taking into consideration that the amortization period for buildings are ten or more times greater than the amortization for computer hardware. Digital *preservation*, however, is much broader than mere storage.

Digital preservation requires systems to establish control over the records being preserved and to monitor them in the storage environment. There is considerable work involved at the time of acquisition to establish control over them and transform them into a state in which they are capable of being preserved by an archives. The other major variable cost that is dependant on the volume of material being preserved is the cost of labour associated with acquiring, appraising, and processing the records before they are sent to storage; automating these processes could help lower the per-unit costs for processing digital records. Consider arrangement and description. At higher levels of description, the time it takes to arrange and describe records is a function of the complexity of the records, not the volume of the records, and is independent of the form

36 According to the Gmail home page, http://www.gmail.com (accessed on 7 March 2011).

37 These equivalencies were estimated on the following basis: one 8"x10" print with a resolution of 300 pixels per inch is equivalent to a TIFF file of ~6.9MB (black and white) or ~20.6MB (colour). The size of a typical MS Word document depends on a number of things (e.g., the presence of graphics or special formatting), but typically ranges from 3-7kB per 8.5"x11" page; a generous upper limit of 10kB/page was used in this calculation.

38 Derived from conversion figures provided in Tennessee Archives Management Advisory, "Archival Containers: Tables Of Cubic-Foot Equivalents For Containers, Shelving, And Cabinetry Commonly Found In Archives," http://www.tennessee.gov/tsla/aps/tama/tama02containers.pdf (accessed on 4 May 2011).

of the records. However, at lower levels of description, the creation of file and item descriptions and inventories is not only easier and faster, it yields much richer information. Metadata is extracted from files as they pass through Ingest processes. That metadata travels with the access versions of the files and is used to automatically populate file- and item-level descriptions. What takes hours or days to perform in the paper environment takes mere seconds to accomplish in the digital environment, and is done to a greater level of detail. Of course, this is an idealization that will not frequently be realized in practice. Conclusions about the efficiency of automated processing are predicated on assumptions about the quality of the accompanying metadata. The quality of metadata accompanying the VANOC records was sometimes good, occasionally questionable, and often non-existent. Metadata quality is already a consideration when attributing value to the records because the presumption of authenticity is linked to the adequacy of record metadata.[39] It is also relevant for resource allocation reasons because of the resource-multiplying effect of successful automation: more work can be accomplished with the same or fewer resources, but only if the quality of the accompanying metadata is good enough to be reliably used as inputs to automated services.

In the CVA's case, it is becoming readily apparent that the per-unit costs of digital preservation are much lower than in the paper environment. There is no reason to expect this should be any different for institutions with a similar size and mandate. These costs should be even lower for larger archives that are better able to take advantage of economies of scale. If the above assumptions about differences between preservation costs for traditional and digital records are valid, and a lower per-unit cost for digital preservation can be realized, one of the consequences will be that archives can afford to preserve a much larger portion of digital records from any given creator using the same, or fewer, resources. The obvious follow-up question is: Just because archives can preserve more, does that mean they should?

Another goal of appraisal is to reduce the volume of records in order to increase the ability of researchers to discover records relevant to their enquiry by removing records of low perceived value, thereby increasing the usability of what remains. This is one of the differences between an organization's records and its archives.[40] While the need for this type of filtering to support usability

39 Terry Eastwood, Barbara Craig, Du Mei, Philip Eppard, Gigliola Fioravanti, Normand Fortier, Mark Giguere, Ken Hannigan, Peter Horsman, Agnes Jonker, Leon Stout, and Su-Shing Chen, "Appraisal Task Force Report," pp. 9–10 in *The InterPARES Project: The Long-Term Preservation of Authentic Electronic Records: The Findings of the InterPARES Project*, ed. Luciana Duranti (San Miniato, 2005), available online at http://www.interpares.org/book/index.cfm (accessed on 7 March 2011).

