



Process and progress: working with born-digital material in the Wendy Cope Archive at the British Library

Jonathan Pledge and Eleanor Dickens

Contemporary Archives and Manuscripts, Politics and Public Life, British Library, London, UK

ABSTRACT

This article considers the processing of the born-digital archive of the English poet Wendy Cope, deposited at the British Library in 2011. Using the Cope Archive as a template, the authors set out a six-part workflow to acquire, preserve, process and make it accessible. The Cope Archive, which contains several types of digital media, presented various problems. The authors, however, were able to successfully apply their workflow and outline the various software and methods used at each stage. They conclude that, though each born-digital archive presents a unique set of issues, the proposed workflow is a successful framework that would be applicable to most.

KEYWORDS

Digital preservation; archives; literature; born digital; contemporary archives

The papers of Wendy Cope were deposited at the British Library (BL) in March 2011. The archive is a hybrid collection containing both paper-based and born-digital archive material. The paper-based material, consisting of 298 files arranged into 10 series, was catalogued and released to researchers in 2015.¹ The born-digital material – comprising 76 floppy disks of two types (89.3 MB), as well as an email archive and further documents (11.2 GB) saved onto a USB flash drive – remained uncatalogued.

It would perhaps be fair to say that the collection, processing and delivery to researchers of archival born-digital material has long been perceived within the archival profession as a somewhat intractable problem.² This has possibly been caused by a reliance on overly technical workflows with curatorial input removed, either accidentally or intentionally, from the process.

In 2008 the eMSS (eManuscripts) department at the BL successfully pioneered an approach to the processing of born-digital material based upon ‘digital forensics’.³ This approach focused mainly on the successful acquisition and capture of material for preservation. By 2015 it was felt that there were sufficiently mature hardware and software solutions available so that a stable production workflow, from acquisition to access, could be developed to accommodate the growing backlog of born-digital material the BL had acquired.⁴ One of the key aims of this workflow was to provide a correlation between the processing of born-digital material and that of physical archives and manuscripts, using as a starting point the proven digital-forensics approach for capturing born-digital material.⁵

The workflow consists of six main stages:

1. Acquisition: the archive is assigned a catalogue number and digital media is sorted by type for processing. At this stage an assessment is made of whether the digital media is to be processed or set aside. Material set aside might include system disks or program installer disks.
2. Capture: the various types of digital media are captured, where possible as forensic captures.
3. Extracted capture: digital objects are extracted from the forensic captures preserving the original folder and file arrangement and metadata. Any material not considered useful for the final arrangement, such as empty folders, can be removed upon the curator's/cataloguer's discretion.
4. Metadata extraction: the extracted captures are processed with the file-profiling tool DROID to extract metadata.⁶ The metadata is arranged and then copied to an Excel spreadsheet for uploading to our Integrated Archives and Manuscripts System (IAMS). Once the final arrangement is approved, this is then published to the BL Explore Archives and Manuscripts Catalogue.
5. Migration: extracted captures are migrated, where possible, as PDF/As (PDF/A-b (RGB) variant) to create access copies. Any files that cannot be migrated are noted in a production manifest for further action.⁷
6. Access: PDF/A access copies are checked for data-protection clearance.⁸ Cleared PDF/As are then saved to a named directory on an FTP server. This server can be accessed, with restrictions, through dedicated terminals only by researchers using the BL Manuscripts Reading Room.

It is important to note that the workflow presented here is one that we at the BL have found fits within our current strategy for working with born-digital collections. We expect that other archivists and institutions will have their own views on which hardware and software solutions best meet their own requirements.

Processing born-digital material in the Wendy Cope Archive

Wendy Cope, OBE (b. 1945), is a contemporary English poet. Her work is perhaps best known for its wit and her humorous takes on the literary establishment and is often seen as an interesting female perspective to the (predominantly) male poetic canon. Her poetry collections include *Making Cocoa for Kingsley Amis* (1986), *Serious Concerns* (1992) and *If I Don't Know* (2001), which was shortlisted for the Whitbread Poetry Award. Her latest collection, *Family Values*, was published in 2011. Cope has received several critical awards for her poetry and was awarded an OBE in 2010.

