DALHOUSIE UNIVERSITY
Dalhousie Libraries

DIGITAL ARCHIVES COLLECTION ASSESSMENT

CREIGHTON BARRETT AND DOMENIC ROSATI

January 2017

Dalhousie University, Communications and Marketing

DALHOUSIE LIBRARIES | University Archives
5th Floor, Killam Memorial Library
6225 University Avenue | PO Box 15000 | Halifax, NS B3H 4R2 Canada
902.494.4089 | Fax: 902.494.2062 | archives@dal.ca | http://dal.ca/archives

Table of Contents

1.0 Introduction 3
2.0 Snapshot of results 3
3.0 Archives Permanent Collection 3
4.0 Methodology 4
5.0 Search strategy 4
6.0 Physical separation of digital media carriers 6
7.0 Data collection 7
8.0 Action plan 8
Action #1 – Establish processing priorities for digital archives backlog 8
Action #2 – Update accessioning procedures 9
Action #3 – Develop digital forensics workflows and procedure manual 9
Action #4 – Establish procedures and infrastructure to support the monetary appraisal of digital archival material 10
Action #5 – Establish technical specifications for archiving born-digital audio and video 10
Appendix A – Register of digital media carriers 11
Appendix B – List of digital media carrier formats in the Archives Permanent Collection 13
Appendix C – Scripts and tools used in search strategy 14
Regular expression (REGEX) and AstroGREP 14
XQuery script and legacy EAD XML files 14
AtoM EAD and DC harvester JavaScript + JSON formatter 15
Appendix D – Mail merge separation sheet template 16

1.0 Introduction

In May 2016, the Dalhousie University Archives (DUA) initiated an assessment of the Archives Permanent Collection for born-digital archival material stored on digital media carriers. The collection assessment was conducted by Domenic Rosati, Archives Student Intern, with guidance from Creighton Barrett, Digital Archivist.

The goals of this collection assessment were to:

- locate and physically separate digital media carriers in the Archives Permanent Collection; and
- produce an inventory of separated digital media carriers

The collection assessment was initiated to help DUA establish processing priorities for its backlog of digital media and to inform the overall development of DUA's digital archives program.

The collection assessment was finished in October 2016. This report outlines the assessment methodology; provides a summary of the data; and establishes a preliminary plan for how to act on the assessment findings.

2.0 Snapshot of results

- Number of digital media carriers: **2,914**
- Most commonly identified digital media carrier: **3.5 inch floppy disk**
- Approximate total digital storage requirements: **325 terabytes***
- Fonds with largest number of digital media carriers: **imX Communications fonds, Centre for Art Tapes fonds, Joan Cummings fonds, Solar Audio Recording Studio Collection, Marine Affairs Program fonds, and Neptune Theatre fonds**

*Actual storage requirements could vary considerably depending on video reformatting specifications and digital preservation file format policies. It is also believed that many of the DVCAM tapes identified in this collection assessment are actually MiniDV tapes recorded using the HDV encoding specification. Further assessment of digital video and audio formats would produce a more accurate estimate of digital storage requirements.

See Section 7.0 for more information about the collection data gathered through this assessment.

3.0 Archives Permanent Collection

DUA acquires administrative records of Dalhousie University and materials related to Nova Scotia visual and performing arts, literature, labour, medicine, business, community organizations, and other thematic areas that support learning and teaching at Dalhousie University. The Archives Permanent Collection consists of approximately seven kilometres of textual records, graphic material, cartographic material, technical drawings, moving images, and sound recordings. DUA also maintains approximately 85 TB of born-digital archival material and digital material produced through digitization projects. Some fonds and collections have detailed finding aids, but much of DUA's holdings are uncatalogued. Finding aids are published in the Archives Catalogue and Online Collections.

4.0 Methodology

The digital archives collection assessment was based largely on Ricky Erway's 2012 research report for OCLC, *You've got to walk before you can run: first steps for managing born-digital content received on physical media*. The report outlines basic steps for surveying and inventorying digital media carriers in an institution's current holdings:

1. Locate existing holdings.

2. Count and describe all identified digital media carriers:

   a. Gather relevant information about collection.
   b. Photograph the digital media carrier.
   c. Remove the digital media carrier.
   d. Assign new unique reference code to each item.
   e. Record information about each item (e.g., hardware/operating systems, maximum storage capacity, etc.).
   f. Add summary information of the digital media carriers to any existing accession record, collection-level record, or finding aid.

