MANCHESTER 1824

The University of Manchester

Carcanet Press Email Preservation Project Phases 2-3

Final Report

Fran Baker
April 2014

Table of Contents

1. Project Summary

Phase 1 of the Carcanet Press Email Preservation Project was completed in May 2012. This was a seven-week JISC-funded project, and the final report can be found at: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:165096.

This initial project was invaluable in giving a kick-start to our practical digital preservation efforts, and focused on one of the most challenging formats to preserve. Much was achieved during that seven-week period, but inevitably there was much which could not be covered in such a short space of time. Phases 2-3 of the project aimed to address outstanding issues and take work forward.

Phase 2 of the project ran from July 2012 to January 2013, and focused on archiving email at 'sequence', or PST, level, including:

- Writing new code and developing formal workflows for processing email archives at sequence level.
- Developing content models and metadata profiles.
- Fully documenting the processes involved in sequence-level acquisition and preservation.
- Successfully ingesting both of the 2012 accessions at sequence level.

The drawback in preserving at sequence or PST level is that each PST is a single file which contains thousands of individual emails and attachments. Viewing its content requires it to be downloaded into a version of the Outlook client, which is time-consuming and less than ideal for preservation and management. Phase 3 of the project addressed this issue. This ran from February 2013-February 2014, although other strategic commitments meant that much of the work was carried out from August 2013 onwards. It included:

- Breaking down PST files to individual email level.
- Migrating emails to different preservation and access formats whilst retaining their 'significant properties'.
- Developing new code, metadata profiles, and procedures for preserving at email level, and fully documenting these.
- Acquiring a new accession to the Carcanet Email Archive (containing a further 53,262 emails).
- Successfully ingesting the entire collection into Manchester eScholar, our institutional repository, resulting in 282,375 digital objects (effectively doubling the size of the repository).
- Risk analysis of 65,500 email attachments spanning multiple formats.
- Analysis of researcher requirements for accessing and using email archives.
- Design of a curatorial tool for managing the archive.
- Some visualisation experiments based on the archive.

2. Background

Whilst the JISC-funded element of the project provided a valuable kick-start, it was impossible to complete all the necessary work to the level we had hoped in such a short timescale. At the end of this initial phase, particular issues which remained to be addressed included the following:

- Coding languages: the temporary developer who worked on the project used PERL to develop scripts, but PERL is not widely used in the Library, so scripts needed to be re-written in XML.
- Work was hampered by having too little storage space for experimenting with format migration, too little processing power for dealing with large collections of emails, and by the restrictions imposed by the University's managed desktop.
- The limited timescale meant that we did not have time to complete key tasks like verification of format migrations and ingest, or to fully test some of the tools identified as potentially useful.

Phases 2-3 aimed to address these issues and to build on the work carried out during the funded project.

3. Objectives

The objectives of Phases 2-3 of the project were as follows:

- Finalise work achieved in the first, funded, phase of the project focusing on sequence-level preservation.
- Augment, finalise and document workflows, tools and procedures for dealing with any future accessions of email from Carcanet Press – at both sequence and individual email level.
- Ensure workflows can be adapted to deal with email in other formats.
- Transfer knowledge of digital preservation to other relevant staff members.
- Ingest Carcanet Press emails at both sequence and email level.
- Augment the existing Carcanet Press digital archive by taking in a new accession of material.
- Test and assess new digital preservation tools.
- Ensure our digital preservation 'lab' is fit for purpose.
- Develop tools which facilitate the management of email archives by curators.
- Through dealing with email attachments, develop a body of knowledge and procedures which will aid work on preserving other types of digital material.
- Gain an insight into the needs and requirements of researchers who will be the ultimate users of archival emails.
- Update our knowledge and understanding of the current landscape in the field of email preservation.
- Establish a solid foundation for the Library's planned strategy project focusing on digital preservation.

4. Outputs

Primary outputs of the project are as follows:

- A series of workflows for processing and managing email in PST format (see Appendix 1).
- 120-page manual fully documenting every stage of acquiring, processing, preserving and managing email in PST format.
- Full content models, metadata profiles and templates.