1. Acquisition

The born-digital material in the Wendy Cope Archive consists of 57 3.5-inch IBM PC floppy disks (75.0 MB), 19 3.0-inch Amstrad floppy disks (14.3 MB)⁹ and a 16 GB USB flash drive, containing Microsoft (MS) Outlook.dbx email folders along with additional MS Word documents.¹⁰

The floppy disks were sorted by separating disks listed as containing system software from those containing files we might process. The simplification both in the number of disks to be processed and in the corresponding reduction in the number of files has resulted in savings in time and storage space. At this early stage, the archive is entered into our production manifest and all actions and decisions made in relation to processing from then on are recorded.

2. Capture

The capture of the 3.5-inch floppy disks was made using Kryoflux (Figure 1).¹¹ The settings we use provide us with a raw file (.img) and a log file (.txt) (Figure 2).¹² This capture became our master copy and each individual capture was saved to the folder 'Captures'. Once a successful capture was made, each item of digital media was placed in a labelled, acid-free envelope.

In the case of the Cope Archive, the handwritten content descriptions on the floppy disks enabled us to sort them into chronological order, allowing us to remove eight system and software-installer disks from those that needed processing.

The 3.0-inch Amstrad disks (Figure 3) presented a different set of problems. One of the operating systems that Amstrad computers used was 'Locomotive Basic' and the word-processing software provided with this system was LocoScript. We are fortunate that we have an InterMedia system (originally acquired from a publishing house) that allows us to copy files from various legacy digital media. The files from each individual disk were saved to a folder labelled sequentially from '-001'. The primary problem with this approach is that by copying only the files we lose some of the technical metadata associated with the original files. After



Figure 1. Capturing 3.5-inch floppy disks using Kryoflux.

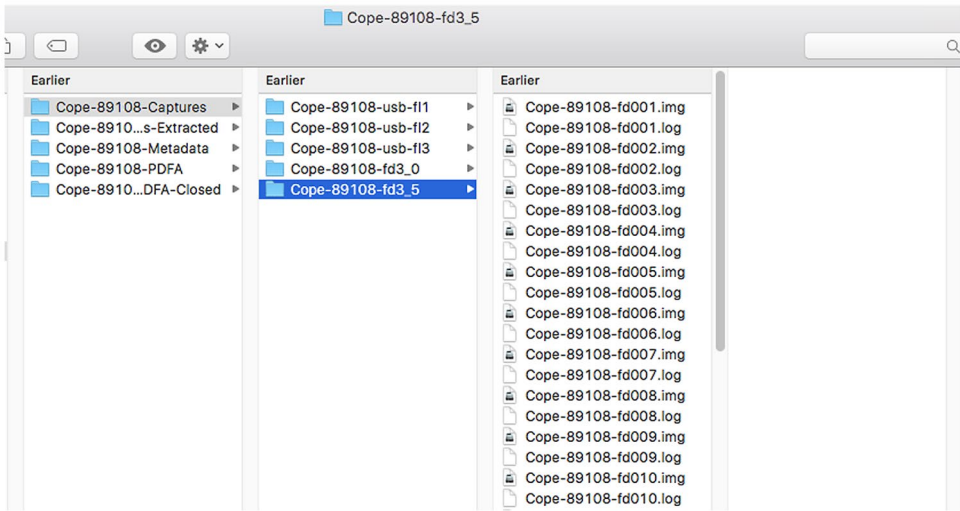


Figure 2. Forensic captures and log files for the Wendy Cope Archive.



Figure 3. Amstrad 3.0-inch floppy disk with original cover-slip.

some consideration we decided to proceed with this process as it gave us a useable output.¹³ We made sure to record our chosen action and justification on the production manifest.

The 16 GB USB flash drive had already been captured previously as an .E01 file when the Cope Archive was acquired.¹⁴ Capture of digital media such as hard-disk drives or

