ARCHIVES AND MANUSCRIPTS
2019, VOL. 47, NO. 3, 305–326
https://doi.org/10.1080/01576895.2019.1622138

Routledge
Taylor & Francis Group

ARTICLE

Appraising, processing, and providing access to email in contemporary literary archives

J. Schneider[a], C. Adams[b], S. DeBauche[a], R. Echols[b], C. McKean[c], J. Moran[d] and D. Waugh[e]

[a]Stanford University Libraries, Stanford University, Stanford, California, USA; [b]Harry Ransom Centre, University of Texas at Austin, Austin, Texas, USA; [c]British Library, London, UK; [d]Alexander Turnbull Library, National Library of New Zealand, Wellington, New Zealand; [e]Stuart A. Rose Manuscript, Archives, and Rare Book Library, Emory University, Atlanta, Georgia, USA

ABSTRACT

The email of contemporary literary figures is ripe for research by scholars, and of broad interest to the general public, but can also present many challenges to cultural memory institutions that seek to appraise, process and provide access to this rich archival material. This article explores how five institutions across the United States, United Kingdom and New Zealand are using ePADD, free and open source software developed by Stanford Libraries that incorporates artificial intelligence and machine learning to help meet these challenges for email in contemporary English-language literary collections. Authors and institutions represented include British poet Wendy Cope (The British Library), British novelist Ian McEwan (Harry Ransom Center, University of Texas at Austin), British Indian novelist and essayist Salman Rushdie (Emory University), American poet Robert Creeley (Stanford University) and New Zealand poet and critic Ian Wedde (National Library of New Zealand). The use cases are followed by a discussion identifying lessons learned and areas for further research.

KEYWORDS

Email preservation; archival processing; contemporary literary archives; natural language processing; machine learning

Introduction

Email offers singular insight into and evidence of an individual's self-expression, as well as records of their collaboration, networks and transactions. Email communications of prominent individuals, including writers, politicians, scientists and scholars, reveal not only their professional and personal actions, decisions, and creative output, but also relationships within society and communities. The appeal of email collections extends beyond historians to all manner of researchers, journalists, and the general public. Archives have long recognised the documentary value of correspondence, and as email has become the primary form of professional and often personal communication, archival institutions of all types are increasingly seeking to acquire email collections.

Yet collecting email is difficult and poses a host of logistical and ethical challenges. Archives have struggled to develop reproducible strategies and methods to both screen email for sensitive, confidential or legally restricted information, and provide effective access

CONTACT Josh Schneider jschne@stanford.edu

to email, due to the sheer volume and complexity of the material. For these reasons, in 2016, a Task Force on Technical Approaches to Email Archives was formed with the support of the Andrew W. Mellon Foundation and the Digital Preservation Coalition; the group's 2018 report, published by the Council on Library and Information Resources, explores many of these challenges, and strategies to overcoming them, in depth.[1]

ePADD is free and open-source software designed specifically to address these challenges for large volumes of email of potential historical or cultural value, and in a manner that is both customisable and scalable.[2] The software accomplishes these goals by incorporating techniques from computer science and computational linguistics, including machine learning, natural language processing and named entity recognition.

This article begins with an introduction to ePADD, including an overview of the history of the project and the core functionality included in the software. Following that introduction are five use cases demonstrating how institutions in the United States, United Kingdom and New Zealand are using ePADD to support archival workflows for email in contemporary English-language literary collections. Authors and institutions represented include British poet Wendy Cope (The British Library), British novelist Ian McEwan (Harry Ransom Center, University of Texas at Austin), British Indian novelist and essayist Salman Rushdie (Emory University), American poet Robert Creeley (Stanford University) and New Zealand poet and critic Ian Wedde (National Library of New Zealand). The use cases are followed by a brief discussion reiterating lessons learned and areas for further research.

About the ePADD software

ePADD is developed by Stanford Libraries' Department of Special Collections & University Archives and partners,[3] with funding provided by the US Institute of Museum and Library Services (2015-2018), US National Historical Publications & Records Commission (2012-2015) and Stanford Libraries (2012-current).

Over the past six years, ePADD has pioneered and refined the application of machine learning and natural language processing to confront the challenges inherent in donating, administering, preserving, or accessing email collections. These include screening email for confidential, restricted, or legally protected information, preparing email for preservation, and making the resulting files (which incorporate preservation actions taken by the repository) discoverable and accessible to researchers. ePADD offers an intuitive and user-friendly interface, ensuring that collection donors, archivists and researchers can use the tool effectively without exhaustive training or technical expertise.

ePADD incorporates several automated functionalities that simplify screening and optimisation of access to an email archive's intellectual content. These processes, which takes place during the initial import of email into ePADD, support all subsequent activities undertaken by users.

First, ePADD resolves names and email addresses associated with a single correspondent. Resolved correspondent names can be browsed and graphed alphabetically or by volume of messages exchanged with the email account holder. Second, ePADD employs a custom fine-grained named entity recogniser that extracts categories of entities bootstrapped from DBpedia. These include persons, organisations, locations, government entities, political parties, companies, universities, diseases and awards. Extracted entities can be browsed alphabetically or by volume of messages. Named

entity recognition effectively converts unstructured data (the email message body) into structured data which can be queried and browsed.

ePADD also provides email creators and the staff of memory institutions with an integrated toolset to efficiently carry out appraisal and processing of email in archival collections (via the **Appraisal module** and **Processing module**, respectively). These tools include keyword search, advanced search, regular expression search, a customisable 'lexicon search', which enables tiered thematic searching across several categories, 'multi-entity search', to aid in comparative entity analysis between the archive and any other textual corpus, correspondent list search, image attachment browsing, mailing list identification, graphical visualisation, user annotation and an interface for assigning authorised headings to correspondents (see Figure 1). Messages can also be annotated and tagged with labels indicating terms of restriction (see Figure 2).

ePADD's automated functionalities, including the resolution of correspondent names and recognition of fine-grained entities, also directly benefit researchers. Discovery of email in archival collections is another challenge, since email is not often made available online to the public due to donor and third-party privacy and copyright concerns. As description made available online in archival finding aids and catalogue records is often very limited, it can be difficult for researchers to determine if they should travel to the host institution to view the materials in person.

To address these issues, ePADD introduced an online **Discovery module**, which enables researchers to access a redacted version of the email archive via a public web server prior to visiting the repository (see Figure 3). This conceit inverts the traditional model of redaction

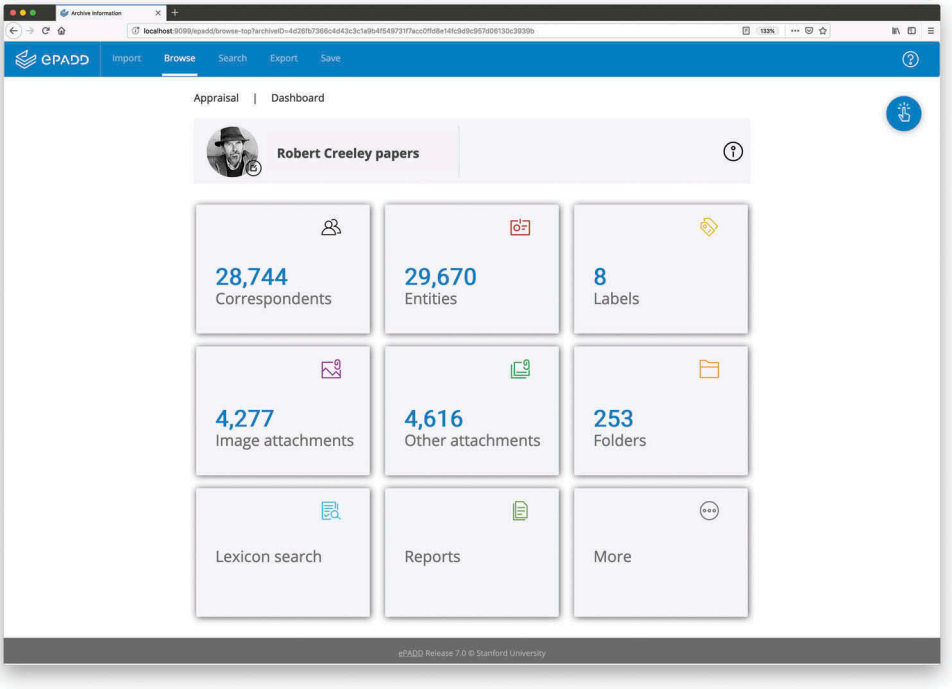

Figure 1. ePADD v.7 Appraisal Module – Browse Menu.