- Library of code for every stage of work, from acquisition to ingest.
- Some initial visualisation experiments.
- Digital preservation 'lab', with appropriate hardware and specialist software.
- Initial functional design for a curatorial tool which will enable the archivist to manage the archive.
- Sophisticated index which forms the foundation for many different ways of exploiting the material.
- A large body of digital archive material preserved to the highest digital preservation standards, consisting of 282,375 digital objects in total.

5. Outcomes

Key outcomes of the project are as follows:

- Heightened profile for the UML in the digital preservation, digital humanities and library/archive communities.
- Insight into researcher requirements and potential research uses of email archives through detailed interviews with academics working in several different disciplines.
- In-house expertise in digital preservation processes, tools and metadata schemas.
- Effective practical application of digital preservation principles at a large scale.
- Awareness of, and initial links made with, related initiatives in other institutions via detailed literature review and correspondence.

6. Project Methodology

Phases 2-3 of the project were both completed by existing staff without the assistance of external funding, and demonstrate how effective such an approach can be if a project is prioritised and formally scheduled into participants' workloads. The project also provided a useful model of cross-departmental working, with staff drawn from three of the University Library's teams: Digital Technologies and Services, Special Collections, and Collection Management.

Phase 2 of the project ran from July 2012 to January 2013 and involved Fran Baker (Archivist) and Phil Butler (eScholarship Manager) working over approximately 16 days in order to complete the 'sequence' or PST-level preservation challenges and documentation which were initiated as part of the JISC-funded project.

Phase 3 of the project ran from February 2013-February 2014. For the first six months, this was small-scale and informal in nature due to other strategic commitments. From August 2013-February 2014, the project was administered more formally, with a project charter and a steering group which met every three weeks, as well as benefiting from input by additional staff.

The Project Steering Group consisted of:

- Sandra Bracegirdle (Head of Collection Management)
- John Hodgson (Manuscripts and Archives Manager)
- Fran Baker (Archivist)
- Phil Butler (eScholarship Manager)

The operational element of Phase 3 primarily involved Fran Baker and Phil Butler working over approximately 52 days in total from February 2013-February 2014. Additional input came from:

- Caroline Martin (Digital Preservation Co-ordinator): work on risk analysis of attachment formats, and researcher interviews.
- Ben Green (Digitisation Infrastructure Manager): work on data transfer and security.

As with Phase 1, the approach was incremental and iterative, including frequent meetings between key staff, ongoing monitoring of progress and modification of goals as necessary.

7. Project Overview

This section provides a detailed account of the project's work, which it is hoped will be useful to other institutions undertaking email preservation.

7.1 Collection development

Carcanet Press is one of the UK's premier poetry publishing houses, and its archive is one of the most important held by the Library. Alongside authors' manuscripts and proofs, the core of the archive consists of correspondence with well-known poets, writers, translators, editors, critics, publishers, artists and many more from across the world. Increasingly this correspondence is conducted digitally, and few emails are now printed out for the correspondence files.

The JISC-funded phase of the project enabled the Library both to take its first steps in practical digital preservation, and to rescue a vast body of key research material in the form of email which had been residing on hard disks and local networks at the Carcanet office for 11 years. It also enabled us to put together the initial version of a workflow which would ensure that annual accessions of email could be acquired straightforwardly alongside the hard copy archive, ensuring the Library's ability to continue collecting in the digital age.

A key aim for the latter stage of the project was to take in another accession of email from Carcanet in 2013, documenting the Press's business since the initial accessions were made in early 2012. This would also test that our workflows, which had been refined since 2012, were fit for purpose. As we worked, we documented the processes in more detail for our in-house manual, to ensure greater self-sufficiency for curatorial staff acquiring such accessions in future.

The focus of this new accession was:

- Email of the Managing and Editorial Director of the Press, created between February 2012 (when the first accession was taken in) and October 2013.
- Email of the former Managing Editor from February-August 2012, when she left the Press.
- Email of the current Managing Editor from September 2012, when she started work at the Press, to October 2013.

This material totalled approximately 53,300 individual emails and 20,100 attachments.

A future phase of work will focus on the email correspondence of the Sales and Marketing team at the Press. Traditionally, sales and marketing material has not been acquired as a matter of course, although ad hoc acquisitions have been made. As sales and marketing reflects the final phase of an

author's interaction with the Press in relation to any single publication, this function will be taken into account in future. Initially, a version of the record-keeping questionnaire developed at the beginning of the JISC-funded project will be used to ascertain record-keeping practice of the relevant staff, and a more detailed records survey initiated before material is acquired.

