Born-digital archives at the Wellcome Library: appraisal and sensitivity review of two hard drives

Victoria Sloyan

Wellcome Trust, London, UK

ABSTRACT

Digital preservation has been an ongoing issue for the archival profession for many years, with research primarily being focused on long-term preservation and user access. Attention is now turning to the important middle stage: processing born-digital archives, which encompasses several key tasks such as appraisal, arrangement, description and sensitivity review. The Wellcome Library is developing scalable workflows for born-digital archival processing that deal effectively both with hybrid and purely born-digital archives. These workflows are being devised and tested using two hard drives deposited within the archives of two genomic researchers, Ian Dunham and Michael Ashburner. This paper examines two specific and interconnected stages of archival processing: appraisal and sensitivity review. It sets out the Wellcome Library’s approach to appraisal using a combination of several appraisal methods, namely functional, technical and ‘bottom-up’ appraisal. It also demonstrates how tools such as DROID can be used to streamline the process. The paper then goes on to explore the Wellcome Library’s risk management-based approach to the sensitivity review of born-digital material, suggesting there is a viable balance to be struck between closing large record series as a precaution and sensitivity reviewing at a very granular level.

KEYWORDS

Born-digital; appraisal; sensitivity review; archives

Introduction

The question of how to deal with born-digital records is a hot topic for the archival profession and shows no sign of being resolved any time soon. Until recently, the focus has been on long-term preservation and the provision of access to born-digital records. But attention is now turning to the important middle stage: the processing of digital archives and all that this encompasses including appraisal, arrangement and description.

The Wellcome Library is no exception to this shift. In recent years, the priority has been to implement a secure digital repository and ingest our existing holdings from their original storage media. Now that we have embedded ingest workflows and the born-digital records are safe from imminent threats, we are considering how best to process and catalogue these records.

CONTACT Victoria Sloyan v.sloyan@wellcome.ac.uk

© 2016 Wellcome Trust. Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Whilst there is a wealth of literature on issues surrounding long-term digital preservation, the archival profession has not, as yet, been so vocal in discussing the processing of born-digital records. The profession needs to communicate and promote discussion of the issues surrounding born-digital processing, as it has successfully done with long-term preservation, in order to refine ideas, develop best practice and encourage all archivists to tackle the issues. Communication can involve sharing successful projects and procedures, but archivists should also have the confidence to broadcast works in progress, failed efforts and superseded ideas, as these can generate fresh perspectives and inspire new ideas.

The Wellcome Library has taken up its own challenge and this paper contributes to the hopefully ever-growing archival literature on digital archival processing by setting out the approaches used to appraise and sensitivity review two external hard drives.

The Wellcome Library is a reference library for the study of health in historical and cultural contexts. In 2010, the Library moved to align itself more closely with the specific biomedical research interests of its parent body, the Wellcome Trust, and consequently began a concerted effort to survey and collect archives that document the Human Genome Project and surrounding developments in genomic research. The Library adopted a broad, documentation strategy-based survey approach whereby all identifiable individuals and organizations involved in the Human Genome Project were contacted about their potential archives. As a result, six collections were deposited with the Library and are being catalogued as part of the Library’s Collecting Genomics project. The majority of the collections are hybrid archives comprising both paper and born-digital records. Two include external hard drives which pose several challenges, particularly regarding the scalability of the Library’s existing appraisal and sensitivity review workflows which were originally developed around much smaller born-digital deposits.

The two external hard drives come from the archives of Ian Dunham and Michael Ashburner. Ian Dunham is a geneticist who worked on the sequencing of human chromosome 22 as part of the Human Genome Project.1 His collection comprises his professional personal papers, consisting of 17 archival boxes of paper records, four 3.5“ floppy disks and one external hard drive containing 4559 digital files (5.6 GB in size). Michael Ashburner is a fly geneticist famous for his work on the genetic structure and genome sequence of the fly Drosophila melanogaster.2 His collection also documents his professional life but is far larger comprising over 300 archival boxes of paper, 58 3.5“ floppy disks, three optical disks and one external hard drive containing 16,378 digital files (12.7 GB in size). Both hard drives contain a variety of records including reports, meeting minutes, grant applications, presentation slides, working papers for published articles and some research data.

The hard drives were used as testing grounds to refine Wellcome Library approaches to the appraisal and sensitivity review of born-digital records, resulting in the development of procedures that can be applied to all born-digital records regardless of format or quantity. That is not to say this paper claims to provide all the answers. Rather, it is hoped this paper will act as a starting point that will encourage conversation within the archival community, leading to further refinement and improvement.

Digital appraisal

The concept of archival appraisal needs no introduction to archival professionals. Digital appraisal is no different to paper appraisal in that it is underpinned by the same archival
theory, but there is a tendency for it to be seen as a distinct process rather than the same process being undertaken on a different record format. As such, some practitioners question its necessity and validity. Terry Cook is a chief critic, proposing the adoption of a neo-Jenkinsonian attitude towards digital records whereby nothing is appraised and everything that is deposited is retained by the archival institution. He claims that the ongoing decrease in storage costs coupled with a continuous increase in capacity means that archival institutions do not face the same storage pressures as with physical archives and so are able to retain as much as they want. Easy discovery and retrieval by researchers is achieved through a combination of sophisticated search engines, metadata and archival description. Cook is not alone in re-evaluating appraisal for the digital sphere. The Paradigm Project highlights several barriers to effective appraisal of digital material, though it does not reject it outright. Paradigm argues that the ability to appraise digital records is undermined by the disorder in which many records are kept, particularly personal archives, where there is not the same level of enforced behaviour regarding the use of directories and filename conventions as there often is with corporate archives. This lack of structured organization means appraisal is only achievable through a very granular file-level appraisal, which is prohibitively time-consuming. In both arguments, quantity is a key argument against appraisal: there is too much to deal with, so it should not be attempted. In actual fact, quantity is one of the more compelling reasons for undertaking appraisal.