3. Prioritize collections for further treatment.

4. Repeat steps every time new media is received.

Between 2013 and 2015, sixty-one institutions throughout the United States participated in the "Jump in Initiative," a program organized by the Society of American Archivists' Manuscript Repositories Section that encouraged members to conduct collections surveys according to the basic process outlined in Erway's research report. DUA reviewed reports and inventories from the "Jump in Initiative" and then designed and implemented a search strategy for systematically locating digital media carriers, and workflows for physically separating and inventorying the carriers.

5.0 Search strategy

The general goal of the search strategy was to identify collections with material created after 1975 and then identify individual digital media carriers within those collections. In practice, the hierarchical nature of archival description and the redundant nature of available data sources meant that each search produced both collection level and file- or item-level metadata that required manual sorting and deduplication into separate "unprocessed" collection-level and item-level inventories. These inventories helped guide the physical separation process and were further refined into a "register of digital media carriers" (see Appendix A).

To improve redundancy in search results and increase the overall reliability of the item-level inventory data, the search strategy utilized several data sources and search methods (see Appendix C for more details):

Data source	Search method	Notes
Legacy Encoded Archival Description (EAD) XML files on Archives' shared drive	Use Domenic Rosati's XQuery script to collect collection-level metadata from legacy EAD files.	Produced a tab-delimited text file. This search method supported the initial goal of identifying collections with material created after 1975. EAD XML was current as of January 2015.
Archives Catalogue and Online Collections	Use Domenic Rosati's AtoM EAD and DC Harvester JavaScript to harvest EAD and Dublin Core metadata from AtoM.	Produced a JSON inventory of all items in search query. JSON was converted to CSV using JSON Formatter.
Archives staff "data staging" catalogue	Use Domenic Rosati's AtoM EAD and DC Harvester JavaScript to harvest EAD and Dublin Core metadata from AtoM.	Produced a JSON inventory of all items in search query. JSON was converted to CSV using JSON Formatter.
Electronic "case files" on Archives' shared drive (includes inventories, appraisal reports, monetary appraisal reports, deeds of gift, and other documentation about each fonds or collection)	Use AstroGREP and a regular expression (Regex) to perform keyword searches of the case files, collection inventories, and other documents on Archives' shared drive	Regex was tested using an online regex validation too at www.regex101.com
MySQL dump of Archivists' Toolkit database	Use regular expression to perform keyword searches.	Database includes accession records and finding aids.

Data sources were searched using the following list of keywords:

1. floppy
2. floppies
3. disk
4. disks
5. disc
6. discs
7. diskette
8. diskettes
9. cd
10. cds
11. cd-rom
12. cd-roms
13. dvd
14. digital
15. usb drive
16. hard drive
17. flash drive
18. email
19. emails
20. electronic message
21. electronic messages

6.0 Physical separation of digital media carriers

Digital media carriers were physically separated from the analog material. A new series of “Born Digital Boxes” was established so that separated digital media carriers could be assigned new, temporary unique reference codes and physical locations.

In general, the following procedures were used to physically separate digital media carriers:

1. Locate digital media carrier in archives storage.
2. Physically remove digital media carrier from the analog file folder.
3. Photograph digital media carrier (*item-level photography was abandoned part way through the physical separation workflow).*
4. Assign new, temporary unique reference code to digital media carrier. Temporary reference codes must include a “BD Box” number and an item number.
5. Create a detailed record of the digital media carrier in the **register of digital media carriers** (see Section 7.0 for more details).
6. Place digital media carrier in polypropylene sleeve unless carrier has hard-plastic container (e.g., MiniDV, DVCAM, DVD case).
7. Print temporary unique reference code on mailing address label and affix label to polypropylene sleeve or container.
8. Mail merge select item-level data from the r*egister of digital media carriers* into separation sheet template. Print sheets in duplicate (See Appendix D for mail merge separation sheet template).
9. Cut separation sheets as necessary.
10. Insert one separation sheet into analog file folder.
11. Insert one separation sheet into polypropylene sleeve.

The order of steps varied depending on the number and physical organization of digital media carriers within a fonds or collection. The collection assessment resulted in 2,914 separated digital media carriers that are now stored in 23 Hollinger boxes.

7.0 Data collection

The collection assessment produced a *register of digital media carriers* and a *list of digital media carrier formats* in the Archives Permanent Collection. See Appendices A and B for more information.