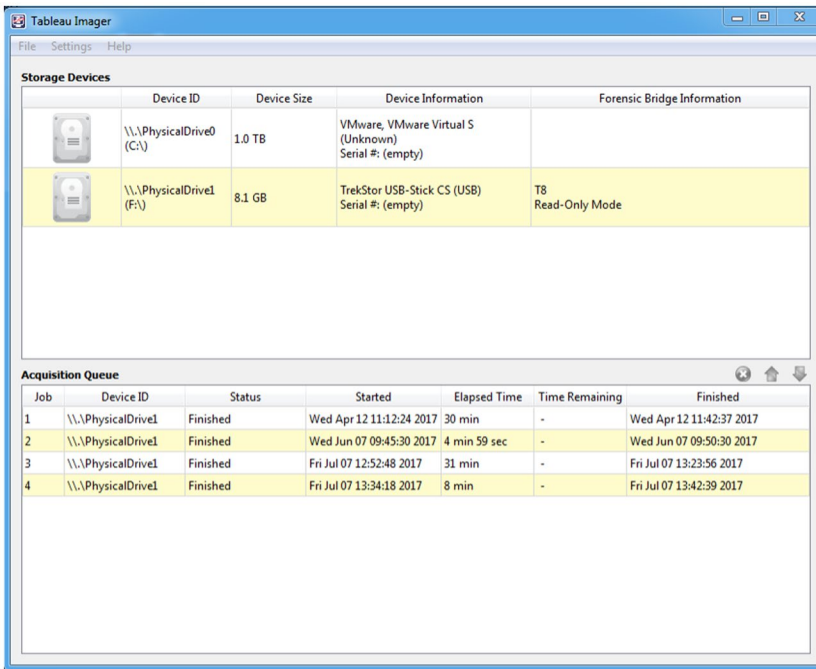


Figure 4. User interface for Tableau Imager software – an excellent tool for basic hard-drive processing using Tableau write-blockers.

flash drives must be done using a write-blocker, which preserves the integrity of the hard drive's data upon capture. We currently use Guidance Software's Tableau Imager software for the bulk of our forensic disk capture as it has a relatively intuitive user interface and gives reliable results (Figure 4).¹⁵

3. Extracted capture

Once we had successfully captured all the digital media, we then extracted the folders and files from each disk image, in turn saving them to a designated hard drive on our main processing workstation.

They were processed as follows:

1. 3.5-inch disk: the .img file for each disk was prepared for copying by mounting it in the operating system finder on our main workstation. It was then saved with the sequential prefix '001-' along with the original name of the disk.
2. 3.0-inch disk: LocoScript files were processed using 'LocoLink for Windows' to convert them to MS Word documents.¹⁶
3. USB flash drive: the .E01 file was exported as an .img file using FTK imager. The .img file was then mounted in the finder, as per the 3.5-inch disks, and saved. In the case of the MS Outlook .dbx email folders, these were converted using Aid4Mail to .mbox format for import into ePADD.¹⁷ The decision to use ePADD was made owing to the number of emails (25,556). Had there been fewer emails then we might have simply exported these direct to PDF/A using Aid4Mail.

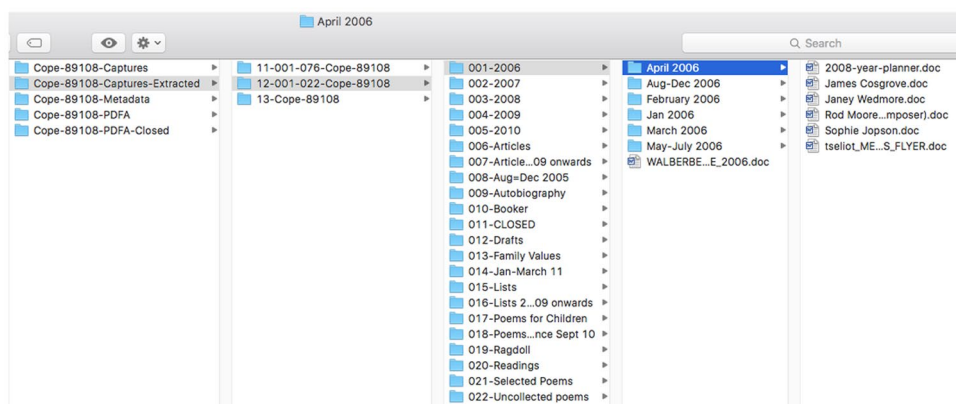


Figure 5. The more complex folder structure on Series 12 meant we had to use sub-sub-series for the catalogue entry.

The extraction process (as documented) was relatively simple. No additional work was required as we were sure, from viewing the extracted captures, that all the files were text documents of some description.