Cook’s central argument against appraisal focuses on storage costs, but he has misinterpreted the actual situation. Whilst it is true that storage costs have decreased on a per-byte basis, this does not take into account the fact that digital content is growing at an extraordinary rate. Capacity is increasing but crucially, people’s expectations are rising with it and they produce more digital content to fill the additional capacity as it becomes available. Consequently, the increased quantity of digital records is nullifying any decrease in storage costs, meaning the overall cost for a digital repository will remain the same, if not increase. Furthermore, digital preservation is an active process that requires ongoing action in order to maintain the accessibility of digital records in the face of file format, software and hardware obsolescence. Every digital file held by an institution will affect the overall digital preservation cost and arguments in favour of storing more digital files than necessary are unlikely to find much traction with archivists struggling to manage budgets. Appraisal does come with its own costs, primarily in staff time, and these should always be weighed against preservation costs when deciding on appropriate action. However, providing appraisal is undertaken as efficiently as possible, appraisal is, in most cases, worth the investment.

The Paradigm Project is correct in stating that disordered digital collections are hard to appraise, but totally disordered collections should not be accepted as a prevailing trend. Many people still rely on some form of organization of their digital records and use directory structures to store and locate their files, rather than rely on search functions. The experience of the Wellcome Library suggests that file organization is more likely to be present the larger the capacity of the storage media. Some 3.5” floppy disks we have processed contain no directory structure. But a set of disks, particularly labelled ones, is in itself a form of organization and the small capacity means even disorganized floppies contain a manageable number of files. In contrast, both the Dunham and Ashburner hard drives have some level of directory structure and an informal survey of Library staff indicates the overwhelming majority of hard drives and shared drives, used both in a personal and professional capacity, are organized to some degree. Furthermore, Cook’s claim that organization is unnecessary as users can
retrieve records through search functions fails to recognize the number of researchers who utilize browse functions. Not everyone is looking for particular records or has search terms to use; some people want to browse and explore. Vast quantities of data can be intimidating and unwieldy and can prevent efficient research, especially when much of the data serve no purpose. Few archivists would accept a paper deposit without some level of appraisal and digital archives should not be treated differently just because of quantity.

**Appraisal frameworks**

Having established that born-digital records would undergo some form of appraisal, the Wellcome Library assessed standard paper appraisal methods to determine a way forwards. Professional opinion on this seems divided. Some believe old concepts are inadequate to deal with digital records and new ones need to be devised. Others argue that paper and digital records are not so different and traditional concepts still apply. The Wellcome Library takes the latter view, with the caveat that whilst the appraisal concept remains the same, the difference in construction and accessibility needs to be acknowledged. This may necessitate new practices, but these remain underpinned by archival theory.

Modern archival appraisal has been extensively written about and detailed analysis does not fall within the scope of this paper. However, key appraisal approaches are here defined so as to clarify how they are used in this paper.

- **Macro appraisal**: undertaken at a high level, assessing large sets of records rather than individual ones. Consideration is given to the process, function and structure of records, rather than content and informational value.
- **Functional appraisal**: a subset of macro appraisal involving the analysis of the functions of the record creator and the retention of records created as a result of these functions.
- **Micro appraisal**: very granular process whereby each record is appraised individually.
- **Bottom-up appraisal**: an approach suggested by the Paradigm Project whereby micro appraisal is used on a sample of records to ascertain a broad classification and determine whether file and folder titles are accurate.

There is much archival literature on appraisal frameworks, and these are increasingly including born-digital records in their scope. The frameworks analysed by the Wellcome Library all promote an integrated appraisal process involving a blend of different approaches. A framework should be built around an institution’s own selection criteria, legal issues, technical considerations, preservation factors and the presence or lack of information and metadata from which value judgements can be made. Some frameworks provide more detail and link specific appraisal methods to the different considerations, such as using micro appraisal to determine significant properties of individual records and macro appraisal to assess functions and surrounding context. Importantly, these frameworks do not distinguish between paper and digital records in their fundamental attitude towards appraisal; although they do recognize there is a technological element that has an impact on digital appraisal decisions, both in terms of the value of certain technical characteristics and the feasibility of preservation.

Appraisal frameworks do overwhelmingly focus on corporate records and cannot always be mapped satisfactorily to personal papers. Nevertheless, they did help us begin to consider the kinds of questions that need to be asked of the records, and reflect on how technical considerations should impact appraisal decisions. The Paradigm Project provides one example
of a successful appraisal methodology for born-digital personal papers, though due to the limited scope of the project the focus is on political papers and the functions and record types particular to that area.\textsuperscript{11} By consulting both kinds of framework in tandem we were able to extract enough guidance and useful tips to enable us to devise a framework of our own that seemed appropriate for our needs. This framework combined high-level functional appraisal geared around specific career functions with bottom-up appraisal undertaken at folder level, rather than individual record level. There was also an element of technical appraisal to aid file format identification. This framework then needed to be tested with the hard drives.

**Practical application**

Existing literature is useful when considering appraisal frameworks and policies, but there is little in the way of advice on the practical application of such frameworks. Given that quantity is widely acknowledged to be a hindrance, it is surprising to find there is little published advice on how to tackle this or how to appraise in an efficient way. Once again, the Paradigm Project is the best source of advice, though that too is limited.\textsuperscript{12} Therefore, as well as devising and testing an appraisal framework, the Wellcome Library used the two hard drives to experiment with appraisal approaches and tools to identify efficient practices.