Before taking in the new Managing Editor's email, the archivist conducted an interview about how she manages her email account in order to ascertain how much appraisal was likely to be required, and adding to our body of overall information about record-keeping practice; her practice fell somewhere between the two extremes of behaviour represented by her predecessor and the Managing and Editorial Director (outlined in the JISC project report).

Other broad behavioural trends were observed while processing the new accession of emails, including an increased proportion of attachments being exchanged during the more recent period – reflecting a shift towards more work (e.g. the exchanging of manuscripts and proofs) being carried out digitally, and perhaps a result of increased storage capacity and processing power in the current technical environment.

7.2 Data transfer, security and hardware

We chose to continue using the encrypted removable hard drive we had used in the first phase of the project for the transfer of new PST files from the Carcanet offices to the Library. We continued to use TrueCrypt software[1] to encrypt the hard drive, and Jacksum[2] to run fixity checks on PST files before and after transfer.

We instituted new folder structures and filenaming conventions which ensured consistency throughout the processing of each PST. Essentially a top-layer folder contains everything relating to a single accession; subfolders each contain a single PST file, along with all the files generated during the processing of that PST (extracted metadata, fixity readings and so on), with appropriate naming conventions.

The files were transferred to the newly equipped Digital Preservation Lab at the Main University Library on campus, which was established after the funded phase of the project had completed.

Some of the problems identified during that first phase related to hardware and network storage space, including:

- Issues with downloading and running certain types of software because of the restrictions imposed by the University's 'managed desktop' system.
- Lack of storage space: although 100GB of secure network storage was allocated for the project, ultimately this did not provide enough space for carrying out processing and migration on a large scale and work had to be divided across two network drives and the hard drive of the Workbench PC.
- Lack of processing power: extracting metadata and breaking down PST files, especially if running more than one of these processes at once, required more powerful computers.

As a result, during Phase 2 of the project, new computers were acquired which were not linked to the University's managed desktop system. Both the Quarantine PC (where initial virus checks and

[1] http://www.truecrypt.org/

[2] http://www.jonelo.de/java/jacksum/

appraisal are carried out) and the Workbench PC (for subsequent processing) have 16 GB (4x4GB) 1600 MHz DDR3 Non-ECC memory; 1TB 3.5 inch Serial ATA III (7.200 Rpm) hard drive; and 3rd Gen Intel Core i7-3770 (Quad Core, 3.40GHz Turbo, 8MB) processor. Their hard drives have been partitioned, so that multiple projects or tasks can be carried out on different partitions. At the conclusion of any project, secure deletion software is run on selected partition/s without the need to wipe the entire hard drive. Processing now takes place on the hard drive of the Workbench PC, with content only being moved to secure network storage on ingest.

The two PCs are equipped with all the necessary digital preservation software, and have been imaged to allow complete reconstruction after a secure wipe if necessary.

In addition we have acquired a third PC which has been fitted out as a Linux workstation for the installation and operation of Linux-based digital preservation software.

After transfer to the Quarantine PC, new PST files are fixity checked to ensure no changes have occurred during transfer. They are then virus checked at individual message and attachment level using McAfee embedded within an Outlook account. This checking was more granular than that carried out on the 2012 accessions – something which was highlighted at a later stage when Aid4Mail software failed to process one attachment from a 2012 accession due to a Trojan virus.

We only encountered two viruses in the 2013 accession, both of which were Trojans embedded in spam email attachments; these were consequently stripped from the archive.

7.3 Digital forensics and archival appraisal

Assessing the content of PST files for the purpose of carrying out archival appraisal proved to be a major challenge during Phase 1 of the project, requiring a dummy Outlook account to interrogate material, and making multiple screenshots to try and capture a record of folders targeted for deletion which could be presented to the donor/depositor if necessary. We recognized that forensic software, which is increasingly being used by digital preservation practitioners, could both help to overcome some of these problems, and facilitate some other preservation activities we are keen to pursue.