40 "As appraisal frees the registry from dead weight [*Ballast* in German], extracts the essential material of the organism and thereby enhances its clarity and useability, so is it, in the final analysis, a part of those activities which transform the registry into an 'archival body'."

may have been undisputed in the paper environment, the increasing availability and sophistication of tools that support discovery calls it into question in the digital environment. The threshold of what might be considered useable has also been changed by technology. Paper records that never crossed the archival threshold because they were deemed to have such a low information density that researchers would have neither the time nor the inclination to go through them may – in their digital form – be welcomed into the archives. This is not because they contain any more information than their paper counterparts. It is because they are more accessible and their content is in a more usable form. As discussed previously, automated metadata extraction allows the creation of much richer lower levels of description, creating the potential for improved discovery. Retrieval times for digital records are typically measured in fractions of seconds rather than minutes, hours, or even days; arbitrary numbers of records can be retrieved simultaneously, rather than restricted to one box at a time. Visual analysis software[41] can rapidly mine content from thousands of records and summarize the results in novel ways, in contrast to an archivist wading through paper documents one at a time, taken from box after box. A low value-to-volume ratio among a series of digital records does not have as great an impact on their overall usability as it would for paper records. Record series that may previously have been subject to sampling in order to retain a representative sample can be retained in much greater numbers – even in their entirety. After all, the most representative sample of a population is the entire population.

Again, much of the preceding discussion assumes the achievement of ideals in order to best take advantage of digital efficiencies. The VANOC case, however, is far from ideal. The VANOC recordkeeping environment was in a constant state of flux, from the time of VANOC's founding to its effective dissolution at the conclusion of the Games. Record metadata ranged from poor and inconsistent to non-existent, and the process of extracting and copying files for transfer altered it further. Records classifications were in a perpetual draft status and applied only at the top levels in a small portion of the recordkeeping

Adolfe Brenneke, *Archivkunde: ein Beitrag sur Theorie und Geschitche des europaischen Archiwesens,* ed. Wolfgang Leesch (Leipzig, 1953), p. 38. The translation here is by Rick Klumpenhouwer in his MAS thesis "Concepts of Value in the Archival Appraisal Literature: An Historical and Critical Analysis" (Master's thesis, University of British Columbia, 1988), pp. 46–47.

41 Visual analytics is the use of software tools with interactive, visual interfaces to address problems related to synthesizing information from large volumes of data that may exist in a multiplicity of formats. James J. Thomas and Kristin A. Cook, eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (Los Alamitos, 2005.) pp. 25–28, available at https://www.icts.uiowa.edu/confluence/display/ICTSit/IlluminatingThePath (accessed on 2 August 2011). An example of its use in an archival context can be seen in Stanford University's application of visual analytics to the emails of poet Robert Creeley, http://dhs.stanford.edu/visualization/robert-creeley-e-mail-correspodence-network/ (accessed on 7 March 2011).

systems. The final transfer of records consisted of a mix of hard drives, optical media, database exports, and website crawls. In some cases, storage media were discovered in boxes months after the original transfer of custody. This created a sizeable amount of up-front work for the digital archives team to inventory the various transfer media, assess the authenticity and reliability of records, and parse transferred records into SIPs for Ingest. The state of the VANOC records and the transfer process diminished some of the anticipated advantages of conducting appraisal in a digital environment.

At the CVA, we intend to use a combination of open-source archival and digital forensic analysis tools to form the SIP.[42] Digital forensics experts have finely honed the process of examining digital files while maintaining their integrity as evidence, but their tools are built for digital forensics experts and not archivists. In the spirit of any good open-source project, we have been seeking help from digital forensics experts and a handful of archivists and computer scientists with experience using digital forensics tools for archival purposes. While it may not be perfect, some version of the pre-Ingest module will be created before the end of the VANOC Acquisition Project so we can compile SIPs for Ingest.

Having based our initial pilot on the OAIS model, very little early work went into planning for the preparation of the SIP. In OAIS, the archivist is presumed to start with some preformed information package. In reality, though, born-digital records do not come in neat packages ready for Ingest. The digital portion of the VANOC records was no exception. As mentioned, shortly after the Games, VANOC transferred the following to the Archives: a DROBO 4-bay redundancy array device (RAID-like) with 8TB capacity; eleven 1TB and 2TB external drives; an external drive containing SharePoint configuration data along with another containing VANOC's SharePoint sites; and hundreds of DVDs and CDs. Several months later, when the CVA gained official custody of the records, the team unsealed over two hundred boxes and began processing the analogue materials. We found digital discs and videotapes totalling at least 1TB of data filed with the paper.