When being captured, each of the three kinds of digital media had been given sequential numbers starting at '001' for each type of digital media. This made it easier to keep track of where we were with the captures, done on different systems and on different days. In the case of the extracted captures, however, we renumbered the various extracted folders with contiguous numbers creating a final arrangement for each series. The 3.5-inch and 3.0-inch disks, Series 11, were numbered 001–076. Because of a more complex folder structure, the MS Word files on the 16 GB USB flash drive were allocated Series 12 and numbered 001–022 (Figure 5). As the MS Outlook .dbx files on the flash drive were processed through ePADD, they were not felt to be part of the existing series arrangement. They were therefore assigned their own Series, 13.

4. Metadata extraction

We used the profiling tool DROID to harvest metadata from the extracted captures by pointing it at the top-level folders in the archive.¹⁸ In the case of the Cope Archive, we were only interested in the metadata from the files in Series 11 and 12. DROID didn't need to be used for Series 13, containing Cope's emails, because ePADD was used to collect this metadata.

The DROID metadata, saved to a .csv file, was then opened in MS Excel (.xlsx) and prepared for uploading in the following manner:

1. Unnecessary metadata fields were deleted.¹⁹
2. Unnecessary rows including invisible files or system files were deleted.
3. Metadata that was to be retained was arranged by either disk or folder, using the sequential numbers added at extraction.

Some metadata was concatenated either to be incorporated into a single field, while some was concatenated to provide labels.²⁰

Once the metadata was organised it was transferred to the BL's archive catalogue system – IAMS – via automated data migration. We use an Excel spreadsheet which outlines all

compulsory metadata compliant with both IAMS and the ISAD(G) cataloguing standard as a template.²¹ The DROID metadata is copied into this spreadsheet and BL catalogue reference numbers are added. It is important at this stage that the catalogue numbers are checked against the extracted captures to ensure that they are catalogued in the correct original order.

Once complete, the IAMS import document was uploaded to IAMS, where it was checked and then published to the British Library Explore Archives and Manuscripts Catalogue.

5. Migration

As stated, the bulk of the files across the three types of digital media processed appeared to be text documents, either WordPerfect (.wpd), MS Word (.doc) or text documents (.txt), and were therefore ideal for migration to PDF/A. Had we found unidentifiable or specialist proprietary document types – or a collection of files such as a bespoke computer program sometimes found in scientific archives – unsuited to migration, then we would record this in our production manifest for action at a later date.²²

Once we were satisfied that we had identified the documents suitable for migration, they were processed to create PDF/As. This is done using either:

1. QuickViewPlus – an excellent tool for viewing and converting many types of files.²³ Files can be batch processed without needing to open the file in its native program.
2. Adobe Acrobat Pro – provides a number of ways to convert files to .pdf and is very good for the bespoke processing of problem files.²⁴
3. Foxit PhantomPDF – has useful batch-processing features (Figure 6).²⁵