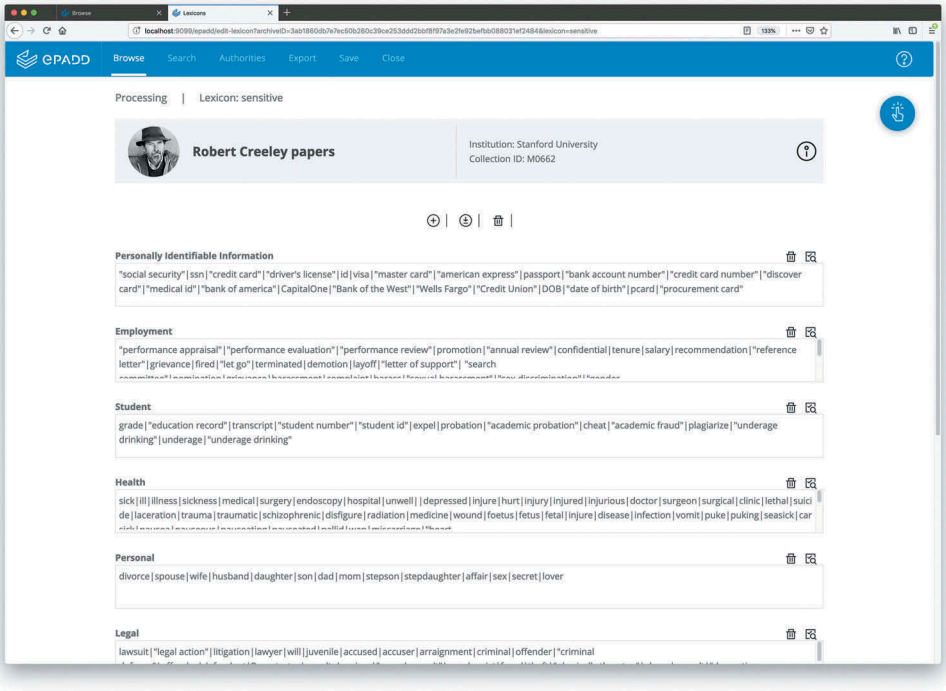

Figure 2. ePADD v.7 Processing Module – Sensitive Lexicon.

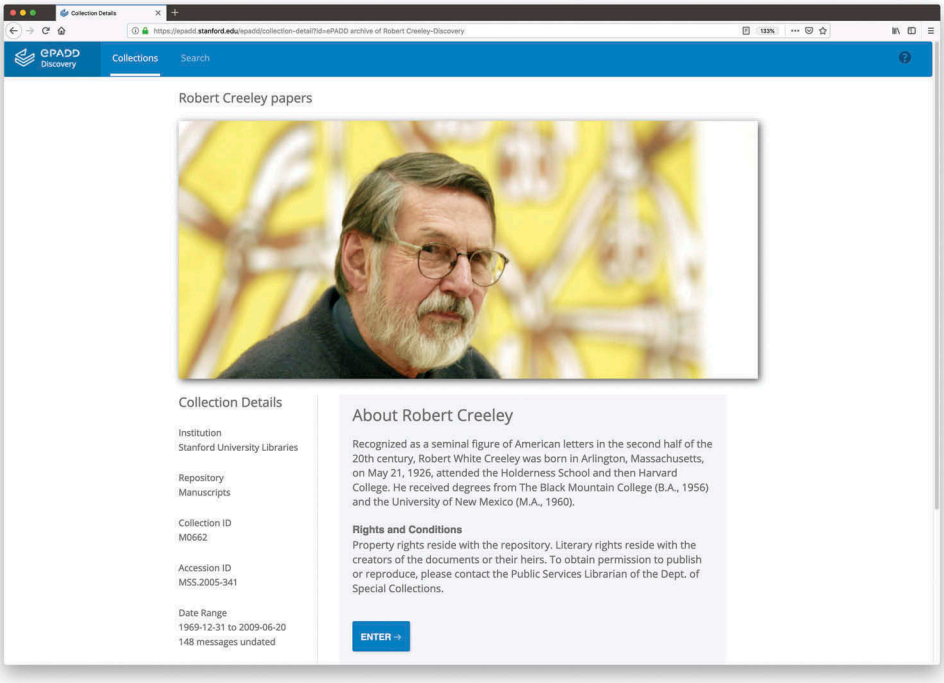

Figure 3. ePADD v.7 Discovery Module – Landing Page.

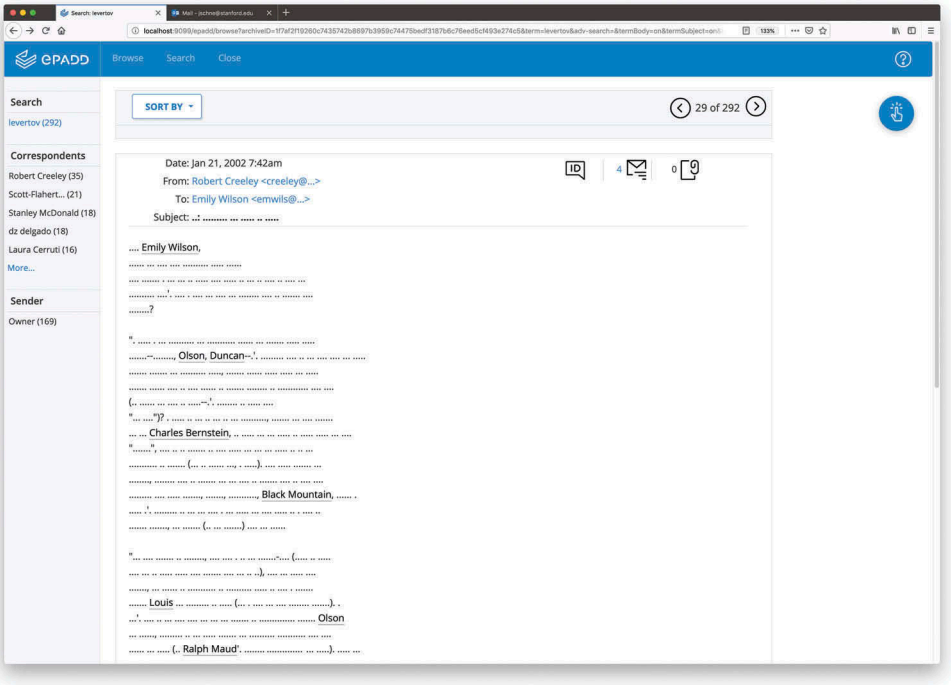

Figure 4. ePADD v.7 Discovery Module – Message Browsing.