The timing of appraisal within the wider processing workflow was carefully considered and occurred at two stages. As recommended by the Paradigm Project, the Wellcome Library always aims to involve depositors in early, macro appraisal prior to records being deposited.\textsuperscript{13} Conversations with Dunham and Ashburner enabled the archivist to identify key record series on each hard drive and explain the obvious types of record that would fail appraisal, allowing the depositors to weed their drives prior to deposit. There is some risk in allowing depositors too much opportunity to retrospectively weed and arrange digital records, but no more so than there is with paper records. Moreover, this risk is outweighed by the benefits of depositor involvement. Information provided by both depositors allowed us to place the hard drives and their various contents within the context of each entire archive. Moreover, Ashburner provided a one-sentence description of each top-level folder, which proved extremely helpful. The usefulness of involving the depositor will depend on their willingness and ability to be involved, but we would definitely recommend it is an option to be explored with every new depositor.

New digital deposits received by the Wellcome Library are immediately virus checked before the data is copied and pasted from the original storage media onto a shared drive holding area on a networked server. Here is where the second stage of appraisal takes place, before records are ingested into our digital preservation system, Preservica (http://preservica.com). Paradigm suggests detailed appraisal can be done at a later stage once descriptive metadata has been generated through cataloguing.\textsuperscript{14} This is logical and turns archival processing into an iterative process, but it is not compatible with the Wellcome Library's digital preservation system. Removing digital files from Preservica is a very arduous and manual process and as such is best avoided. It does not seem sensible to ingest files into Preservica that are then going to fail appraisal, as they then need to either be deleted or left in the system for no real reasons. This can be accepted when it involves a handful of additional files, but when it is a case of thousands, or even millions, it means significant storage and preservation resources are being wasted. Furthermore, Preservica has been synched to the cataloguing software CALM so that catalogue records are automatically generated for each
digital file upon ingest into Preservica. If these records then fail appraisal, the CALM records need to be deleted, adding another task to the cataloguing process. In short, our systems and workflows mean we find it most efficient to undertake appraisal prior to the records being ingested in our digital repository.

The single most useful tool we found for appraising born-digital records was the export produced by DROID. DROID (Digital Record Object IDentification) is a free file profiling tool developed by The National Archives. It produces a range of metadata including filenames, last modified dates, file formats and extensions and can assign each file an MD5 checksum. The metadata can be exported as a CSV file which can then be interrogated to assist with appraisal, particularly technical appraisal. Some of these interrogation practices are detailed below.

**Functional appraisal**

First, we used the DROID export to get an overview of the directory structure on each hard drive and identify key functions. This work was supported by information received from the depositors, particularly with regards to deciphering the organisational structure used within the directories and identifying key functions of Dunham’s and Ashburner’s careers. Through this, we identified common folders likely to fail appraisal, such as ‘Download’ and ‘Trash’ folders. The filenames of the contents of these folders were briefly assessed to ensure the folder titles were accurate, and where necessary a small sample of files were viewed. These folders were then deleted from the copies of the hard drives on the Wellcome Library network drive. From here, we moved on to technical appraisal.

**Technical appraisal: identifying unwanted file formats**

Previous appraisal work on smaller born-digital deposits had raised the possibility that there were certain file formats that would nearly always fail appraisal. The overview of each hard drive suggested a similar thing, and it was thought feasible to appraise certain records based entirely on their file format. We used the DROID export to identify the various formats each hard drive contained and after evaluation devised a list of file formats that commonly held little value and could be weeded out with only a light assessment. The formats include the following:

- Thumbs.db (.db)
- Application help files (.hlp)
- Icon files (.ico)
- Temporary files (.tmp)

Both hard drives contained unfamiliar file formats that were unidentified by DROID and had to be identified manually using the file extensions and the online database www.fileinfo.com. These file formats tended to be very specific to genomic research, such as UCSC BED Annotation Track files which hold genome annotation track data and FASTA Sequence Files, which hold nucleic acid and protein sequence data. These types of files met our functional appraisal criteria, though they were identified and the appraisal decision was made based on the file format rather than file and folder names.

Technical appraisal had limited success with the Ashburner hard drive due to a problem with some digital files that were originally created on a SUN system and had later been migrated to Mac. Ashburner originally used full stops in filenames, which modern operating
systems and DROID interpreted as denoting the file extension. For example, one file entitled OMIM.960131 was stated as the file OMIM with the file extension .960131. In actual fact, the numbers are a date: 31 January 1996 and the real file extension has disappeared. The file formats were unidentified by DROID and manual identification could not be carried out as there was no accurate file extension to use. Therefore, technical appraisal had limited use on this occasion and more reliance was placed on the other forms of appraisal.

Technical appraisal: identifying duplicates
The DROID export was used to detect duplicates by sorting the data by MD5 checksum and identifying duplicate values. All duplicates were then evaluated and after doing this for both hard drives, certain patterns started to appear. In some cases, every file in a folder was duplicated elsewhere, suggesting it was a back-up. In other instances, certain organizational techniques could be identified, including the use of deliberate duplication to allow records to serve different functions. One record might exist in a folder about a particular project and also in a folder of meeting minutes. Instances like this underline an important point: duplicates were not automatically deleted simply because they were duplicates; further analysis was first undertaken to ascertain why they existed. Where duplicates serve different purposes, deleting one of the records can destroy meaning and value in its folder. Therefore, it is important to analyse the context surrounding a record, rather than deleting it purely for being a duplicate. Where it was decided to delete a duplicate file, we used the last modified date to help establish which file to delete.