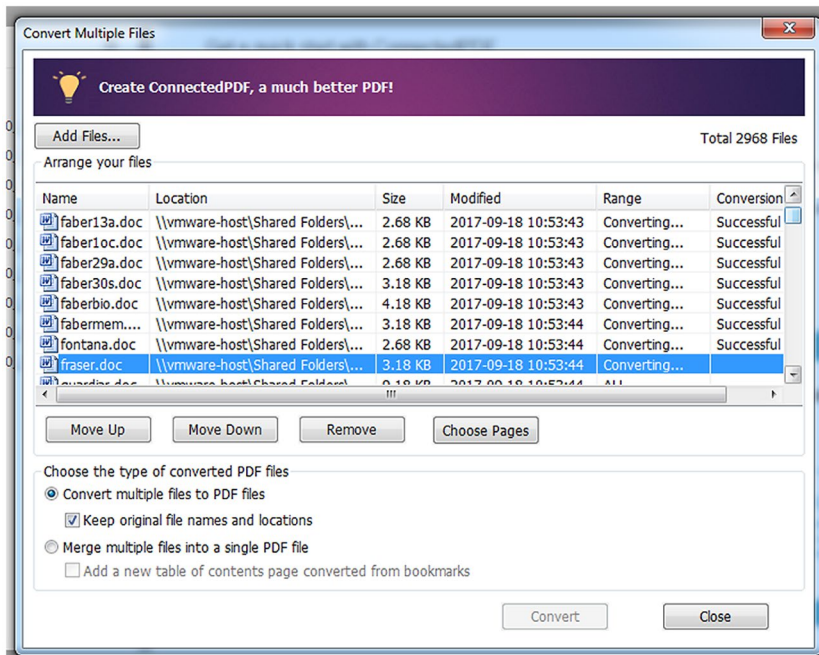


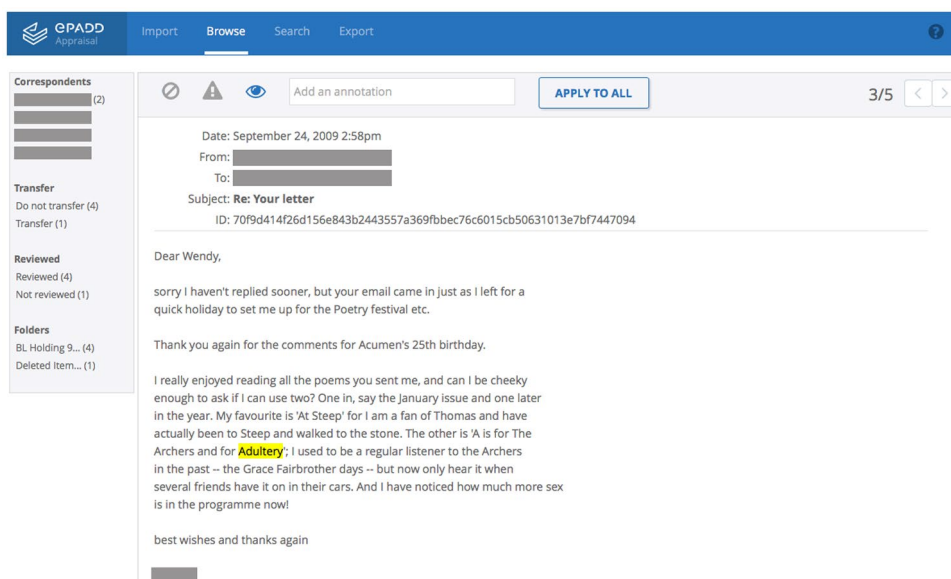
Figure 6. Bulk PDF processing interface for Foxit PhantomPDF.

If PDFs are not saved directly to PDF/A, then this conversion is done as a batch conversion via Acrobat Pro. Converted PDF/As are saved to a folder structure that replicates the catalogue hierarchy uploaded to the archives catalogue, IAMS. They are then renamed sequentially, from Add MS 89108/11/1, using Advanced Renamer software.²⁶ Some care is needed at this stage to make sure the numbers assigned to the files at the metadata stage correspond with those assigned to the access PDF/As. It is these files that will be provided as access copies to researchers.

ePADD

As mentioned, the Cope Archive contains a large corpus of email dating from 20 August 2004 to 9 March 2011. The decision was made to use ePADD as this provides a flexible platform for processing, curating and providing access to large collections of email. ePADD is comprised of four modules: Appraisal, Processing, Discovery and Delivery. The first three modules are all handled within ePADD, with the fourth being an extensible module that can require some IT customisation to use.

From the Cope Archive acquisition document, we were able to ascertain that the original Outlook .dbx files had been saved from at least two separate email accounts. The .dbx files were converted to .mbox files. This conversion, performed using Aid4Mail, changes the single .dbx email folder to separate mailboxes contained, in our case, in a 'local folder', and the .mbox files were then imported into ePADD. Given the nature of email, one of the most useful features in ePADD available in the Appraisal stage is the lexicon search, which separates email into categories via a keyword search (Figure 7). Selecting each category allows either individual or groups of categorised emails – via the three options: 'Message reviewed', 'Transfer with restrictions' and 'Do not transfer' – to be tagged for export to the 'Processing' module.



The screenshot displays the ePADD Appraisal interface. The top navigation bar includes 'Import', 'Browse', 'Search', and 'Export'. On the left sidebar, there are filters for 'Correspondents' (2), 'Transfer' (Do not transfer (4), Transfer (1)), 'Reviewed' (Reviewed (4), Not reviewed (1)), and 'Folders' (BL Holding 9... (4), Deleted Item... (1)). The main content area shows an email header with 'Date: September 24, 2009 2:58pm', 'From:', 'To:', and 'Subject: Re: Your letter'. The email body contains a message to Wendy, mentioning a poetry festival and a birthday. The word 'Adultery' is highlighted in yellow in the text. At the bottom, there is a 'best wishes and thanks again' and a redacted signature.