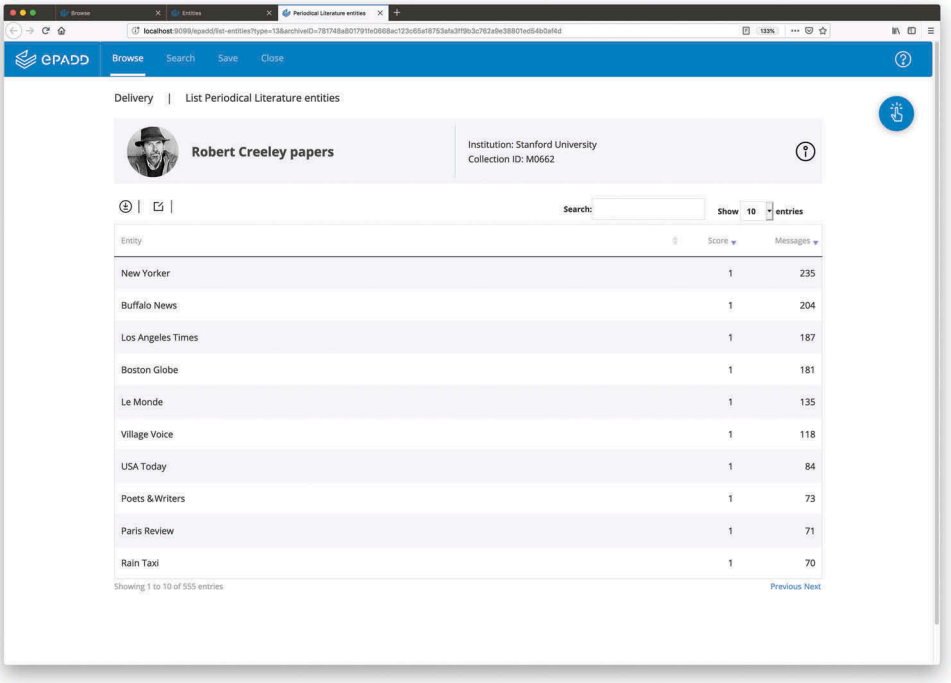

Figure 5. ePADD v.7 Delivery Module – Periodical Literature entities.

by displaying only the correspondents and named entities, and relies heavily upon ePADD's automated resolution of correspondent names and extraction and classification of fine-grained entities. The Discovery module enables even those remote users who are unfamiliar with a collection to browse across it via these same correspondents and entities (see Figure 4).

Once a researcher has discovered the email archive and decided to view it, a repository can use ePADD to enable access to the full archive of unrestricted messages, including attachments, via a **Delivery module** that offers many of the same functionalities included in other modules (see Figure 5). ePADD also supports a variety of exports of controlled data from the archive, to support accessing collections as data, including topic modelling and network analysis.

Case studies

ePADD enables a broad and diverse constituency of libraries, archives and museums to begin to collect, preserve and provide access to vital cultural heritage materials that might otherwise be unavailable for research. The five case studies below will describe strategies that archivists have employed to process email collections, and the functions within ePADD that support those strategies. Please note that because the software is in active development, and these case studies took place over several years, there may be some inconsistency in how the tooling is described due to differences between releases; discussion of migration between platforms is also common.

Case study 1: Wendy Cope's email, the British Library

Wendy Cope (b. 1945) is a British poet known for poetry collections such as *Making Cocoa for Kingsley Amis* (Faber & Faber, 1986) and *Serious Concerns* (Faber & Faber, 2002). She was awarded an OBE (Order of the British Empire) in 2010. Cope's archive was purchased by the British Library in 2011. The born-digital component of the archive – including around 25,000 emails - remains uncatalogued as of writing. This case study reflects work performed by Callum McKean using ePADD v6.1 in 2018.

When the Library acquired Cope's archive in 2011, the importance of collecting literary email archives was not in dispute, but consensus around strategies for acquiring, arranging and providing access to this material proved more elusive. Despite stringent data protection controls in place at the Library, Cope was understandably worried about the consequences of making this vast collection of intimate material available to the public. In order to assuage these concerns, Library colleagues offered to help Cope to vet to her emails using her desktop-based email client (Outlook) and her web-based email client (BT Internet).

This process proved to be time-consuming and frustrating. For both web and desktop email clients, the archival function is incidental. Search functionality is optimised for specificity and speedy retrieval of recent messages, with very limited options for browsing and visualisation. These design choices placed the onus on Cope to remember conversations which she had years prior, and to know how to search for them effectively, or to browse through her entire archive in list form. Both options were logistically and physically straining for her.

If ePADD had been available in 2011, most of these problems could have been mitigated: entity browsing would have allowed Cope to remove particular entities deemed sensitive as a whole, such as those of a purely financial or legal nature; 'lexicon' search would have eliminated the need for highly specific single-search terms and allowed for more 'woolly' searching around personalisable subjects, like health and children; a versatile labelling system would have provided her with additional assurance that no important material had been passed over; and image browsing would have allowed her to identify private visual material without the need to open every attachment individually. In addition to these technical benefits, the continuity of ePADD's user experience from ingest to access would have allowed Cope to visualise her email as an archival collection rather than a live work environment.

Despite some difficulties, Cope's email archive was eventually acquired as a legacy PST file on a USB Flash Drive, which was then backed up to a secure server and stored physically in the Library's eMSS Lab. The Library prefers to acquire email collections as discreet files (usually MBOX or a file-type easily convertible to MBOX)[4] with a commitment to future accruals, rather than acquiring IMAP credentials from donors. This approach maximises continuity with the Library's existing workflows for other born-digital material[5] and creates minimal friction in the context of the European Union's General Data Protection Regulation legislation (GDPR), which makes the collection of live data like IMAP credentials more complicated with no added benefit for our purposes.[6] Another advantage of MBOX is that ingesting these files using ePADD is intuitive, providing colleagues with quick access to these collections for internal communications purposes, allowing us to build a profile for email collections in the department, or in the Library as a whole.

GDPR impacts the appraisal and processing of archival material at the Library, as well as its ingest. Although allowances are made within GDPR for material made available 'in the public interest', what this means precisely remains a matter of discretion for archivists and curators (and the process of using that discretion effectively can be incredibly time-consuming). Lexicon searching and customisable labels in ePADD represent invaluable tools in this legal context.

Labelling allows the archivist to move through the collection quickly, and to identify problematic material and highlight it, for example, using a traffic light system. Used in conjunction with Lexicon searches, which can be customised to leverage the archivist's deep knowledge of the archive to greatest effect, it is possible to make the process of Data Protection compliance checking much more efficient.

The Library makes born-digital material available to researchers through a 'call-up' mechanism in the Reading Room, where PDF/As are displayed in a secure, read-only environment. Email archives, being voluminous and threaded, problematise this delivery method which privileges the discrete 'orderable' unit, analogous to the paper file. The Delivery module in ePADD retains the structure and form of email messages in the context of a collection with threads. Because research interest in Cope's email spans disciplines, from traditional literary critics, academics and biographers to those in emerging fields, such as the study of email itself as a communications medium, ePADD gives researchers in the Delivery module freedom to approach and engage with the collection as a whole or from any particular angle, without requiring the user to make individual requests for each email through the Library's online ordering

system, *Explore the British Library*. Equally, the Library's increasingly international community of researchers, who may have accessed email collections through ePADD in other repositories, will not be required to learn a new system.

Overall, ePADD provides a means of guiding email collections – which have very particular requirements, even in the already specialised context of born-digital material – from ingest to access. In the case of Wendy Cope at the Library, it has allowed us to think critically about how email was explained to and acquired from donors in the past, and to envision a future where an international workflow exists to provide access to email collections as a matter of course.