Bottom-up appraisal
After technical appraisal, the next step was ‘bottom-up’ appraisal. The Dunham hard drive was arranged into 82 top-level folders containing over 4500 digital files, with 73 additional ‘loose’ files. Functional appraisal had already weeded out irrelevant folders and identified key functions. Bottom-up appraisal through the spot-checking of folder contents allowed us to ensure the folder names were accurate before final appraisal decisions were made. More detailed examination was also given to certain sub-folders where it was felt to be appropriate. Some failed the Wellcome Library’s specific appraisal criteria (for instance, we do not keep student references) and some sub-folders proved not to contain anything of value as the folder name suggested: one sub-folder for a grant file contained a software application and tutorial file but no actual grant records. Dunham’s hard drive was very well organized. He had made good use of file directories to keep relevant groups of records together and he gave the majority of files and folders accurate and descriptive names, making appraisal a fairly simple and quick process. After appraisal, the hard drive contained 3132 files (4.69 GB), a reduction of more than 1400 files.

The Ashburner hard drive differed in many ways. It had 22 top-level folders containing 16,304 digital files, with 74 loose files. The directory structure was more hierarchical than Dunham’s and included many more sub-folders. This, combined with the fact that Ashburner had already provided a one-sentence description of each top-level folder, led us to focus on the next tier down, the sub-folders. Another contrast was that the Ashburner file and folder names were not as descriptive, since Ashburner made more use of acronyms and used dates in the filename as a means of organization. Knowledge gained from cataloguing the paper records helped identify some acronyms and thus some functions, but each sub-folder required more detailed spot-checking for accurate appraisal.
With both hard drives, basic descriptive metadata such as date ranges and brief content descriptions were documented whilst appraisal was undertaken. This added time to the appraisal procedure, but made the entire processing workflow more efficient as the folders will not have to be re-assessed during the cataloguing phase in order to write catalogue descriptions.

Early on in the appraisal process, it became apparent that Ashburner stored extensive amounts of research data on his hard drive, alongside other work files. This raised the suggestion of using record sampling as an appraisal method.

**Sampling**

Record sampling does not appear to have wide application these days. It had a period of popularity in the 1980s and early 1990s, but application has since declined, or is at least not being written about. Misgivings are understandable. One critic views sampling as a necessary evil forced upon archivists due to storage and other resource pressures, that ultimately leads to the loss of information regardless of the care taken to ensure a representative sample remains.\(^{16}\) Terry Cook has also highlighted the difficulty that can be had in ensuring a truly representative sample is taken and questions the validity of taking a sample that is not truly representative.\(^ {17}\) One danger is that an unrepresentative sample can result in misinformed research, if the sampling method and the extent of its representativeness are not made clear. Despite these challenges, sampling can have a place in archival appraisal, if applied with due diligence and planning.\(^ {18}\)

Most critics agree that sampling is most appropriate for use with homogenous record series.\(^ {19}\) Evelyn Kolish goes further, arguing sampling works best with records that also contain informational value that is ‘both relatively consistent and not extremely high’.\(^ {20}\) An additional support for sampling is instances where the information contained in a record series is readily available elsewhere.\(^ {21}\) Therefore, there is strong argument for sampling in specific cases where record series meet the criteria. Care needs to be taken over the sample methodology and certain approaches do lack rigour. However, that is a weakness of those approaches, not of sampling as an appraisal tool, and should not be used to rule out sampling.

As has been noted, the Ashburner hard drive contained a large proportion of research data files either created by or sent to Ashburner as part of his genome sequencing work. Raw research data do not fall within the Wellcome Library’s collecting policy as they often lacks contextual information and cannot be interpreted by Library researchers, who on the whole lack the specialist knowledge required. Those who can and wish to interpret the data can access it through publicly accessible databanks, such as GenBank (http://www.ncbi.nlm.nih.gov/genbank/), since much of the genome sequencing information produced in the last 25 years or so has been deposited in such places. Furthermore, reducing the number of data files on the hard drive would make it easier not only to catalogue but also to navigate and exploit for research purposes. We did not wish to dispose of all the data records because they do provide some secondary evidential value in terms of documenting how large-scale genome sequencing collaboration and data sharing was undertaken on a day-to-day basis. Nevertheless, the consistency in the records means this type of evidence need only be witnessed in a handful of records, not the entire set. Following Kolish’s criteria outlined above, the data contained in these hard drive records are easily accessible elsewhere, the records
existed in homogenous series, the type of information included in these records was consistent and, due to the lack of surrounding context, the informational value was not particularly high, thus sampling seemed to be appropriate.

Therefore, the decision to sample these files was made and work began on identifying those record series suitable for sampling. This had already been done to some extent during earlier stages of appraisal, which was when the idea to sample first originated. A more systematic approach was now taken, revisiting the series highlighted earlier and analysing the DROID export and directory structure to identify other potential series. A judgement was made on the amount of time taken to identify appropriate series versus the number of files likely to be weeded. Hence, only obvious folders were identified and there are likely to be files on the hard drive that could have been sampled and were missed, but we believe this potential omission is justified.

When selecting the specific sampling approach, we consulted Terry Cook’s work on archival sampling and the distinction made between three methods often all referred to as ‘sampling’: actual sampling; selection; and exampling. Exampling involves selecting one or two specimens as illustrations. Selection is where specific records are chosen because they contain a predefined significant characteristic, such as being created by women. Sampling involves the retention of a percentage of a record series that reliably represents the whole. Selection distorts rather than reflects the whole series and exampling has limited evidential value and is not wholly representative, therefore true sampling is what we aimed for. Cook sets out two methods for this that avoid personal bias influencing the choice of records. One method is termed ‘simple random sampling’ whereby every record is given a number, 50 numbers (or however many files you wish to retain) are then randomly generated and those records that correspond to the generated numbers are retained and preserved. The second method, ‘Systematic random sampling’, uses a random number generator to select the starting position (i.e. 3rd record) and then every following nth record is retained. Systematic random sampling is easier to implement and so it was selected as the method. It can be biased in cases where file organization follows a cyclical pattern and retaining every nth record results in the same type of file being retained, but this did not seem to be the case with the Ashburner files. Sampling was undertaken at either digital file level or folder level. On most occasions, individual digital files were sampled, but folder sampling was used on two occasions where a series of sub-folders existed, all of which contained the same information but from different dates.