Based on the functional and recordkeeping analysis, it was clear that records belonging to any given series were spread over multiple media; however, as mentioned before, the first step in our transfer workflow is to immediately copy the transferred digital records. The CVA, in preparation for the transfer of the VANOC records (and uninformed about the nature of the transfer), pur-

42 For instance, the CVA is evaluating AFFLib (including fiwalk and bulk extractor), http://afflib.org/; Sleuthkit and Autopsy, http://www.sleuthkit.org/; the University of North Carolina's Curator's Workbench, http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/; ArchiveSpace's accession specification, http://archivesspace.org/; and the TAPER submission agreement templates, http://dca.tufts.edu/?pid=49&c=70 (all accessed on 8 March 2011).

chased numerous 1TB external drives to use for copying, imaging, backup, and processing. The copying process further obfuscated the physical arrangement of the transfer because files from the original transferred media had to be copied across multiple drives or grouped together artificially onto one. For instance, the 2TB drives containing daily Torch Relay photograph files were sorted in folders containing every image from that date and location. It was impossible to simply copy the drives to two 1TB drives since the folders would have had to be taken out of sequence to total just under 1TB per drive. Instead we had to copy, for example, the first 680GB onto one drive, the next 515GB onto another, and the remaining 542GB onto the last drive. In other cases, the 1TB drive might contain copies of every DVD in a case, a 30GB series of records from another drive, and several gigabytes of data from another series on yet another drive. To date, we have used nearly fifty 1TB drives to copy only 20TB of digital records.

One of the assumptions we made in designing the Archivematica workflow and the information package designs was that there should be a one-to-one correspondence between a SIP and an AIP, that both of these packages should correspond to an arrangement unit within the acquisition. To establish some order and compile SIPs from the nearly fifty drives, some arrangement into archival units or parts of units is necessary. That arrangement must be completed while adhering to strict standards that protect the records from tampering while revealing information about their origin, configuration, and content. The tools that allow for compilation of the SIP will inevitably allow for further selection or culling based on the first close look at the transfer's content.

Stage Three: Selection for Preservation

Finally, Selection for Preservation – handled by Archivematica's "appraiseFor Preservation" micro-service – allows the user to appraise the contents of the SIP and delete unwanted files. A log of the deleted files is added to the information package.[43] At this point in the current system, the archivist decides what components of the SIP to keep before it becomes the AIP. This review is necessary since there can sometimes be insufficient information present at Selection for Submission to make a decision. Processing – in particular format identification and validation, and metadata extraction – provides more information about the files. For example, the original file may have been a format that could not be viewed, but at this point there may be a normalized copy that can be viewed.

This stage of appraisal is intended to result in the formation of an AIP for Archival Storage. The AIP is entirely independent of the processing software and is designed to interoperate with other digital repositories. The AIP consists

43 Archivematica Micro-services, http://archivematica.org/wiki/index.php?title=Micro-services#Archivematica_Micro-services (accessed on 9 March 2011).

of the original files and their normalized preservation copies, packaged together in accordance with the Library of Congress BagIt specification.[44] Review of the AIP at this point includes manually checking whether the previous incremental assessments were correct. Analogue materials are stored and processed simultaneously, so there is no need to check that the storage package meets the expectations of the processing results. Digital materials, on the other hand, cannot be checked until certain processes have been performed.

At this point, the digital archives team's experience with Selection for Preservation has been limited to test scenarios executed using sample sets of VANOC records. The practice so far has been to accept the AIP for Archival Storage in all cases. One example of a case where the team might reject an AIP is if format identification reveals that hundreds of files that appeared different from each other are actually duplicates. The VANOC acquisition, with its many formats and diverse quirks, will likely present cases not yet anticipated. We look forward to those learning opportunities.

Conclusion

The acquisition and continuing appraisal of the VANOC records has been a challenge. In many ways it represents a worst-case scenario: a large organization – with a rapidly evolving organizational structure and a wide diversity of recordkeeping technology – that existed for a limited time and therefore had little need for organizational memory after the close of the Games. The difficult VANOC acquisition is a good test for our digital archives program, and contributes much toward our understanding of how the digital environment affects core archival activities.

The initial tasks in the appraisal of large digital fonds are fundamentally the same as for the appraisal of any other large body of records. This should come as no real surprise because, at the highest level, it is really the records' creator that is being appraised, not the records. Differences emerge as the appraisal becomes increasingly detailed. The first difference to emerge is the need to understand the technological context of the recordkeeping system. In any appraisal, it is important to understand the procedural contexts that relate workflows to the structure of the recordkeeping system. In the digital environment, however, features and limitations of the technologies used for the creation, keeping, and use of records play a greater role in shaping how people interact with records. These technological capabilities therefore need to be taken into account when conducting appraisal, even at a high level.