Figure 7. The sensitivity review page in ePADD. Usefully, the word that has resulted in the email being flagged is highlighted.

In light of the legal obligation we have in the UK owing to the Data Protection Act 1998 when providing access to contemporary archives and manuscripts, this is an invaluable feature.²⁷

6. Access

Our access model uses a resource that was originally developed by the BL's Endangered Archives Programme (EAP). EAP requires a platform to deliver this content to researchers whilst abiding by standard BL practices in relation to copyright and data protection for the delivery of archives and manuscripts. In this case they use an FTP server with access restricted solely to researchers using the BL Asian and African Studies Reading Room. Owing to the relatively small amount of born-digital material we were able to make available, we decided to use this model rather than provide access through our central Library repository.

PDF/As are uploaded, via FileZilla FTP,²⁸ to a named directory on the server (in this case: Cope_Add_MS_89108). Access is provided via a fixed URL which is provided to the researcher upon request. The PDF/As can be viewed only via computer terminals in the BL Manuscript Reading Room. The PDF/As are locked for viewing only and cannot be edited, saved or printed.²⁹

Conclusion

At the British Library we have developed a workflow which matches curatorial expertise with the application of mature technical solutions for processing and providing access to the born-digital material that is increasingly found within contemporary personal archives.

Within the United Kingdom at least, this would seem to be a novel achievement and one that we are keen to build on. Our approach has provided us with experience of the reality of processing different types of digital media and allowed us to reduce our cataloguing backlog. In the process we have gained valuable feedback from researchers as to how best to provide access to born-digital archival material.

We have found that each born-digital archive tends to present its own unique set of problems. Through our work on the Wendy Cope Archive, we are confident that we have established a workflow that will be applicable to most born-digital archives that we expect to acquire in the near future.

Endnotes

1. See: BL Add MS 89108.
2. Where born-digital material is referred to in this article, we are talking about material that was originated on a computer and not physical material that is digitised. For some early discussions and projects to address this problem in the UK context, see the University of Oxford and University of Manchester, Personal Archives Accessible in Digital Media (paradigm) project, available at <<http://www.paradigm.ac.uk/>>, accessed 29 October 2010, and successive Bodleian Library projects available at <<http://www.bodleian.ox.ac.uk/beam/about/projects>>, accessed 29 October 2017; Jeremy Leighton John with Ian Rowlands, Peter Williams and Katrina Dean, 'Digital Lives: Personal Digital Archives for the 21st Century: An Initial Synthesis', British Library, 2010, available at <<http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis01a>>.