The decision on the percentage of records to preserve in each sample was based on their nature. If all the files contained exactly the same kind of genomic data, a lower proportion was kept. If the files all contained genomic data of slightly different kinds, then a larger sample size was kept. For example, one folder contained multiple records that held a combination of contig data, oligo data and blast data. An online random number generator (http://numbergenerator.org/) was used to acquire each starting position and sampling progressed from there. Those records that fell outside the sample were deleted from the network drive copy of the hard drive and their corresponding entry was highlighted on the DROID export to show they had been deleted. In the end, 23 folders were sampled (two of these had their sub-folders sampled, the rest had their files sampled). These folders originally contained 5063 files and after sampling they contained 935. Table 1 details the exact sample size adopted for each folder and the number of file retained.
Researchers can still view some of the data in order to experience the way in which Ashburner worked and collaborated with others. They will be unable to view all the genomic data from all of Ashburner’s research projects, but this data is deposited in accessible data-banks elsewhere.

**Appraisal section conclusion**

We feel our appraisal approach maximized efficiency and yielded good results. It is clear that investing time analysing technical elements of digital files can result in a substantial number being weeded, more so than we initially assumed. Technical appraisal can also support functional appraisal where very niche file formats are present. One key factor is the difference the depositor can make. The Dunham and Ashburner hard drives differ not only in size but in the amount and type of organization used. This had a significant impact on the amount of time it took to appraise each drive, despite Ashburner providing information on the top-level folders. The work has shown that as yet there is no complete solution for dealing with quantity, but by adopting certain approaches and fully utilizing the DROID export the problems generated by quantity can be mitigated to an extent. Quantity is also a key problem when attempting to sensitivity review a collection and so this was our next consideration.

**Sensitivity review**

Sensitivity review is one of the more challenging issues faced when processing born-digital records. Unlike paper records, digital ones cannot be quickly rifled through to get a general
overview of content. Files have to be individually opened and examined, making the process more arduous and often unfeasible given the vast quantities of born-digital records commonly deposited. Nevertheless, archivists have a legal and moral responsibility to ensure sensitive information is not released into the public domain and so sensitivity review cannot be ignored. Other archival tasks are increasingly being streamlined through the use of computers and automated processes, but there has not been such success with sensitivity review due to the very nature of sensitivity: it is hard to define and is very dependent on context. Sensitivity is not just a matter of what is said, but can also depend on who said it, when it was said and where. It is very difficult to programme software to deal with these subtleties and archivists cannot rely on key word searching for known sensitive words to identify all problematic records. This leaves archivists with a problem. Sensitivity review cannot be abandoned, but the nature and quantity of digital records makes it very difficult to undertake.

The Council on Library and Information Resources (CLIR) publication *Born Digital: Guidance for Donors, Dealers and Archival Repositories* covers sensitivity and identifies three actions to take when sensitive content is identified: keep the record open, remove it altogether or impose a restriction. Keeping a record open is at best negligent and at worst illegal. Removing records purely because they contain sensitive information goes against basic archival principles. Imposing an access restriction should be done when sensitive material has been identified, but there is a danger that quantity encourages archivists to close huge swathes of records as a precaution. Tim Gollins and colleagues argue that this approach is not ‘morally, ethically or politically acceptable in an era of increasingly open government’ and are trying to tackle these issues with Project Abacá ([https://projectabaca.wordpress.com/tag/project-abaca/](https://projectabaca.wordpress.com/tag/project-abaca/)). This aims to develop a framework for digital sensitivity review and increase the ability of computers to accurately assess records. The project outcomes and their utility remain to be seen and it may take years, but archivists cannot afford to stop catalogue processing work and wait. Sensitivity review needs to be tackled now and archivists will be doing users a disservice if they close records as a precaution until suitable tools are developed.

Whilst archivists have plenty of advice on identifying sensitive data and applying access restrictions, there is little guidance for those struggling to review large volumes of born-digital material, though the CLIR guidance does provide two suggestions of moderate helpfulness. The first is keyword searching of known sensitive terms. As has been mentioned, keyword searching does not address the problem of context when identifying sensitive material. However, it can provide limited help in situations where there are known sensitive topics, events or people likely to be referred to in a set of born-digital records. Thus, it should not be discounted, but should be used with caution. The second suggestion is to utilize the depositor’s knowledge by asking them to highlight potential issues. This can be effective in situations where the depositor is the archival creator and has good knowledge of the deposit. However, there are many situations where this solution may not be appropriate, such as if the depositor is not the creator or if the depositor’s memory is unreliable. Relying on the depositor to highlight problematic records also hangs on the depositor understanding the concept of sensitivity in all its forms, such as personal data, commercial confidentiality and business sensitivity. Both keyword searching and utilizing the depositor have merit and should be used where appropriate, but neither of these methods provides a complete solution on its own and even taken together still leave many gaps. These are gaps the Wellcome Library has attempted to address when sensitivity reviewing the two hard drives.
The first thing to note is that effective appraisal aids sensitivity review by sorting the wheat from the chaff and reducing the total quantity of records requiring review. Appraisal also highlights file formats that are highly unlikely to contain sensitive information, such as UCSC BED Annotation Track files. After assessing the risk, an archivist may feel confident to ignore these records when sensitivity reviewing, thereby again reducing the quantity of records included in the review. It is also recommended to undertake appraisal and sensitivity review in tandem. As has been mentioned, one stage of the appraisal of the hard drives undertaken by the Wellcome Library involved a more granular assessment of a sample of individual files. In this instance, it was efficient to judge them for appraisal purposes and review them for sensitivity at the same time, rather than looking at a record series twice.