During Phase 2 of the project, the team visited colleagues at the Bodleian and British Libraries to discuss and look at some forensic software in action. As a result, we purchased Forensic Toolkit (FTK) and Paraben's Email Examiner software. We identified several advantages offered by this type of software, including:

- The ability to deal with multiple different email formats without requiring the relevant email clients.
- It enables you to connect to, and view, a file without in any way compromising the fixity value or changing its properties; by contrast, simply opening a PST file in an Outlook account can change its checksum reading.
- It enables you to preview attachments in multiple formats within the program, so the relevant software (e.g. MS Word) is not required.
- It has quite sophisticated searching functionality.
- It can extract, and export, useful metadata.
- It can run fixity checks.

- It allows you to 'bookmark' or highlight particular folders and messages for any user-defined reason (e.g. archival arrangement; highlighting sensitive data; creating access sets).
- It maintains an audit trail which is important for ensuring authenticity of data.
- Some packages can carry out visualisation of data.

However, forensic software does not provide all the answers: it does not necessarily produce metadata in exactly the format we require it to be in, so further transformations would be necessary; as a tool utilised by law enforcement agencies, by definition it does not facilitate secure deletion of material from within a PST file (this still had to be done using the compaction process within Outlook, as detailed in the initial project report); and it does not provide a solution for long-term archival storage.

By the time of purchasing this software, considerable progress had already been made on our processing workflows using existing tools like Aid4Mail,[3] PST Reporter[4] and Jacksum, which were doing things that forensic software could potentially do. Aid4Mail, for instance, can deal with email on a server (it does not require a PST file); it can deal with many different formats, both web- and server-based; and it can process individual message files.

However, Paraben's Email Examiner was identified as something which could greatly facilitate the appraisal process, and it was employed for this purpose during Phase 3 of the project when the new accession of email (comprised of three PST files) was acquired. The software was used to view/analyse content, to create 'bookmarks' against messages and folders highlighted for secure deletion, and to generate HTML reports of these which can be presented to the donor/depositor, and retained as a record of appraisal actions.

We also noted that the software could be used post-compaction to extract key metadata and thus form an additional layer of verification against other metadata extraction tools.

We therefore adopted forensic software for this project to complement, rather than replace, other key software we were already using.

As with Phase 1 of the project, appraisal was carried out largely at folder level – although some more detailed appraisal focused on messages stored directly within the Inbox of each PST. Further, more granular, appraisal is likely to be carried out post-ingest.

The appraisal reports generated by the Email Examiner software are being stored digitally in EMu collection management software,[5] which was acquired by the Library after the first phase of the project was completed. EMu provides a useful secure storage mechanism for files like appraisal and virus reports which are not being ingested with the digital archive, but do need to be retained for reference. They are attached as multimedia files to the relevant accession record, and access to them is restricted to two staff members.

[3] http://www.aid4mail.com/

[4] http://www.nucleustechnologies.com/outlook-pst-reporter.html

[5] http://emu.kesoftware.com/about-emu/overview

7.4 Migration and preservation planning

7.4.1 Significant Properties

As the email acquired was in the form of proprietary Microsoft PST files, we aimed to migrate these files to preservation-friendly platform-neutral formats if possible. We envisaged preservation taking place at two levels:

- The initial focus was on 'Sequence' (or PST) level, i.e. migrating a single PST file to a more neutral single-file format which captured all the essential properties of the PST at the same level.

- We also hoped to break down each sequence (or PST) to individual email message level, and migrate to appropriate single-message format/s.

Following best practice, we therefore undertook an analysis exercise to identify the 'significant properties' or salient characteristics of the email which we felt should be preserved through any subsequent format migrations. At PST file/sequence level, these were high-level, easily quantifiable properties which we knew could be measured using the tools we had, and consisted of:

- Number of folders
- Folder path names
- Number of messages
- Size of folders
- Number of attachments
- Mimetype of attachments

At individual message level, properties were much more granular and took into account the context and content of this particular email archive, generated by a poetry publisher. Some of the properties we noted included:

- Different ways of indicating book titles, e.g. by use of italics, bold, or capitalisation.
- Centring, line breaks and indentation, where (for instance) the text of poems is included in emails.
- Idiosyncratic spacing, paragraph breaks and layout, which reflect a writer's style. Some correspondents laid out their emails very formally in the same way they would treat hard copy correspondence; others treated the medium very informally, with minimal punctuation, paragraphing or capitalisation.
- Deliberate use of unusual fonts, which in some cases were referred to in the text of the message – the references becoming meaningless if the font is lost.
- Colour or indentation indicating extracts of text from an original message to which the recipient is responding.
- Use of font colour in extracts from proofs which had been pasted into an email, with red text indicating printer's errors, and blue indicating authorial emendations.
- Foreign characters and scripts, reflecting the international nature of Carcanet's network of correspondents.