Lower-level appraisal activities, such as confirming or refuting higher-level

44 Library of Congress BagIt specification, http://www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf (accessed on 9 March 2011).

appraisal assumptions, purposive or random sampling of records, and identification of specific content that should or should not be retained, necessitate the viewing of records and the record-related metadata. When designing workflows, archivists must be aware of existing dependencies that must be accounted for within the workflow. For example, in cases where the file format is unknown, selection of an appropriate file viewer depends on knowing the file format, and in some cases the format version, or the audio or video codec used to encode the file. This means that format characterization and metadata extraction must precede the selection of an appropriate viewer. Furthermore, local policies – such as the CVA's policy of accepting original files from the records creator rather than normalized copies – can affect an institution's ability to appraise records early on. Any system – human or technological – will inherently impose constraints on the archivists' regular workflow.

Any decision to sample records, either randomly or purposively, in order to reduce their bulk should be considered within the context of the cost of the work required to devise and implement sampling methodologies. Will the cost of the up-front work involved in sampling be recouped over time through lower storage costs? Archivists should keep in mind as well that sampling in order to enhance the discoverability and usability of the remaining records may be less desirable in the digital environment, because of the improved capability to index digital records and harvest information from them. In our experience, there are three factors that should be taken into consideration: 1) the consistency of the recordkeeping structure (i.e., the tendency for recordkeeping units of a similar nature to be located at the same depth within a directory structure, or to have the same level of metadata applied to them); 2) the homogeneity of file formats; and 3) the absolute volume of the records. Records with consistent recordkeeping structures and homogeneous file formats are easier to analyze, and it is much easier to apply selection methodologies to records with these characteristics. Very large bodies of records will see the greatest storage cost-benefits from their size being reduced.

The Torch Relay video footage (many terabytes of video documenting every day of the event) met all of these criteria; the appraisal decision, however, was not to perform selection on these records. The scale of the Torch Relay, the role it had in linking the Games to communities across the country, and the amount of planning required from VANOC to conduct the event led us to conclude that it would be appropriate to preserve all of the footage. This serves as a reminder that technological aspects of the records may influence appraisal decisions, but that the content and context of the records are still the primary drivers of appraisal decisions.

The VANOC project is a work in progress and there are still many unknowns. This paper is our attempt to describe some of the problems we encountered in our first major digital acquisition in the hopes of providing examples for other institutions about what to expect as they extend their acquisitions into the digital

realm. A complete analysis of the costs associated with acquiring and preserving the VANOC records will not be feasible until after the project is complete. We believe that, regardless of whether or not an analysis confirms our assumptions about the lower per-unit cost of digital preservation, it would be a valuable contribution to the community as a yardstick against which similar institutions can measure costs associated with digital preservation. The unique nature of the Olympics as an event, and of VANOC as an organization, would seem at first to mark this particular acquisition as an outlier, rather than a representative example of a primarily digital fonds; private fonds, by their very nature, are often atypical. Although digital preservation literature asserts that consideration of digital preservation should begin at the point of records creation, the reality is that archives will seldom have opportunities to influence the recordkeeping practices of organizations that they themselves are not a part of. During the six years that we had regular contact with VANOC, we were only able to influence their practices in a very minor way; we were, however, able to exert more influence over VANOC than we could reasonably expect with any other private donor because of their recordkeeping staff's willingness to collaborate with us.

When considering digital preservation, Canadian archives with a total archives mandate should hope for the best, but expect the worst. The worst, however, may not be as bad as anticipated. There is a growing network of digital preservation expertise that archivists can tap into, as well as a growing body of tools to facilitate this work. Without a doubt we have made (and will continue to make) mistakes in the course of this project; but building expertise is as much about mistakes as it is successes. Despite the many wrinkles and flaws present in the VANOC project, we are confident that the end product will be a body of records that will be preserved into the future, accessible and usable by researchers, and document an important event in the history of the city and the nation. We encourage other archives to share their own mistakes (hopefully different ones than the ones we have made), so that we can all learn from each other and build better solutions together.