Case study 2: Ian Mcewan's email, Harry Ransom Center, University of Texas at Austin

Ian McEwan (b. 1948) has become one of the most influential British contemporary writers. McEwan's writings cover a range of creative output including short stories; essays; lectures; scripts for radio, television and stage; screenplays; and libretti. His novels, including *Atonement* (Jonathan Cape, 2001), have garnered great critical and popular acclaim, and many have been adapted into successful films. McEwan frequently contributes to social and political discourse including issues related to censorship, climate change, science and the humanities. His archive was acquired by the Harry Ransom Center, a humanities research library and museum at the University of Texas at Austin, in 2014. This case study represents work performed by the Ransom Center's Digital Archivist, Chance Adams, and Processing Assistant, Reid Echols, using ePADD v4.1-5.1 in 2018.

McEwan's papers contain manuscript and typescript drafts, personal and professional correspondence, notebooks, photographs, clippings and family papers. In addition to the 71 document boxes of physical material, the collection also includes one computer hard drive containing document files, email correspondence and photographs; six compact discs; and four 3.5-inch floppy diskettes. Many of McEwan's writings (including novels, screenplays, essays, and lectures, as well as unidentified documents), outgoing correspondence, photographs, and personal and professional documents exist in born-digital form and are available to onsite researchers.

The Ransom Center acquired Ian McEwan's born-digital records alongside the physical collection, which in addition to the materials described above include an extensive archive of email correspondence. This email archive consists of approximately eighty thousand emails (9.4 GB) Apple Mail email messages spanning the years 1997 to 2014. McEwan's technical assistant filed the emails separately as sent and received, broken down by year initially and quarterly as volumes increased, and this file structure has been preserved. Once McEwan's born-digital drafts were processed and made available to researchers, we began processing the email archive in the summer of 2017. At over thirteen thousand correspondents, this task proved overwhelming, given the prospect of refining thousands of names formatted as a single column list with groupings separated by double hyphens.[7]

After several failed attempts at transforming the data into a more user-friendly format that could be manipulated in bulk with the data clean-up tool OpenRefine, we consulted with the ePADD developers and archivists at Stanford to find out what strategies they employed for this step in the appraisal workflow. Once we discovered

they had focused primarily on only cleaning up the top tier of correspondents based on message volume, we decided to change tactics and concentrate on the first few hundred names, sorted in the browse screen by incoming message volume. Since then, the correspondents list has been edited and the bulk of these issues were manually resolved. Fortunately, this list is saved as a separate text file, which can be preserved and restored over any upgrades or new installations: a practice certainly recommended. It is worth noting that the most common errors seem to be derived from inconsistent reading of unicode characters associated with non-English languages and names. A discussion with the ePADD developers suggests they are still working to resolve these encoding issues, though many of them, no doubt, are unavoidable, based on converting legacy files from defunct operating systems.

In February 2018, Reid Echols joined the project as a processing assistant, with the primary goal of assessing the email archive and identifying messages for restriction. The creator had stipulated that material regarding specific correspondents and some topics of a personal nature be restricted during his lifetime and that of his wife, in addition to any personal identifiable information (PII) still present. While it appears that his assistant made an initial assessment of these emails prior to accession, McEwan's requests for restriction still necessitated a substantial evaluation of the collection for sensitive and confidential data before making it available to researchers, as well as the development of a clear access policy governing the responsible use and reproduction of email correspondence. Essentially, since the accession was made before adopting ePADD, work began in the appraisal module to flag emails for export, restriction, or temporary embargo. Ideally, the workflow would allow the creator to use ePADD to appraise their own correspondence prior to transfer, but in this case that was not possible.

For the first several weeks, we focused on building robust lexicon searches within ePADD aimed at identifying sensitive or confidential information. The first step was to run searches for PII such as passport, social security and credit card numbers in addition to other financial accounts. We had limited success using common regular expressions (regex) for certain sequences of numbers (US social security numbers, credit cards, and so on), attributed largely to the international scope of McEwan's archive. Different government standards meant that results were found for UK tax identification numbers (TINs) or business registries that overlapped with the social security number (SSN) sequence, or conversely that several US phone numbers showed up in searches for UK passport numbers. The extraneous hits were significant enough to prompt a more strategic approach, in which several keyword searches were performed and a list of McEwan's actual personal information (US TIN, McEwan's and his immediate family's passports, and so on) was compiled. This allowed us to quickly identify and batch-restrict emails containing these number strings.

This approach would be particularly useful for creators preparing their own archives for processing, as they will be most aware of their own sensitive data, but if archivists find themselves needing to reverse-engineer the process, they should plan on a significant amount of time spent researching the creator and the collection, or on outlining a dossier of personal data (including sensitive number strings, nicknames, and so on) for the creator to generate and include with the accession. ePADD's search functions and entity analysis proved invaluable in performing this research, as they likely would for a scholar examining the collection for less sensitive data.

After identifying and flagging sensitive personal information, we moved on to restricting materials based on the creator's requests, which was a far more nuanced and involved process. Since McEwan's email archive was appraised by his personal assistant prior to delivery, confidential messages that could have been easily found and restricted in bulk had been removed from the original acquisition: a simple correspondent search found no results for one of the individuals flagged for restriction, for example. However, a deeper search revealed that many emails that referred to this individual, legal proceedings, or private reflections on what had transpired – all of which McEwan had requested be restricted – still remained in the archive. This is where we spent the bulk of our time, as many of the search threads required actually reading the full text of emails that appeared in search results.

Common nicknames or recurring correspondents were noted, which contributed to narrowing the search results: emails with certain correspondents often dealt with specific personal issues, for example. It is worth pointing out that the typical human reticence when discussing sensitive topics, the tendency to use euphemisms, and the idiosyncrasy of individual speech all seem to be a major roadblock to any attempt at automating this process via keyword search. During this phase of the project, we maintained a log of pertinent terms and correspondents to perform searches for, always remembering to select 'Not Reviewed' from the filters on the left-hand side of the page when presented with results to avoid covering the same ground.

After completing this process, flagging nearly a thousand emails for restriction, we imported the email archive into the processing module to add metadata and began assessing name authorities, cleaning up the correspondents list and preparing the collection for onsite access. Here we ran into an issue with unicode characters described above, and contacted the ePADD developers to see what was causing the problem. As it turned out, the issue was addressed in ePADD v5.1, which was soon to be released, so we worked with Peter Chan, Digital Archivist at Stanford Libraries, and ePADD Project Manager, to migrate the archives and upgrade the Center's installation of ePADD v4.1 to the new version. In what was later determined to be an issue with the operating system encoding (a test version of v5.1 was installed on a Mac laptop, and v4.1 had been installed on a Windows workstation), several errors in correspondent formatting and the transfer of annotations (now called 'Labels' in v5.1) were discovered. However, with Chan's assistance, these issues were eventually resolved, and would not have occurred if v5.1 had been available to install from the beginning. To date, the archive, edited correspondents list, and all restriction labels are successfully implemented within the v5.1 processing module, being prepared for onsite delivery to patrons.

The ePADD delivery module and McEwan's email archive will be available to researchers onsite in the Ransom Center's Reading and Viewing Room in 2019.[8] At this point, there are no plans to implement the ePADD discovery module. ePADD will be served up on a performance desktop computer with a Windows operating system. Wireless access and downloads will be disabled. Based on user demand, additional workstations may be installed in the Reading and Viewing Room.