The collection assessment found that digital media carriers are predominately found in fonds and collections that document theatre, music, film, and other artistic practices:

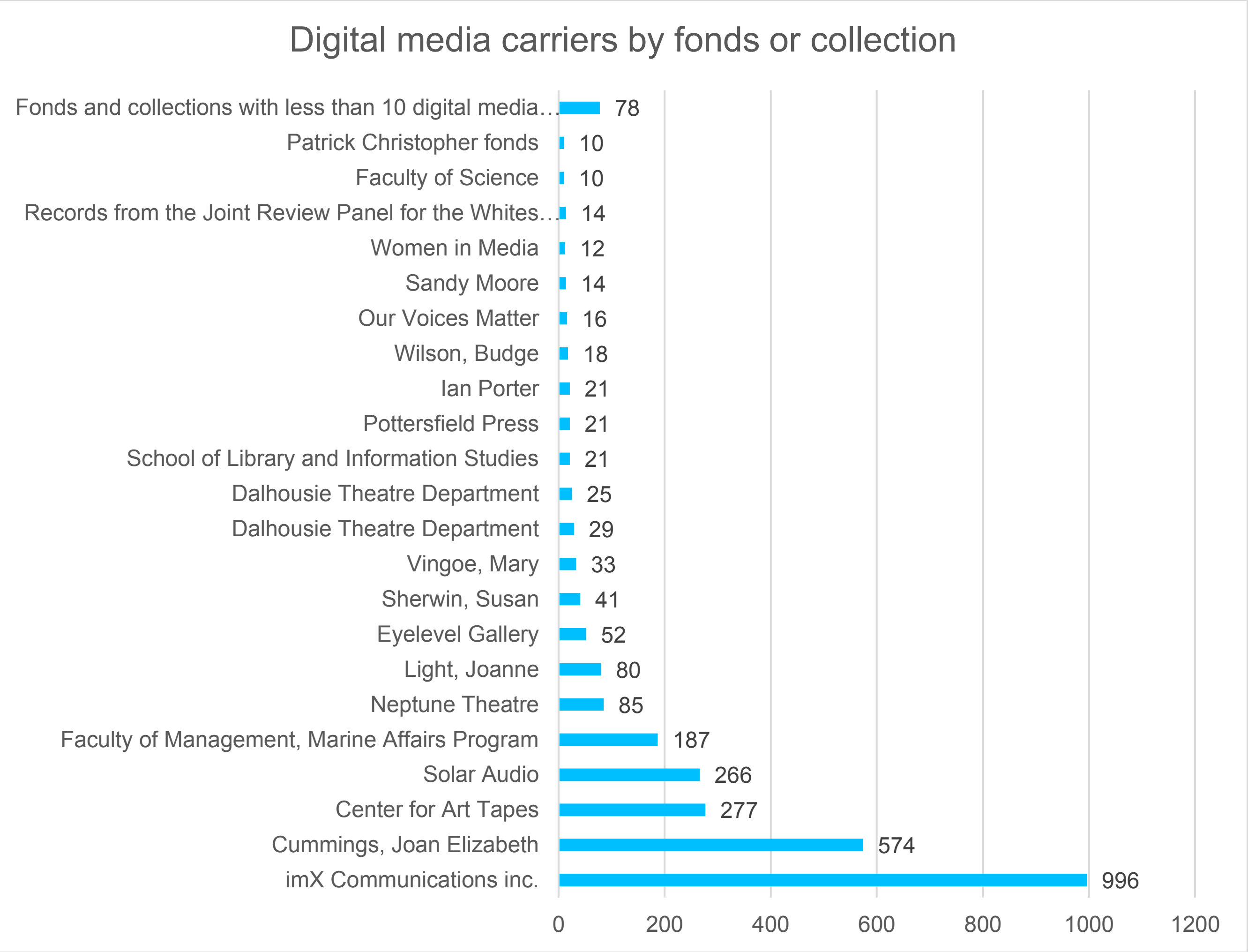

Figure 1 – Distribution of digital media carriers by fonds or collection

The collection assessment found that 3.5 inch floppy disks are the most predominate digital media carrier format in the Archives Permanent Collection (978 items) followed by DVCAM videocassettes (481 items):