Undertaking the sensitivity review of the Dunham and Ashburner hard drives enabled staff at the Wellcome Library to refine our thoughts on the process, to consider the risks associated with too much and too little review and to devise an approach based on risk management. Due in part to the lack of literature on sensitivity review, we looked to other Wellcome Library policies and procedures for inspiration. Of particular use was the section on sensitivity review of digitized material available online within our access policy.30

According to these guidelines, all archival material classed as ‘open’ during the cataloguing process is assigned a risk category based on various criteria and a sample is sensitivity reviewed before the records are made available online. The sample size is determined by the risk category (Table 2).

The sample sizes for risk categories B and C are flexible to take into account the quantity and nature of the specific material. For instance, small quantities and/or records that are ambiguously catalogued may be reviewed in their entirety. In contrast, material that has been catalogued in detail and/or comes in a large quantity may only have a percentage reviewed. When determining a risk category the following are considered, alongside the age of the material and presence of identifiable, living individuals:

- How comprehensive and detailed is the catalogue?
- Do files appear to be very mixed in content or homogenous?
- Was information provided on the understanding that it would be kept confidential?
- How, if at all, is the information structured (e.g. alphabetically)?

This method of sensitivity review is undertaken for digitized material being made available online. It is catalogued, therefore has already undergone some sensitivity review that has identified obvious access restrictions and the catalogue descriptions help determine the appropriate risk category. As such, this exact system is not appropriate for uncatalogued born-digital material. Nevertheless, the concept served as a useful starting point from which to devise a suitable system.

### Table 2. Wellcome Library sensitivity risk categories for digitized material made available online.

| Risk category | General nature                                                                 | Sample size (%) |
|---------------|-------------------------------------------------------------------------------|-----------------|
| A (low)       | • 100+ years old<br>• No living individuals<br>• No risk of damage or distress to family | 0               |
| B (medium)    | • 50–99 years old<br>• Some living individuals<br>• Little risk of damage or distress<br>• Less than 50 years old | Up to 100       |
| C (high)      | • Less than 50 years old<br>• Some living individuals<br>• Potential damage or distress | 50–100          |
Firstly, we retained the three risk categories but modified the criteria and resulting sample size. The criteria were shaped by considering the following questions:

1. How comprehensive and detailed is the accession information, metadata and associated contextual information?

   Detailed information increases the likelihood that the nature and content of records can be identified without having to view them. Accession information and metadata can either highlight potentially sensitive records or indicate that they are low risk. A lack of information provides no such assurance, so increases the risk. Contextual information can include disk labels, information provided by the depositor, and file directory names.

2. How is the information structured, how homogenous are the record contents?

   Records kept in well-organized directories are easier to sift. Clearly non-sensitive sub-folders can be passed over and attention paid to more suspicious or obscure sub-folders. A lack of discernible file structure means any file could be anywhere and so everywhere needs to be assessed. Homogenous sub-folders also reduce risk: if a set of records are very similar only a few needed to be checked to be confident of the appropriate access status. Very varied content means such assumptions cannot be made and so a greater sample needs to be reviewed.

3. Do folder and/or file names raise suspicions?

   Certain words might be known as likely relating to sensitive content. For instance, ‘grants’ can indicate grant applications and the names of particular individuals or the use of known controversial subject terms can highlight potentially sensitive content.

4. Do file formats seem non-sensitive?

   As previously mentioned, some file formats are highly unlikely to contain sensitive information due to the nature of the format and the data it holds.

   After considering these four questions, each set of records was allocated one of three risk categories (Table 3).

   The term ‘up to’ a certain percentage is used, partly to retain the flexibility regarding quantity, as previously mentioned, but also to allow for some nuance within the categories. For instance, two sets of records could be assigned category B status but sit at either end of the spectrum: one set almost meeting category A criteria and the other nearly but not quite meeting category C criteria. More importantly, reviewing ‘up to’ a certain percentage takes into account the fact that this approach works on a model whereby review stops once the highest level of access restriction is found (closure for 100 years in accordance with the Data Protection Act 1998). If the first record reviewed requires such a closure there is little point reviewing the remaining records as nothing more excessive will be found.

   Before this approach could be applied and tested, we had to decide on the level at which closure would apply. This is done at file level (the deliverable unit) for paper records: all records within a particular file are either open or closed, we do not close fractions of a file. The hard drives did not have an obvious deliverable unit level, so this needed to be decided upon. One option was to make the entire hard drive the deliverable unit, but this would have been too large and unwieldy to serve as a useful deliverable unit for researchers. Moreover, any access restriction would result in the entire hard drive being closed, barring access to a lot of rich research material. This seemed disproportionate, much like closing an entire archive. Making each individual file a deliverable unit was also rejected as being impractical as it would require each digital file to be individually reviewed. A middle level had to be found that struck the
balance between breaking down the hard drive into deliverable units of a manageable size and encompassing large enough groups of digital records to make sensitivity review feasible. This was achieved by analysing the directory structure of each hard drive. Dunham’s drive contained 82 top-level folders, some of which had sub-folders but the majority did not. These top-level folders seemed to form a natural deliverable unit and effective level at which to apply access restrictions. In contrast, the Ashburner hard drive was much more hierarchical. There were 22 top-level folders containing several layers of sub-folder: one top-level folder contained 42 sub-folders and 121 folders within those, holding over 900 digital files in total. In this instance, the balance between a manageable deliverable unit and feasible sensitivity review was not at the top-folder level, but one step down at sub-folder level.