Currently, the Center requires researchers viewing any born-digital collection material to sign a separate use policy that restricts downloads, printouts, screenshots, forensic images and original media in addition to requiring users to submit a notification of intent to use the materials. The general use policy for all collection materials includes a statement that materials may contain sensitive or confidential information protected under federal or state

laws and regulations, and researchers are advised that disclosing certain information pertaining to living individuals may have legal implications. Differing or additional language may be needed for researchers accessing the ePADD delivery module due to the presence of email addresses for living individuals, many who are well known in the arts and entertainment industries, and the possibility of unidentified confidential and private subject matter.

ePADD is an incredibly powerful tool for managing the overwhelming size and scope of data in McEwan's email archive. The automatic entity analysis, even with the additional labour of examining the correspondents lists, provided a robust and intuitive means of engaging with the collection.

The following notes on this specific use case are predicated on appraising a collection post-accession:

When presented with the specific challenges of restricting personal materials, while automation and targeted lexicon searches are useful, they cannot substitute the close attention of a processing archivist, particularly when the items flagged for restriction go beyond the usual scope of personal data. Locating and flagging restricted items in email archives like McEwan's relies on many of the same skills needed by cataloguers working in physical archives: namely, careful research into the creator and their biography, and close attention to emails flagged by search results.

ePADD provides a variety of preset options (with common terms for pharmaceutical and recreational drugs, for example), but creating custom lexicons based on the individual creator's unique profile was necessary. The software is extremely effective at detecting recipients, but it provided less help in searching for references in the content of emails (entity searches, while incredibly useful, do not always capture things like nicknames, or informal signatures).

Also, it is likely that the conversational tone of email correspondence may render many common keyword searches less effective than desired: few people refer to close friends by their full names, for example (that is 'Hitch' for Christopher Hitchens, 'Ish' for Kazuo Ishiguro, and so on). The exception to this is professional correspondence with lawyers, agents and the like who typically employ full signatures and formal language, so proper name searches here proved useful. By and large, however, locating restricted material in this massive archive has been most effective after performing some preliminary research and building a custom lexicon for the particular creator, adding to this lexicon as the project progressed.

Case study 3: Salman Rushdie's email, Emory University

Salman Rushdie (b. 1947) is a British Indian novelist and essayist. His second novel, *Midnight's Children* (Jonathan Cape, 1981), won the Booker Prize in 1981. Rushdie's archive was acquired by the Stuart A. Rose Manuscript, Archives, and Rare Book Library at Emory University in late 2006. In addition to more than one hundred linear feet of traditional manuscript material, Rushdie's collection included a significant born-digital component, specifically four computers, a hard drive and a handful of disks.[9] Although we had received pieces of digital media in previous collections, this was the first time that the Rose Library had acquired entire computers. This case study examines work performed by Manuscript Archivist, Laura Carroll, between 2006 and

2009 and reflects subsequent work performed by Digital Archivist, Dorothy Waugh, using ePADD v6.1 in 2018.

Local copies of emails were stored on two of Rushdie's four computers, plus the external hard drive. Archivists began processing one of those computers, Rushdie's Performa 5400, soon after the collection's arrival at the Rose Library. The donor agreement stipulated a fairly extensive list of restrictions intended to protect the privacy of Rushdie and his close family and friends. In addition to specifying a list of correspondents from whom email should be restricted outright, Rushdie also requested that all correspondence remain closed pending his review and approval.

In practice, this required that archivists opened and viewed each email, checking for content identified as restricted. Following this initial review, any email that either did not appear to fall under Rushdie's set of restrictions or for which the status was unclear were printed and shared with Rushdie in hard copy for his final review. Processing of the Performa 5400 was completed in 2009 and files are available to view at dedicated laptops in the Rose Library's reading room. Nearly 10 years on, the remaining components of Rushdie's born-digital collection are as yet still unprocessed and, crucially, inaccessible to researchers. Multiple factors have contributed to this delay, not least loss of technical support for the existing access interface. For the email specifically, however, the review process required in order to lift restrictions on access has also been a significant contributing factor.

Faced with these restrictions, ePADD's Discovery module might offer an interesting compromise. In lieu of a fully processed collection made available in our reading room, the Discovery module would enable online access to a redacted set of email, in which visitors can view only correspondent names, message dates and entities extracted from the email corpus. Examples of entities include persons, locations, publications and organisations. For the already processed email taken from Rushdie's Performa 5400, this online environment could provide a useful discovery tool that might inform later onsite research trips. For email as yet unprocessed, we are currently exploring the possibility of using the module to provide limited access to material that would otherwise be entirely unavailable due to restrictions lasting until Rushdie's death.

The Discovery module supports browse and search, allowing researchers to identify particular entities and view them within otherwise redacted email messages. Graphs provide visual representations of entity distribution across collections. ePADD's multi-entity search function analyses large blocks of text in order to find and highlight any matching entities within the email collection in question.

Similarly, the export of message headers and entities, a function of ePADD's Processing module, is also of interest. If such datasets could be made available, they would offer an opportunity for Rose Library staff to expose new information through text analysis and topic modelling, and provide examples to encourage new methods of research that better leverage the born-digital format. Given these possibilities, we are beginning to consider how revised language in our donor agreements could address if and how the library might use such extracted data to support discovery and research for collections subject to donor-imposed restrictions. An equally important part of this work, once we are at the point when we can share it with researchers, will involve gathering feedback as to the utility of this level of limited access.

Our interest at the Rose Library is in how these functions can provide discovery mechanisms for remote researchers and support limited analysis of an email corpus without violating donor-imposed restrictions and privacy concerns. So far, we have only explored this possibility in a test environment and will have to secure Rushdie's permission before proceeding further. By considering this approach, we are making the assumption that exposing correspondent names and other entities included within Rushdie's email would not violate the restrictions outlined in his donor agreement. Rather, this approach, which would redact the contextual information surrounding those correspondents and entities but not the correspondents and entities themselves, assumes that it is this contextual information to which Rushdie's restrictions apply. If further conversation with Rushdie proves these assumptions to be mistaken, we will need to rethink our approach. Even given the possibility that we do not receive permission from Rushdie, however, we hope this test case might set a precedent as to how we leverage the born-digital format in order to provide access to data, or entities, about email collections for which the content itself is either closed or requires an onsite visit due to as yet unexpired restrictions.

Case Study 4: Robert Creeley's email, Special Collections & University Archives, Stanford University

Robert Creeley (1926-2005) was a leading American poet of the twentieth century and core member of the 'Black Mountain School' of poetry. His email documents his work as a poet and professor of English at the University at Buffalo and Brown University, as well as his correspondence with other poets, friends and family. Creeley's papers were initially acquired by Stanford Libraries' Department of Special Collections and University Archives in 1993, and his email collection was deposited between 1999 and 2011 on floppy disks, optical discs and internal and external hard drives.[10] This case study reflects work performed by Sally DeBauche using ePADD v5.1-7.0 in 2018.

Stanford Libraries has implemented ePADD as its primary tool for acquiring, appraising, processing and providing access to email, and several email collections are currently available via Stanford's instance of the ePADD Delivery module.[11] This case study will describe some of the strategies employed to process this collection and functions within ePADD that supported them.

The main goal when processing a collection of email is usually to identify and restrict messages that contain sensitive content. This could include government-issued personal identification numbers, credit card numbers, personal health information and other protected information related to either the email account owner or their correspondents. Although the goal is straightforward, the task of finding that information within a large collection of email, and tracking those messages, is a challenge.