At individual email level, then, our set of significant properties consisted of all message header information (as documented by the InSPECT Project[6]), and the following properties relating to the email body:

- All textual content
- Body background
- Line and paragraph breaks
- Horizontal rules
- Tabs
- Text alignment
- Formatted text
- Headings
- Emphasis
- Bold
- Italics
- Underline
- Strikethrough
- Font type
- Font size
- Font colour
- Subscript
- Superscript
- Lists
- Tables
- Character set
- Authoring device where stated
- URLs for links retained, but hyperlinks not traversible

Significant properties were recorded at the relevant level using the PREMIS metadata schema[7] (see Section 7.5.3).

7.4.2 Format migration

7.4.2.1 Migration

In order to carry out migration experiments and verify their results, we created a structured PST file containing a set of approximately 150 emails representing all our identified significant properties, copied from across both of our initial accessions.

For sequence level, we had already experimented during Phase 1 of the project with migration to MBOX and some of its variants, and felt that none of them offered an exact equivalent to the PST – either failing to retain key properties like folder structure post-conversion, or requiring further compression to preserve them as a single file. In light of this, we opted simply to preserve the PST files in their original format at sequence level. As PST is a current, well-supported and widely used

[6] Gareth Knight, *Significant Properties Testing Report: Electronic Mail* (30 March 2009). http://www.significantproperties.org.uk/email-testingreport.html

[7] Preservation Metadata: Implementation Strategies, http://www.loc.gov/standards/premis/.

format, the files are unlikely to become obsolete in the near future. We will keep a watch on technology and look to migrate them at a time when obsolescence looks likely, but at this stage we decided to focus our main migration efforts at individual email level.

Our original aim in Phase 1 of the project was to split the PST files down to email level, and to preserve each email in three formats:

- EML, which is more platform-neutral than Microsoft formats.
- XML, which is the preferred format for preservation, and which can also preserve formatting markup instructions.
- MHT: we identified this as a possible access format for the future. It generates a browser-based view of emails without the need of an email client. However, it is not standard HTML, and is tied to Microsoft for full rendering; this means that in Internet Explorer, MHT messages display to the user in a way which is visually similar to the way they appear in an email client, but in other browsers they do not. In future, it might be possible to write a program to render the MHT versions more consistently across all browsers, so we felt that this format was worth retaining.

During Phase 2 of the project, we also decided to retain each email in MSG format, as the format which most closely replicates the 'original' representation of each message (MSG being the MS Outlook format for individual messages).

The most reliable migration tool we experimented with during Phase 1 was Aid4Mail, and we found the Aid4Mail help desk extremely responsive. However, we also investigated Xena[8] and PeDALS Email Extractor[9] as tools which offer conversion to XML for preservation.

We encountered problems with running PeDALS at scale, and at the time of Phase 1, it did not appear to have been widely adopted, although now the Sourceforge page reports over 4,000 downloads. Xena was specifically written to convert emails to XML for preservation; it seems promising, although the software is not well documented.

Whilst XML was our preferred preservation format, as outlined above, we also wished to migrate to several other formats. Aid4Mail can break down PST files to message level and migrate to XML very effectively: the XML version of each email includes the body of the message with full formatting instructions (showing font details, linebreaks, colour and so on), as well as a plain text version of the message body. The software can also migrate to the three other formats we had identified. This, and the helpful support we received from them during the first phase of the project, led us to continue using Aid4Mail for the latter phases of our work.

7.4.2.2 Verification

Aid4Mail migrates to multiple formats and in each case it offers various different output options. Our initial tests simply involved testing all these options and rejecting those which (a) obviously failed to retain the properties we had identified; and (b) did not output results in a way which met our requirements for subsequent processing and ingest.