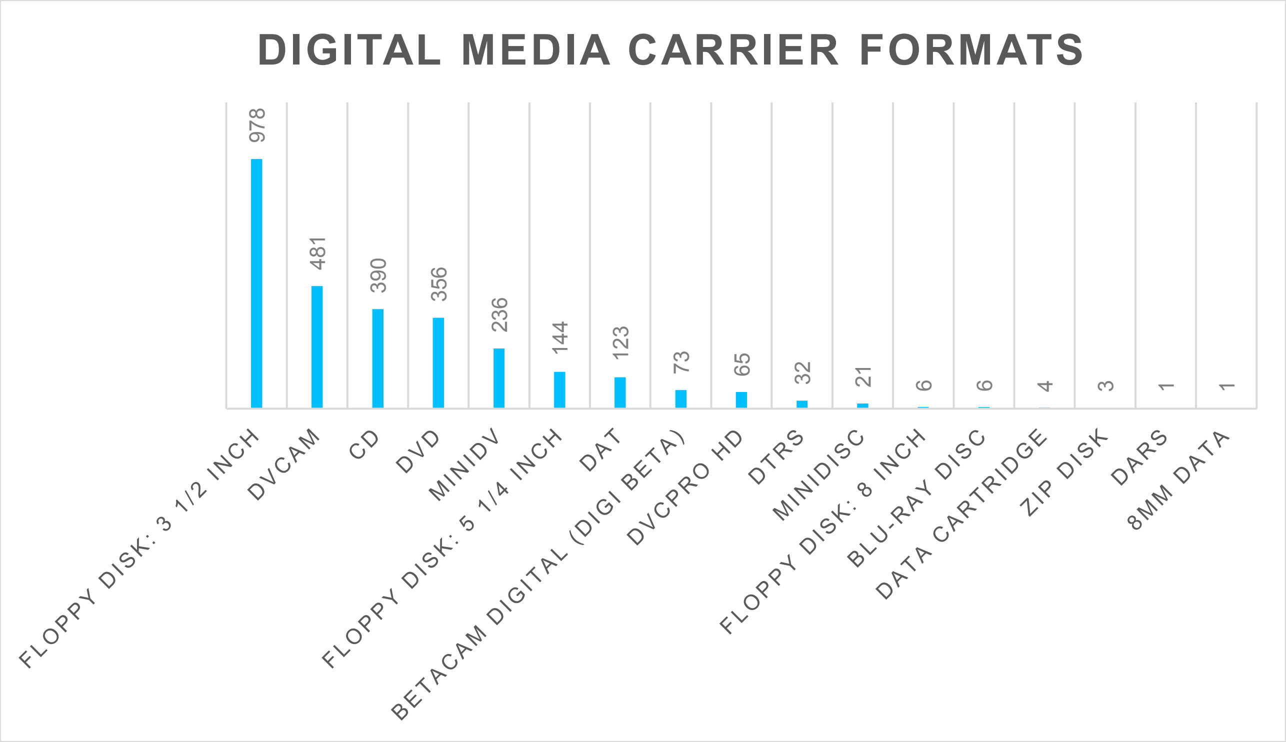

Figure 2 - Distribution of digital media carriers by format

The assessment highlighted a number of inconsistencies with DUA's catalogue records for moving image formats, particularly digital video tapes (DVCAM, MiniDV, and DVCPRO). For example, it is believed that many of the 481 DVCAM tapes identified in the assessment are likely MiniDV tapes recorded in HDV mode. The distribution of digital media carriers is based on the existing descriptive metadata and the collection assessment did not allow for the correction of metadata with incorrect physical description information.

8.0 Action plan

The digital archives collection assessment has shown a remarkable diversity of born-digital material and digital media carriers in the Archives Permanent Collection that are currently at risk of data decay. It is critical that the Archives continue to prioritize digital forensics, reformatting, and digital preservation as key areas for growth and development.

The results of the collection assessment have helped establish a five-step preliminary action plan:

Action #1 – Establish processing priorities for digital archives backlog
Dalhousie University Archives should establish priorities for physically processing items inventoried in the register of digital media carriers. Processing priorities should be established with the following considerations:

1. Macro-appraisal and functional analysis of creator(s).

2. University Archives' acquisitions policy and selection criteria.

3. Age and/or condition of the obsolete digital media carriers.

4. Learning, teaching, and research activities that could be supported by the digital archival material.

5. Sustainability of the digital material, including storage requirements and technical requirements for long-term preservation of the digital archival material.

6. Availability of digital content in analog form or in other collections or institutions.

7. Processing status of the analogue material in the fonds or collection.

Action #2 – Update accessioning procedures

Dalhousie University Archives should update its procedures for accessioning. It is critical that the Archives continue to maintain the register of digital media carriers and end the problem of storing digital media carriers amongst unprocessed and processed collections. Accessioning procedures should include:

1. Instructions on basic accessioning of hybrid acquisitions.

2. Workflows for physically separating digital media carriers from analogue material and inventorying items in the register of digital media carriers.