One of the most powerful tools for identifying sensitive content in the email collection is the lexicon, which contains predefined, but customisable, sets of terms organised under themes such as 'personal', 'medical', or 'academic'. ePADD lexicons are easy to edit; terms can be added or deleted as necessary, and we chose to add several terms to the lexicons that were specific to the Creeley email collection. Using lexicon analysis, we were able to identify and restrict messages that fell into these categories. One technique that proved especially effective was the 'test' function, which allows the user to limit their search to messages related to

specific terms within a chosen lexicon. Since the lexicons cast a wide net and include many terms that did not relate to the types of messages we were looking for, searching the collection in this more targeted way made our process more efficient.

Keeping track of restrictions placed on messages is an essential task in processing an email collection, to provide an accurate record of content that cannot be made available to researchers, and to document when, if ever, it can be released to the public. ePADD allows the user to attach labels to messages individually or as a bulk action, which can be used to note those restrictions. These labels can also be customised, so for this project, we created several new labels, including 'do not publish' and 'restrict for 80 years'. In the latter case, ePADD automatically determines the individual restriction period for each message based on the message date. We chose these restriction periods because we felt that they would protect the privacy of individuals discussed in the messages while still committing to providing access to the content when it no longer has potential to damage any individual's personal or professional lives.

While there are common types of information that we tend to restrict and tools geared towards identifying them, we also discovered sensitive information in some unexpected places.

Messages sent through email listservs can usually be safely ignored as they are typically aimed at wide audiences and do not contain personal information. This assumption held true until we stumbled upon a message sent through the Brown University English department listserv that included an attachment of the department's meeting minutes where specific job applicants were named and their candidacy discussed in detail. Since information related to job applications is typically considered private, we decided that these messages and attachments should follow Special Collections guidelines and be restricted for 80 years after the date that the message was sent, when, presumably, those candidates would be deceased. The lesson learned from this case was that not all listservs are created equal, and those targeted at smaller communities where members participate actively require greater scrutiny.

The Creeley email collection is available to researchers via the ePADD Delivery module in the Special Collections reading room, and the metadata is available online via the ePADD Discovery module. Researchers interested in browsing through the entire email collection need to visit the reading room in person. Users requesting copies of specific email messages are required to complete and sign a digital reproduction copyright declaration form, as they would for any archival material that is protected by copyright. We have already had one remote request for correspondence between Robert Creeley and a specific individual. After reviewing the request, we determined that we could release the selected messages to the user in accordance with our institutional policies. ePADD enabled us to export the set of messages as an MBOX file, and email that file to the researcher. As researcher awareness of using email as a primary source increases, we should anticipate greater use of our email collections both in person and remotely and ensure that our infrastructure and access models support that increased use.

Case study 5: Ian Wedde's email, Alexander Turnbull Library, National Library of New Zealand

Ian Wedde (b. 1946) is an Aotearoa New Zealand author, curator, editor, and critic who has published poetry and prose works since 1966. He was co-editor of the *Penguin Book*

of New Zealand Verse (1985), and editor of the *Penguin Book of Contemporary New Zealand Poetry* (1989). From the mid-1980s, he increasingly worked as an art critic and curator, taking leading roles in the Arts and Humanities at the Museum of New Zealand Te Papa Tongarewa between 1994 and 2004. He was awarded an Order New Zealand Merit in 2010, and the Prime Minister's Award for Literary Achievement in Poetry in 2014. Wedde was also New Zealand Poet Laureate from 2011 to 2013. Wedde's archive was acquired by the Alexander Turnbull Library, the Special Collections library within the National Library of New Zealand Te Puna Mātauranga o Aotearoa, in 2013. This case study reflects work performed by then Digital Archivist, Jessica Moran, in collaboration with Assistant Manuscript Curator, Seán McMahon, and Digital Materials Librarian, Dolores Hoy, using ePADD v2 in 2016.

Wedde's email archives were part of his larger hybrid papers that spanned 50 years and 11 meters, and includes correspondence, literary works, non-fiction projects, working papers relating to publishing, editing and translation work, and personal papers. The email archives the Library received were in fact only a small slice of his email. It appears that in 2005, while backing up a computer hard drive as part of moving to a new computer he saved copies of his email archives between 2003 and 2005 to optical disks. These were transferred to the Library as part of his papers in 2013.

The email archives were made up of five PST files from his Outlook email client and in total there were 2,152 individual messages from both the sent and received folders. Due to the discrete and manageable nature of these files, we determined they would make an excellent pilot test of the ePADD software.

In order to ensure that knowledge about ePADD and its potential integration into the Library's workflows was uniform among staff we decided to convene a small group to work together to process the email archives using ePADD. Staff involved included a digital archivist working with the technical processing of the digital component of the files, an arrangement and description librarian with expertise with digital materials, and the manuscript curator. After the digital archivist converted the PST files to the open MBOX format, and loaded the email into the ePADD Appraisal model, the group came together to view the email collection in ePADD.

Initial processing steps were focused on understanding how entity extraction, identification of potentially sensitive records, lexicon searchers, and advanced search worked. We spent time cleaning up the list of stop words, or words we wanted the NLP entity extraction to ignore, for example, 'papa' which in this case was not another name for father, but the name of New Zealand's national museum, Te Papa Tongarewa. We also spent time exploring the functionality of ePADD's regular expression (regex) searches, especially to find potentially sensitive personally identifiable information (PII). Using the regex expressions already available in ePADD, we identified a number of credit card numbers and accompanying information in emails that we then restricted. New Zealand's bank accounts use a different string pattern than US bank accounts so we wrote regex's for New Zealand bank accounts and New Zealand Inland Revenue Department or IRD numbers. These personal identification numbers are used for income tax, retirement savings, student loans, and buying or selling property in New Zealand, and can also be searched as a regular expression. After writing these we were able to find and restrict a small number of additional emails containing personally identifiable information.

We then developed and adapt specific lexicons to help us delve more deeply into specific topics and understand the extent to which Wedde's creative and professional projects were represented in his email. We created a lexicon for poetry and made small adjustments to a lexicon already available with ePADD for a faculty member. In the initial purchase agreement between Wedde and the Library for his papers, there were a number of sensitive topics that were to be either restricted for a period of time, or removed if found during processing. Using ePADD's lexicon and advanced search mechanisms, we were able to quickly search and find these topics and restrict or remove any message that fits our criteria.

Prior to introducing ePADD into our born-digital processing suite of tools, when we encountered mailbox files of email in personal or organisation papers the default was to restrict the entire mailbox. This practice was in place to ensure that we did not inadvertently release personal or sensitive information, since we did not have the tools or staff resources to review each individual email and attachment in a mailbox folder. With the introduction of ePADD we have greater confidence in our ability to provide access to mailboxes because we now have a tool to identify potentially sensitive content, quickly visualise the extent of topics and people, and make more informed decisions about what can and cannot be made available to researchers. While we understand that both ePADD and its human operators are not infallible, using ePADD we have a higher degree of confidence in our decisions to make some content available.[12]

Using ePADD has also allowed the Library to imagine how we might process other text-based digital collections in the future using similar tools. Often people point to the value of physical library stacks for serendipitous discovery, and lament that this is lost when working with digital archives, but playing with ePADD allowed us to have that experience. Reviewing the correspondence list within ePADD we discovered a listing for Donna Haraway. Further investigation confirmed that the Donna Haraway whom Wedde was corresponding with was the theorist of science and feminism, known for her work such as *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century* (Berkeley Socialist Review Collective, 1985); the essay was also included in the volume, *Simians, Cyborgs and Women* (Free Association Books, 1991). This led to the discovery of an exchange of letters between Wedde and Haraway about her work, *The Companion Species Manifesto* (Prickly Paradigm Press, 2003), which looks at the relationship between humans and their pets, and Wedde's essay 'Walking the Dog' first published in his volume *Making Ends Meet* (Victoria University Press, 2005), which is included as a born-digital manuscript in Wedde's papers at Turnbull Library. By browsing through the correspondence list using ePADD, I came across this unexpected exchange of letters, which then led back to manuscript versions of both writers works within the archives: unexpected connections made possible through ePADD.