[8] http://xena.sourceforge.net/

[9] http://sourceforge.net/projects/pedalsemailextr/

As a result, we identified our preferred options for converting to all four formats – MSG, EML, MHT and XML.

Given the time constraints of the project, we did not have time to develop a mechanism for automating any of the verification procedures, so this had to be done by manually comparing source and migrated versions. On comparison, we were confident that the message header properties had been retained in all formats.

For the message body and display-related properties, we felt that key properties had been retained, although there were some issues:

Date and time:

There were many discrepancies in the display date/time across formats, even though sent and received times matched up in the email headers. The MSG messages all replicated precisely the 'date sent' as given in the source PST, so comparisons in date and time were made between the MSG files and the three other formats.

The largest number of discrepancies occurred in relation to the EML messages. However, in every case except one, the discrepancy was *either* a precise number of hours ahead or behind the MSG time, *or* one minute ahead or behind. Although there were fewer discrepancies with the MHT files, these followed the same pattern. We therefore concluded that the majority of these discrepancies were probably due to timezone differences (e.g. timestamps based on local timezone information) or server errors (e.g. a failure to alter times when clocks change).

In order to resolve the issue, we decided that in our descriptive metadata, we would record both the original date and time and a normalised version in Coordinated Universal Time (UTC). This still cannot overcome server errors, which we discovered can in some cases result in the sent time of a message appearing to be later than the received time, so users of the archive will need to be warned about this.

Font:

We found that generally if a sender had selected to use a specific or non-standard font, it had been replicated in the migrated versions of the message. Where there were discrepancies between fonts, this appeared to arise from correspondents simply using the default font of their email client; different viewers or email clients may have different defaults. This was therefore not considered to be a major issue. On checking a cross-section of messages which had been converted to XML with full formatting markup, it was evident that information about fonts had been retained; how they appear when displayed depends on whether the fonts are installed on the viewer's computer.

Character encoding and special characters:

In the MHT format, problems were observed with the rendering of certain characters, including single quote marks, accented characters and Chinese characters. We pursued this by creating a new PST consisting of messages containing five Unicode (UTF-8) character test sets in order to re-test the migration to MHT. Scrutinising the MHT version in a text editor suggested that the encoding Internet Explorer uses to render the files was not configured to deal with UTF-8, so in order to ensure the messages display properly in this format, users would need to ensure that UTF-8 is turned on in their browser.

Hyperlinks:

We knew it would not be possible to retain any live hyperlinks to URLs which appeared in the email messages, due to the transitory nature of external web pages. However, we were keen to retain the full path of each URL. We found that although this does not actually display in every format, on inspection of the underlying files, it is clear that full URLs are retained.

Line breaks:

Hard line breaks had all been maintained in migrated versions, although soft line breaks inevitably varied. We did not consider this a problem because it was deliberate line-breaks (e.g. in poems) we wished to maintain.

7.4.2.3 Conclusion

Not every email format retained every single significant property we identified; however, considering *all four formats* of each email as a whole, we are confident that all properties have been preserved.

This is what prompted us to create just one PREMIS 'Representation' record for each email object, rather than treating each format as a different 'Representation' of the same message. See Section 7.5.3 for further detail on PREMIS metadata.

7.4.2.4 Customisation of Aid4Mail

Although Aid4Mail successfully migrated to all our desired formats, a major issue arose from the way the tool uses unique MD5 checksums to identify each message. Our testing revealed that these MD5 checksums were run on the migrated formats rather than the source format. For ingest into Fedora repository software[10], we needed to be able to associate all four formats of a single email message with each other because they form part of the same digital object. The obvious way of doing this was via the checksum, but if different checksums were generated for the same message in different formats, we would be unable to match them up.

Fortunately, Aid4Mail allows you to create your own scripting options, so we produced an Aid4Mail script which would:

- Break down a PST file into individual messages.
- Migrate these individual messages to MSG, EML, MHT and XML formats.
- Create a folder for each message which holds all four different formats of the same message, along with its attachments in their native format in a subfolder. The folder is named with a unique MD5 checksum.
- Create an XML file containing descriptive metadata about each email, also stored in the same MD5-named folder. See Section 7.5.2 for more information about descriptive metadata.