- [pdf](#)>, accessed 29 October 2017; University of Hull, Stanford University, University of Virginia and Yale University, 'AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship', 2011, available at <http://dhcommons.tamu.edu/sites/default/files/aims_final.pdf>, accessed 5 November 2017. Approaches to preserving born-digital cultural heritage using digital forensics are outlined in Mathew G Kirschenbaum, Richard Ovenden and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, Council on Library and Information Resources, Washington, DC, 2010, available at <<https://www.clir.org/pubs/reports/pub149/pub149.pdf>>, accessed 29 October 2017. A range of perspectives are presented in Christopher A Lee (ed.), *I Digital: Personal Collections in the Digital Era*, Society of American Archivists, Chicago, 2011. In the Australian context, see 'Guidelines for Library Staff Assisting Donors to Prepare their Personal Digital Archives for Transfer to NSLA Libraries', 2nd edn, National and State Libraries Australasia, November 2011, available at <http://www.nsla.org.au/sites/www.nsla.org.au/files/publications/NSLA.Guidelines_donor_digital_archives_201111.pdf>, accessed 29 October 2017.
3. Jeremy Leighton John, *Adapting Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools*, The British Library, London, 2008, available at <http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf>, accessed 29 October 2017.
 4. The British Library's born-digital collection of personal digital archives is at present 860 GB in size and is comprised mainly of scientific and literary archives.
 5. Research and development of alternative methods of born-digital processing, including both migration and emulation, is now undertaken by the BL's Digital Preservation department.
 6. Further information is available at <<http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>>, accessed 29 October 2017.
 7. The production manifest is an MS Excel document where each stage of processing, including actions, problems and solutions arrived at, is recorded.
 8. UK Data Protection Act 1998, available at <<http://www.legislation.gov.uk/ukpga/1998/29/contents>>, accessed 29 October 2017. The UK Data Protection Act 1998 was created to bring the UK into line with the EU Data Protection Directive 1995. Currently a new Data Protection Bill is passing through the UK Parliament in order to comply with the EU's General Data Protection Regulation 2016, available at <<http://www.eugdpr.org>>, accessed 5 November 2017.
 9. Centre for Computing History, Amstrad plc, available at <<http://www.computinghistory.org.uk/det/1227/Amstrad-Plc/>>, accessed 29 October 2017.
 10. There were also two additional 8 GB flash drives acquired but upon examination at the capture stage we realised they contained duplicate material of the 16 GB flash drive. In this case they were captured but not exported as 'extracted captures'.
 11. Details of Kryoflux are available at <<https://www.kryoflux.com>>, accessed 29 October 2017.
 12. Kryoflux settings used are: Kryoflux Stream Files, Format Guided; Section Image: MFM; Extension: img. All other settings are default.
 13. An alternative solution would have been to use an external Amstrad disk drive with Kryoflux to create preservation disk copies. This may be a path we take at a future date, although the number of Amstrad disks we have acquired is so small as to make allocation of any further resources to this problem difficult to justify.
 14. .E01 (Encase Image File Format). Proprietary file type developed by Guidance software for forensic investigation of born-digital material primarily for legal or law-enforcement purposes.
 15. Available at <<https://www.guidancesoftware.com/tableau/download-center-tim>>, accessed 29 October 2017. You can also use AccessData's FTK Imager or Guidance Software's Encase Forensic Imager, which provide additional functionality not available using Tableau Imager.
 16. LocoLink for Windows is available at <<http://www.locoscript.uk>>, accessed 29 October 2017.
 17. Information about Aid4Mail is available at <<http://www.aid4mail.com/>>, accessed 29 October 2017; Stanford's ePADD is available at <<https://library.stanford.edu/projects/epadd>>, accessed 29 October 2017.

18. We run DROID over the extracted captures prior to any change to the creator's structure so that the file path metadata reflects the original digital arrangement of files.
19. The decision on which metadata fields captured from born-digital archives would be transferred to the archival catalogue was decided by the British Library Born Digital Cataloguing Group in January 2017.
20. For example, the Checksum data is labelled 'SHA256_HASH:' to denote the type of Checksum used, as the IAMS field only specifies 'Hash Value'.
21. The BL uses a bespoke archive catalogue, the Integrated Archives and Manuscripts System (IAMS).
22. When processing the 3.5-inch disks, we found that 10 contained WordPerfect files. These are usually easily opened with MS Word. In this case, however, they would appear to be pre-WordPerfect 5.0 files, which cannot be converted in this manner. In this case we may look to emulation, at a later date, to provide access.
23. QuickViewPlus is available at <<https://avantstar.com/quick-view-plus-2017#fndtn-overview>>, accessed 29 October 2017.
24. Working on another archive, it was found that the only way to process image-heavy PowerPoint files (.ppt) originally created in Mac OSX 10.7.5 was to print the document as a PostScript file (.ps) and then create the PDF/As by dropping the PostScript files into Acrobat Distiller (included with Acrobat Pro).
25. Foxit PhantomPDF is available at <<https://www.foxitsoftware.com/pdf-editor/>>, accessed 29 October 2017.
26. Advanced Renamer is available at <<https://www.advancedrenamer.com/>>, accessed 29 October 2017.
27. As of October 2017 we are still checking the Cope emails for data-protection issues.
28. FileZilla® is available at <<https://filezilla-project.org>>, accessed 29 October 2017.
29. In March 2017 we held a testing day for our access model with a mix of student volunteers, technicians from the BL Digital Preservation department and academics. The testing was very successful and several suggestions for improved presentation have been noted and are in the process of being implemented.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Jonathan Pledge is a curator for Politics and Public Life in the Department of Contemporary Archives and Manuscripts at the British Library. With an MA in Cultural Heritage Studies from UCL his research interests include the history of science, information technology and computing.

Eleanor Dickens is a curator for Politics and Public Life in the Department of Contemporary Archives and Manuscripts at the British Library. She holds a MA in Archives and Record Management from UCL. She is critically engaged with research into women's history and wider political protest and activism.