Having determined the deliverable units, a sensitivity risk category could be assigned. The records contained within the deliverable unit were then sampled for sensitivity review, prioritizing those files that looked most likely to contain sensitive information and using a sample size appropriate to the risk category. We took the view that sensitive content requiring a lifetime closure was very unlikely to be found on either hard drive. This was based on knowledge of the accompanying paper archives, which contained very little of this type of sensitive content, and indications of hard drive content from the directory structures. Both hard drives were used in a professional capacity, neither appeared to contain personal material and the nature of their work did not involve dealing with sensitive topics, such as human subjects. It was felt that the highest closure period to be required on the hard drives was a 60-year closure to protect careers and professional reputations. Therefore, sensitivity review of a particular deliverable unit halted when this type of sensitivity was identified. Where sensitive material was found that warranted an access restriction of less than 60 years, the deliverable unit continued to be reviewed until either something requiring a 60-year closure was found, or the appropriate percentage of the deliverable unit had been reviewed. In that instance, the strictest closure period identified was then applied. In this way, both hard drives were sensitivity reviewed in a manageable way that did not resort to reviewing every individual file.

This is a risk management approach to sensitivity review that strikes a balance between making sweeping access decisions across large quantities of records and individually assessing each digital file. There is a risk the Wellcome Library will accidentally provide access to
sensitive content, but the likelihood of this happening is at a level we are comfortable with, due to the steps we have taken to mitigate the risk. Given time, projects like Project Abacá may lead to the process becoming more automated and this risk management approach may not be required. But until that day comes, the Wellcome Library has devised an approach to sensitivity review that allows us to continue processing born-digital archives and making appropriate records available to current researchers.

**Conclusion**

Appraisal and sensitivity review are not straightforward tasks in the digital sphere, but then they are not always straightforward in the paper world. The technical dimension and vast size of many digital deposits add a new layer of complexity, but as this paper has shown, they are not insurmountable obstacles to effective archival processing. Transplanting traditional, paper practices to digital records is in many ways effective and serves to breakdown mental barriers many archivists have with regards to digital archives. Functional appraisal, making full use of the depositor’s knowledge of the archive and sensitivity reviewing at deliverable unit level are all practices used by the Wellcome Library to deal with paper records and these have translated well into the digital sphere. They have also been shown to be scalable, as they proved effective in dealing with two hard drives that differed substantially in size and complexity. Nevertheless, archivists should not be limited to only devising digital equivalents to paper practice. The fundamental difference between digital and paper is technical construction and whilst this can add complexity to tasks like appraisal and sensitivity review, it can also provide solutions, so long as archivists broaden their thinking and consider appraisal and sensitivity review from different perspectives. At the Wellcome Library we found DROID to be very useful, not only as a file profiling tool, but also as a means to easily identify duplicates, achieve a good overview of directory structures and weed out unwanted file formats. Technological tools were not so helpful when sensitivity reviewing, though we found that technical metadata can help identify levels of risk.

This paper has also shown that the key to effective appraisal and sensitivity review of born-digital archives is to find a balance between two extremes: making sweeping decisions about large batches of records and undertaking very granular assessment. This balance is what the Wellcome Library has attempted to find through our work on the Dunham and Ashburner hard drives and hopefully we have shown that it is a realistic expectation that all archival institutions can follow. That is not to say the Wellcome Library has solved appraisal and sensitivity review. Each new deposit will allow us to test our procedures, refine our ideas and will undoubtedly throw up new problems we have not yet encountered. For the time being, we have developed practices that make appraisal and sensitivity review as effective as possible to the best of our current ability and hope this paper will stimulate discussion amongst the archival profession, leading to improved practices in future.

**Notes**

1. Ian Dunham: archives, Wellcome Library, [http://archives.wellcomelibrary.org/DServe/dserve.exe?dsqIni=Dserve.ini&dsqApp=Archive&dsqCmd=Show.tcl&dsqDb=Catalog&dsqPos=0&dsqSearch=%28AltRefNo%3D%27grl%2Fidn%27%29.](http://archives.wellcomelibrary.org/DServe/dserve.exe?dsqIni=Dserve.ini&dsqApp=Archive&dsqCmd=Show.tcl&dsqDb=Catalog&dsqPos=0&dsqSearch=%28AltRefNo%3D%27grl%2Fidn%27%29)