That really is the magic of working in the archives and illustrates that serendipitous discoveries and connections are possible in digital libraries and archives. Still in its development, ePADD has demonstrated its usefulness for processing and managing email archives, and potentially for how we might access many other text-based digital collections. It is exciting to think about what the future of research with these born-digital archives may look like.

Discussion

The five use cases shared in this article highlight how institutions with differing missions, policies, workflows and resources are using ePADD to appraise, process and/or provide access to email collections in their care. The cases also surface how particular collection rights and conditions, and donor agreements and concerns, can also impact how tools like ePADD are applied.

ePADD clearly brings new opportunities to help confront issues that have proven to be major challenges to institutions in the past. In the case of the Wendy Cope email at the British Library, the authors write that, 'if ePADD had been available in 2011, most of these problems could have been mitigated [through entity browsing, correspondent list search, lexicon searches, bulk labelling and image browsing], and the continuity of ePADD's user experience from ingest to access would have allowed Cope to visualise her email as an archival collection rather than a live work environment.' A related point is raised in the case of the Ian McEwan email at the Ransom Center, University of Texas at Austin: in this case, since the email was accessioned prior to the adoption of ePADD, the donor's assistant likely removed messages from the archive using a broader brush than would currently be needed (or required according to the terms of the purchase); in place of applying bulk restrictions across messages, many messages were likely simply removed.

Similarly, the authors of the case study on Ian Wedde's email at the National Library of New Zealand reflect that 'while we understand that both ePADD and its human operators are not infallible, using ePADD we have a higher degree of confidence in our decisions to make some content available … we have greater confidence in our ability to provide access to mailboxes because we now have a tool to identify potentially sensitive content, quickly visualise the extent of topics and people, and make more informed decisions about what can and cannot be made available to researchers.'

The Cope case study is unique in that it highlights the dramatic impact of the European Union's General Data Protection Regulation legislation (GDPR) on institutional workflows for affected institutions, yet almost all of the case studies shed light on the need to protect donor and third-person privacy, including any restrictions identified or enacted through a legally binding donor agreement. This can be seen even more clearly in the case of Rose Library at Emory University. In this case, a donor agreement and restriction guidelines are interpreted with the help of consultation with a living creator/donor, in order to ensure the guidelines not only meet legal guidelines and current agreements, but are also interpreted and applied to the donor's satisfaction in light of new technologies for archival processing and research. Especially in cases where the record creator is no longer living, software like ePADD may raise difficult yet important questions regarding the intention of the records creator and the legal and ethical responsibilities of the collecting repository.[13]

Email is not unique among born-digital materials in raising these ethical concerns and questions, and ePADD is certainly not alone or foremost among software projects that attempt to give record creators and archivists the ability to begin confronting them.[14] Still, with that in mind, the highly personal nature of correspondence can make the sharing of email a particularly worrisome consideration for donors, and likewise place stress on cultural memory workers responsible for ensuring the ethical provision of access.

Almost all of the case studies highlight the utility of customisable lexicon searches to improve searching for confidential, protected, or legally protected information (including

information protected by GDPR), proving that even in the context of many additional methods intended to surface this information, including entity analysis, for now search remains considered the primary and most time-efficient method of screening.[15] Likewise, ePADD's customisable bulk labelling functions are identified as further enabling processing as scale. As the Cope study points out, when 'used in conjunction with lexicon searches, which can be customised to leverage the archivist's deep knowledge of the archive to greatest effect, it is possible to make the process of Data Protection compliance checking much more efficient'.

In addition to the effect of GDPR on institutional workflows and practice, the geographical and linguistic context of email collections was another area of focus for several collections, which is not surprising as the correspondence of contemporary literary figures is often international in scope. The McEwan, Cope and Wedde case studies all mention expanding the scope of regular expression searching to accommodate additional government identification numbers. The Wedde study also discussed how staff customised a default lexicon and edited extracted entities to ensure more accurate results for Māori terms.

Several authors highlighted areas where the software either fell short or did not entirely reflect the complexity inherent to screening email collections for confidential information, necessitating more extensive manual review. The McEwan case study especially highlighted, 'that the typical human reticence when discussing sensitive topics, the tendency to use euphemisms, and the idiosyncrasy of individual speech all seem to be a major roadblock to any attempt at automating this process via keyword search'. Although all case studies, to some extent, highlight the value of preliminary research to identify areas of potential sensitivity and to inform approaches to screening for sensitive information, the McEwan case study especially recommends that processing archivists would benefit from spending a significant amount of time, 'researching the creator and the collection, or on outlining a dossier of personal data (including sensitive number strings, nicknames, and so on) for the creator to generate and include with the accession'. One area for future research might be to assess the extent to which a machine learning model, trained on other messages flagged with particular restrictions, might help to surface those sensitive messages which might not otherwise be easy to identify via existing tooling.

Besides focusing on how the Appraisal and Processing modules might assist donors and institutions with collecting, screening and preparing email for access by researchers, several case studies mention adoption of ePADD to support researcher access. Stanford University's Creeley case study shares how Stanford is providing access to that email both via ePADD's online Discovery module, as well as within Stanford's Special Collections reading room via the Delivery Module. The McEwan and Cope case studies both mention upcoming adoption of the Delivery module to support researcher access in the repository reading room.

The Creeley case study maintains that, 'as researcher awareness of using email as a primary source increases, we should anticipate greater use of our email collections both in person and remotely and ensure that our infrastructure and access models support that increased use'. The Cope case further reflects on the value of a uniform experience for researchers which could be provided through mass institutional adoption of ePADD within repository reading rooms, so that 'the Library's increasingly international community of researchers, who may have accessed email collections through ePADD in other repositories, will not be required to learn a new system'. In a similar vein, the ePADD development team is also seeking to expand

the list of institutions that provide access to email via the Discovery module, so as to provide a more uniform discovery experience as well as perhaps a unified, federated, discovery portal. It would be of value to better understand the extent to which researchers' cross-institutional research experience might be impacted (and improved) by consistency in the application and availability of software in these ways.

Conclusion

The Cope study ends by reflecting on how '[ePADD] has allowed us to think critically about how email was explained to and acquired from donors in the past, and to envision a future where an international workflow exists to provide access to email collections as a matter of course'. Along these same lines, the Wedde study highlights that, 'using ePADD has also allowed the Library imagine how we might process other text-based digital collections in the future using similar tools'.

Perhaps the most fitting ending to this compilation of case studies is to draw attention back to these observations and aspirations. The libraries, archives and museums that collect, process and make available email and other born-digital materials benefit from the development of community-driven software such as ePADD, in part because this tool helps to foreground donor concerns, institutional obligations and research needs, in the context of evolving technologies, and in a clearly defined and productive manner. In other words, the application of artificial intelligence as a force multiplier to enhance the ability of donors, archivists, and others to meet these challenges requires a regular re-evaluation of how those concerns and needs are communicated, understood, and applied. Likewise, ensuring the conversation continues to take place via international forums will help make certain the software is flexible enough to meet the needs of an increasingly global cultural heritage and research community.