3. Instructions on receiving and accessioning university records through network file transfers.

Action #3 – Develop digital forensics workflows and procedure manual

Dalhousie University Archives should develop workflows and detailed procedures for creating disk images and migrating archival material from obsolete digital media carriers. The procedures should include:

1. Guidance on when digital media carriers should be imaged (e.g., USB, CD) and when they should not be imaged (e.g., network file transfers, digital audio tapes).

2. Instructions on how to use peripheral hardware components and all software applications found on the Archives' digital forensics workstation.

3. Workflows that allow archival appraisal to occur before, during, and after the disk imaging process.

4. Instructions on how to store and analyze disk images on the Archives' digital forensics workstation.

Workflows and procedures can be adapted from a 2013 OCLC report titled *Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-house*. Dublin, Ohio: OCLC Research.

Action #4 – Establish procedures and infrastructure to support the monetary appraisal of digital archival material

The collection assessment highlighted a growing trend of private acquisitions of digital archival material that require a monetary appraisal. Dalhousie University Archives should develop appropriate procedures and infrastructure to support the monetary appraisal of digital archival material. Action items include:

1. Develop procedures for preparing digital archival material for monetary appraisal, including detailed instructions on all software applications required to extract and collect metadata.

2. Develop space and technical infrastructure to support the viewing of digital archival material.

Action #5 – Establish technical specifications for archiving born-digital audio and video

A significant number of digital media carriers contain digital audio and video files. It is difficult to estimate total storage requirements for these materials. Dalhousie University Archives should establish technical specifications for archiving born-digital audio and video files. Specifications should include:

1. Indication of preferred codecs, file formats, resolution, bit-depth, sampling, and other characteristics of preservation masters and access copies for digital audio and video files.

2. Guidelines for transferring digital audio and video from digital media carriers.

3. Guidelines for editing and providing access to digital audio and video.

4. Guidelines for arranging and describing digital audio and video.

Appendix A – Register of digital media carriers

The register of digital media carriers includes the following metadata elements:

Metadata element	Description	Notes
New box number	The number assigned to the “born digital box” in which the digital media carrier was placed after physical separation.	A new series of “born digital boxes” was established for the collection assessment. The new box number forms part of the temporary unique identifier for the separated digital media carrier.
New item number	The item number assigned to the digital media carrier as it was placed into a new “born digital box.”	The new item number forms part of the temporary unique identifier for the separated digital media carrier.
Collection name	The name of the fonds or collection from which the digital media carrier was physically separated.	
Collection number	The unique identifier for the fonds or collection from which the digital media carrier was physically separated.	
Box number	The number assigned to the box in which the digital media carrier was placed prior to physical separation.	
Folder number	The number assigned to the folder in which the digital media carrier was placed prior to physical separation.	If applicable.
Item number	The number assigned to the digital media carrier prior to physical separation.	If applicable.
Title	A brief description of the digital media carrier.	Titles were modified and/or supplied as necessary.
Extent	Number of digital media carriers described in the register record.	For the collection assessment, each register record describes one digital media carrier. Future practice will allow for multiple carriers of the same format to be described in a single register record.

Metadata element	Description	Notes
Format	Description of the format of the digital media carrier	Terms pulled from the list of digital media carrier formats.
Medium	The general type of digital media carrier.	Automatically populated based on format selected from the list of digital media carrier formats.
Dimensions	Measurement of the physical size of the digital media carrier.	Automatically populated based on format selected from the list of digital media carrier formats.
Condition	Comments on the physical condition of the digital media carrier.	Often “Unknown” but also used to indicate scratches on optical disc
Maximum storage size	Indication of the largest amount of data that can typically be stored on the digital media carrier.	Automatically populated based on format selected from the list of digital media carrier formats.
Manufacturer	Name of the manufacturer of the digital media carrier.	Provide “Unknown” if the manufacturer cannot be determined.
Brand	Name of the brand or model of the digital media carrier.	If applicable.
Information known about creation and hardware requirements	Comments on file systems, operating systems, software requirements, and hardware requirements.	This information was recorded, when possible, to facilitate appraisal and digital forensic workflows.
Date(s)	Date or range of dates of creation of the data stored on the digital media carrier.	Dates were modified or supplied as possible, but many digital media carriers do not have a known date(s) of creation statement.
Descriptive notes	Other pertinent information about the digital media carrier.	Notes were extracted from file- and item-level scope and content notes or supplied based on labels found on the digital media carrier or container.
Separation date	Date on which the digital media carrier was physical separated.	
Separated by	Name of the individual who physically separated the digital media carrier	All items were physically separated by Domenic Rosati.