2. Michael Ashburner: archives, Wellcome Library, [http://archives.wellcomelibrary.org/DServe/dserve.exe?dsqIni=Dserve.ini&dsqApp=Archive&dsqCmd=Show.tcl&dsqDb=Catalog&dsqPos=0&dsqSearch=%28AltRefNo%3D%27pp%2Fmia%27%29.](http://archives.wellcomelibrary.org/DServe/dserve.exe?dsqIni=Dserve.ini&dsqApp=Archive&dsqCmd=Show.tcl&dsqDb=Catalog&dsqPos=0&dsqSearch=%28AltRefNo%3D%27pp%2Fmia%27%29)
3. Cook, “Archival Appraisal Past, Present,” 5.
4. Paradigm, “Appraising Digital Records,” 4.
5. 4C, “Digital Curation Sustainability Model,” 6.
6. Eastwood, “Digital Appraisal: Variations,” 1.
7. Paradigm, “Recommended Approaches to Appraising,” 4.
8. Niu, “Appraisal and Selection,” 69–77; Ross, “Instalment on Appraisal,” 26; and InterPARES, Appraisal Task Force Report, 8–14.
9. Niu, “Appraisal and Selection,” 69.
10. Ibid.; and InterPARES, Appraisal Task Force Report, 8.
11. Paradigm, “Recommended Approaches to Appraising,” 4.
12. Paradigm, “Practical Solutions,” 4; and Paradigm, “Useful Appraisal Tools,” 4.
13. Paradigm, “Practical Solutions,” 4.
14. Paradigm, “Recommended Approaches to Appraising,” 4.
15. File Profiling Tool (DROID), http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.
16. Hull, Use of Sampling Techniques, 6; Kolish, “Sampling Methodology and Its Application,” 62; and Kepley, “Sampling in Archives,” 239.
17. Cook, “Appraisal Guidelines for Sampling,” 27.
18. Kepley, “Sampling in Archives,” 239.
19. Ibid.
20. Kolish, “Sampling Methodology and its Application,” 62.
21. Ibid., 63.
22. Cook, “Appraisal Guidelines for Sampling,” 27–8.
23. Ibid., 38.
24. Kepley, “Sampling in Archives,” 240.
25. Redwine et al., Born Digital: Guidance for Donors, 8.
26. Gollins et al., “Selection and Sensitivity Review,” 4.
27. National Archives, “Step 3: Sensitivity Reviews”.
28. Redwine et al., Born Digital: Guidance for Donors, 7.
29. Ibid.
30. Wellcome Library, “Assessment Prior to Online Publication,” 14–21.

Acknowledgements
The author wishes to thank her Wellcome Library colleagues for their help and support in writing this paper.

ORCID
Victoria Sloyan http://orcid.org/0000-0003-1077-7625

References
Cook, Terry. “‘Many Are Called, but Few Are Chosen’: Appraisal Guidelines for Sampling and Selecting Case Files.” Archivaria 32 (1991): 25–50. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11759/12709.
Cook, Terry. “‘We Are What We Keep; We Keep What We Are’: Archival Appraisal Past, Present and Future.” Journal of the Society of Archivists 32, no. 2 (2011): 173–189. doi:http://dx.doi.org/10.1080/00379816.2011.619688.
Eastwood, Terry. “Digital Appraisal: Variations on a Theme.” Paper presented at the Conference on Appraisal in the Digital World, Rome, November 15–16, 2007.
4C – Collaboration to Clarify the Cost of Curation. The Digital Curation Sustainability Model (DCSM). Accessed November 6, 2015. http://www.4cproject.eu/dcsm
Gollins, Timothy, Graham McDonald, Craig Macdonald, and Iadh Ounis. “On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records.” Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security (PIR 2014), Australia, July 11, 2014. http://ceur-ws.org/Vol-1225/pir2014_submission_9.pdf.

Harvey, Ross. "Instalment on 'Appraisal and Selection." DCC Digital Curation Manual, 2007. http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/appraisal-and-selection/appraisal-and-selection.pdf.

Hull, Felix. The Use of Sampling Techniques in the Retention of Records. Accessed June 25, 2015. http://www.unesco.org/webworld/ramp/html/r9006e/r9006e04.htm

InterPARES. Appraisal Task Force Report. Accessed June 25, 2015. http://www.interpares.org/book/interpares_book_e_part2.pdf

Kepley, David R. “Sampling in Archives: A Review.” The American Archivist 47, no. 3 (1984): 237–242. http://americanarchivist.org/doi/pdf/10.17723/aarc.47.3.kj041746214p6u1v.

Kolish, Evelyn. “Sampling Methodology and Its Application: An Illustration of the Tension Between Theory and Practice." Archivaria 38 (1994): 61–73.

The National Archives. “Step 3: Sensitivity Reviews of Selected Records.” Accessed June 25, 2015. http://www.nationalarchives.gov.uk/information-management/manage-information/selection-and-transfer/sensitivity-reviews-on-selected-records/.

Niu, Jinfang. “Appraisal and Selection for Digital Curation.” International Journal of Digital Curation 9, no. 2 (2014): 65–82. doi:http://dx.doi.org/10.2218/ijdc.v9i2.267.

Paradigm. “Appraising Digital Records: A Worthwhile Exercise?” Chap. 4 in Workbook on Digital Private Papers. 2008. http://www.paradigm.ac.uk/workbook/appraisal/digital-appraisal.html.

Paradigm. “Recommended Approaches to Appraising Personal Archives.” Chap. 4 in Workbook on Digital Private Papers. 2008. http://www.paradigm.ac.uk/workbook/appraisal/appraisal-approaches.html.

Paradigm. “Practical Solutions Advocated by the Paradigm Project.” Chap. 4 in Workbook on Digital Private Papers. 2008. http://www.paradigm.ac.uk/workbook/appraisal/practical-appraisal.html.

Paradigm. “Useful Appraisal Tools.” Chap. 4 in Workbook on Digital Private Papers. 2008. http://www.paradigm.ac.uk/workbook/appraisal/directory-tools.html.

Redwine, Gabriela, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, and Will Hansen. Jeremy Leighton John, Nancy Kuhl, Seth Shaw and Susan Thomas, Born Digital: Guidance for Donors, Dealers, and Archival Repositories. Washington, DC: Council on Library and Information Resources, 2013. http://www.clir.org/pubs/reports/pub159/pub159.pdf.

Wellcome Library. “2.3 Phase 3: Assessment Prior to Online Publication.” Access to Personal Data within Our Research Collections, 14–21. London: Wellcome Trust, 2014. http://wellcomelibrary.org/content/documents/policy-documents/access-to-personal-data.pdf.