Notes

1. See: 'The Future of Archives: A Report from the Task Force on Technical Approaches for Email Archives', available at <*https://clir.wordpress.clir.org/wp-content/uploads/sites/6/2018/08/CLIR-pub175.pdf*>, August 2018, accessed 12 November 2018.
2. The ePADD website is available at <*http://library.stanford.edu/projects/epadd*>, accessed 12 November 2018; ePADD code and software releases are available at <*https://github.com/ePADD/epadd/*>, accessed 12 November 2018.
3. Institutional partners for the grant period covering 2015-2018, funded by the Institute of Museum and Library Services, include Harvard University; the Metropolitan New York Library Council, University of California, Irvine; and University of Illinois, Urbana-Champaign. Partners for the previous grant period covering 2012-2015 and funded by the National Historical Publications and Records Commission (US NHPRC) include Columbia University, New York University, Oxford University and the Smithsonian Institutions. Additional persons and institutions, too numerous to list here, have also contributed time, effort and resources to the development of the software and community.
4. ePADD accepts mail transferred via IMAP as well as mail in MBOX format. The British Library converted Cope's PST file to MBOX using Aid4Mail before ingest in ePADD. Aid4Mail is available at <*https://www.aid4mail.com/*>, accessed 29 October 2018.
5. As outlined in Jonathan Pledge and Eleanor Dickens, 'Process and Progress: Working with Born-digital Material in the Wendy Cope Archive at the British Library', *Archives and Manuscripts*, vol. 46, no. 1, 2018, pp. 59-69.

6. For a more detailed description of GDPR in a UK archival context, see: *Guide to Archiving Personal Data*, available at <*http://www.nationalarchives.gov.uk/documents/information-management/guide-to-archiving-personal-data.pdf*>, accessed 29 October 2018.
7. J Schneider's note: Improving accuracy for name resolution was a focus of a subsequent software release (v7), though the author is correct to note that results are not always accurate – for this reason, ePADD offers the ability for users to manually override the algorithm for names that have been incorrectly resolved.
8. Access was also granted to Lise Jaillant in Summer 2017. See L Jaillant, 'From Letters to Emails: Reading Ian McEwan's Correspondence', *TLS Online*, 2017, available at <*https://www.the-tls.co.uk/articles/public/ian-mcewans-emails-letters/*>, accessed 3 January 2019.
9. See the Finding Aid to the Salman Rushdie papers, available at <*https://findingaids.library.emory.edu/documents/rushdie1000/*>, accessed 10 November 2018.
10. The original count of over 150,000 email messages included many duplicate messages; following deduplication by ePADD and processing to screen for sensitive materials, the total number of unique messages stands at approximately 50,000.
11. ePADD Discovery site, available at <*epadd.stanford.edu*>, accessed 10 November 2018.
12. Turnbull Library supports three levels of access to born-digital materials: open and available online, open and available only in the reading room, and available only through permission. Access to the Ian Wedde email is currently available only by permission, to allow staff time to prepare the collection for researchers. See <*https://tiaki.natlib.govt.nz/#details=ecatalogue.802282*>, accessed 11 November 2018.
13. Recent literature on digital forensics, digital curation, participatory archives and post-custodial archives have foregrounded ethical considerations and the need for more open dialogue with record creators in the context of evolving models of collaboration.
14. Several projects in support of acquiring, preserving and facilitating access to sensitive, community-driven born-digital content have placed ethical considerations front-and-centre. For instance, DocNow is a 'tool and a community developed around supporting the ethical collection, use, and preservation of social media content', <*https://www.docnow.io/*>, accessed 11 November 2018; and Mukurtu CMS is 'a grassroots project aiming to empower indigenous communities to manage, share, narrate, and exchange their digital heritage in culturally relevant and ethically-minded ways', <*http://mukurtu.org/about/*>, accessed 11 November 2018.
15. Previous work by the ePADD Lexicon Working Group has ensured that ePADD provides new users with several examples of how new lexicons might be created or existing lexicons further refined to better screen email or support researcher access for a particular collection. See ePADD Lexicon Working Group page, available at <*https://library.stanford.edu/projects/epadd/community/lexicon-working-group*>, accessed 11 November 2018.

Acknowledgements

The authors would like to thank Lise Jaillant for encouraging us to develop this compilation of case studies into an article, and for the wonderful series of workshops she held in 2017-2018, with the support of the British Academy, to explore how contemporary researchers are interacting with born-digital collection material. We would also like to thank Peter Chan, Digital Archivist, Special Collections & University Archives, Stanford University Libraries, and ePADD Project Manager; Glynn Edwards, Assistant Director, Department of Special Collections, Stanford University Libraries, and ePADD Project Director; Sudheendra Hangal, Professor of Computer Science-Practice, Ashoka University, and ePADD Technical Advisor; and Chinmay Narayan, Developer, Amuse Labs, and ePADD Developer.

Development of the ePADD software has been supported by the US Institute of Museum and Library Services (2015-2018), US National Historical Publications and Records Commission (2013-2015), and Stanford University Libraries.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Institute of Museum and Library Services [LG-70-15-0242-15]; NARA; Stanford University Libraries.

Notes on contributors

J. Schneider is Assistant University Archivist, Department of Special Collections & University Archives, and Product & Service Manager, Digital Library Systems & Services, Stanford University Libraries. He is also Community Manager for ePADD, recipient of the 2018 Digital Preservation Award for Research and Innovation and the 2017 National Digital Stewardship Alliance Innovation Award. Josh serves on the editorial board of the Society of American Archivists (SAA) peer-reviewed academic journal, *The American Archivist*. He holds an MLIS from Simmons College and a BA in Philosophy from Brown University.

C. Adams (they/them) is Digital Archivist at the Harry Ransom Center at University of Texas at Austin and acts in a volunteer role as Digital Archivist for the Electronic Literature Organization. Adams previously worked at Hagley Museum and Library as Digital Projects Coordinator and Access and Electronic Records Archivist in Special Collections at University of Georgia.

S. DeBauche is Digital Archivist in the Department of Special Collections & University Archives, Stanford University Libraries. She is responsible for processing and describing born-digital archival materials and developing digital curation workflows. Sally previously worked at the Hoover Institution Library & Archives as a Digital Archivist.

R. Echols is a PhD candidate in English at the University of Texas at Austin and Research Associate at the Harry Ransom Center, completing a dissertation on ecology and rural nostalgia in twentieth-century British literature. He is the Processing Assistant for the Ian McEwan Email archive and former Graduate Intern at the Ransom Center.

C. McKean is Curator of Contemporary Literary and Creative Archives at the British Library. He has a background in contemporary literature and philosophy, having completed his MPhil at the University of Cambridge in 2017, where he wrote his thesis on auto-fiction and phenomenology. Before his curatorial role, he worked as a Reference Librarian specialising in the Social Sciences and News collections at the British Library.

J. Moran is Digital Collection Services Leader at the Alexander Turnbull Library, National Library of New Zealand. She leads the teams responsible for born-digital and digitised collections and services. Prior to moving to New Zealand she has worked in university and government libraries and archives in California, including most recently the California State Archives.

D. Waugh is Digital Archivist at the Stuart A. Rose Manuscript, Archives, and Rare Book Library at Emory University where she is responsible for the acquisition and management of the Rose Library's born-digital collections. She holds an MLS from Indiana University, Bloomington and an MA in English Literature from the Ohio State University.

ORCID

J. Schneider http://orcid.org/0000-0002-7166-5080
C. Adams http://orcid.org/0000-0001-6095-9705
S. DeBauche http://orcid.org/0000-0002-9239-1576

R. Echols http://orcid.org/0000-0002-0172-7020
C. McKean http://orcid.org/0000-0002-1708-2677
J. Moran http://orcid.org/0000-0003-1206-6755
D. Waugh http://orcid.org/0000-0003-0188-0401