Appendix B – List of digital media carrier formats in the Archives Permanent Collection

The list of digital media carrier formats contains the following elements:

Metadata element	Description	Notes
Format	The specific type of digital media carrier	Terms derived from the PBCore instantiationPhysical controlled vocabulary.
Medium	The general type of digital media carrier.	Values include optical disk, magnetic disk, and magnetic tape.
Dimensions	Measurement of the physical size of the digital media carrier.	Provide in cm.
Typical maximum storage (in MB)	Indication of the largest amount of data that can typically be stored on the digital media carrier.	Typical maximum storage cannot be provided for many formats because of variables such as data encoding, video and audio resolution, and other technical specifications.
URI	Reference to the PBCore instantiationPhysical controlled vocabulary.	The controlled vocabulary is published on the Open Metadata Register.

Appendix C – Scripts and tools used in search strategy

Regular expression (REGEX) and AstroGREP

The AstroGREP utility was used to search the Archives' share drive with the following regular expression (Regex):

```
(flopp(y|ies)\.?|(^|\s)dis(k|c|kette)s?(\.|\s|<)|(^|\s)cd(-
rom)?s?(\.|\s|<)|(^|\s)dvds?(\.|\s)|(^|\s)digital(\.|\s|<)|
(^|\s)(usb|hard|flash)\sdrives?(\.|\s|<))
```

This step was conducted as part of the early efforts to learn about the possible scope and distribution of digital media carriers in the Archives Permanent Collection.

XQuery script and legacy EAD XML files

The following XQuery script generated collection-level information from legacy EAD XML files stored on the Archives' shared drive:

```
(: This xquery will create a tab delimited text that
contains collection title, identifier, physical
description, and dates from all the ead xml files in the
/ead path :)

declare variable $tab := "&#9;";
(:This function will return multiple date ranges as one
string :)
declare function local:dateParser($dateArrayTemp)
{
   for $parseTemp in $dateArrayTemp
   return (substring-before(data($parseTemp),"/"), $tab,
substring-after(data($parseTemp),"/"),
if($dateArrayTemp[2]) then $tab else "")
};
(: For all coll:)
for $collectionTemp in
collection("ead/")/ead/archdesc[1]/did[1]
   return (data($collectionTemp/unittitle[1]),$tab,
       data($collectionTemp/unitid[1]), $tab ,
       data($collectionTemp/physdesc[1]), $tab ,
local:dateParser($collectionTemp/unitdate/@normal)
       , "&#10;")
```

The directory included over 500 XML files exported from the Archivists' Toolkit and a batch of derivative XML files generated through an XSLT script written by Libraries' Systems Developer Margaret Vail during a 2014 project to migrate descriptive metadata from the Archivists' Toolkit to Access to Memory (AtoM v2.1.1).

AtoM EAD and DC harvester JavaScript + JSON formatter

Domenic Rosati developed an "AtoM EAD and DC Harvester" JavaScript tool to perform automated searches for media carrier keywords found in EAD and Dublin Core XML metadata made available via DUA's Archives Catalogue and Online Collections.

The JavaScript tool produced JSON metadata that was converted into CSV data using jsonformatter.org.

Appendix D – Mail merge separation sheet template

Separation sheets were produced by mail merging data from the register of digital media carriers into the following separation sheet template:

UNIVERSITY ARCHIVES – SEPARATION SHEET

The following item has been separated from {{Collection Name}} by {{Name}} for preservation. See records from 2016 Born Digital Collection Assessment project for more details.

ITEM TITLE:	{{Title}}
PHYSICAL DESCRIPTION:	{{Extent}} {{Format}}, {{Medium}} : {{Dimensions}}
DATE:	{{Date}}
OLD REFERENCE #:	{{Collection #}}, {{Box #}}, {{Folder #}}
NEW REFERENCE #:	BD Box {{New Box Number}}, Item {{New Item Number}}
DATE OF SEPARATION:	{{Separation Date}}

Separation sheets were printed in duplicate and cut into strips during physical separation (see Section 6.0).