7.4.2 Risk analysis

Phase 3 of the project included scope for a small risk-analysis exercise based on file attachments. The purpose of this was to:

- Establish whether any of the formats included as file attachments are at risk now.

[10] http://www.fedora-commons.org/software

- Investigate any zipped or compressed files.
- Start work on a preservation plan which would include a technology watch schedule, and identify suitable migration formats for the future.

It also included some research into recommended preservation formats, and the policies of other institutions in this area.

To this end, we created a database containing information about approximately 45,000 attachments received in the first two accessions of the Carcanet Press email; these were all the files for which we had FITS[11] data at the time (see Section 7.5.4 for information about how we used FITS). The database included information about:

- The format and/or mimetype of every attachment, although we did encounter some discrepancies between the version information taken from FITS (and ultimately from the PRONOM[12] registry) and the mimetype, which we did not have time to investigate.
- Last modified date.
- Date of the email with which the attachment was associated (to supplement any information about last modified dates).
- An identifier for each file from the output of the FITS tool.
- The filename.
- A checksum allowing each attachment to be matched up with its email.

We discovered that a high proportion of all the files are in common formats like Microsoft Office documents, JPEG or PDF. We divided them into four classes:

- Those we knew we could read.
- Those we were confident we could read, but felt we should check (approx. 300).
- Zip files: we needed to test some of these so we could be confident that we could at least uncompress them.
- Those about which we had no information at all.

Using working copies, we investigated the very small quantity of potentially high-risk material, including zipped files and unusual formats. It was still possible to open most of these using current software, and we were able to open and view well over 97% of them. Potential significant issues were identified with only 470 of the 45,000 files.

Ultimately we opted to ingest zipped attachments as they were, but these will need to be revisited at a later stage of preservation planning. We also began compiling information about other institutions' policies on dealing with specific formats.

Unfortunately the timescale of the project prevented us from developing our own full preservation plan for this body of material. We recognise that these attachments are likely to include most formats we will encounter in other contexts, so they provide a crucial testbed for preservation planning activity. In future we hope to undertake a more formal preservation planning exercise using the Plato tool.[13] We have ensured that we are capturing sufficient metadata about file attachments to run post-ingest searches in order to identify the file information we require.

[11] http://projects.iq.harvard.edu/fits

[12] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[13] http://www.ifs.tuwien.ac.at/dp/plato/intro.html

7.5 Metadata

7.5.1 Object profiles

Capturing or creating sufficient metadata about every digital object is essential for compliance with ISO 14721:2012 – the Open Archival Information System (OAIS) Reference Model.[14] The OAIS Information Model defines different types of Information Package, each of which consists of:

- The digital object to be preserved.
- The metadata necessary to support its long-term preservation and access.
- Packaging Information, which relates the first two types of content.

OAIS stipulates that the Archival Information Package – the version of each digital object which is preserved and held in archival storage – should contain:

- Descriptive information, which supports discovery of the digital content. This does not have to form part of the AIP, rather it describes it, and so allows scope for there to be multiple sources of descriptive information or 'ways in' to an archival object. We are currently using several different types of descriptive metadata.

- Packaging Information, which binds all the components of an AIP into a single logical unit. This is supplied for us by Fedora software's FOXML schema, which wraps each digital object.

- Content Information: the digital object being preserved, and its associated Representation Information (i.e. information about what is necessary to render the object in a meaningful way). This is being stored using PREMIS metadata schemas.

- Preservation Description Information, which includes Reference Information (i.e. a unique and persistent identifier); Context Information (information about relationships between the digital object and others); Provenance Information (including the chain of custody and a record of any preservation actions); and Fixity Information (which validates the object's authenticity). We are storing most of this information in PREMIS; persistent identifiers are automatically assigned by Fedora software; and relationships are encoded using Fedora's RELS-EXT relationship metadata, as well as by indexing.

As indicated above, Manchester eScholar, which we are using to store the archive, is based on Fedora software. This is digital object based, and facilitates the straightforward creation of AIPs. We developed content models for five different types of AIP for ingest into Manchester eScholar. Inside each AIP there are several datastreams. See Appendix 1 for a diagrammatic representation of the five types of digital object or AIP we have developed. They are as follows:

- Collection object: the highest-level object. In PREMIS terminology, this is an 'Intellectual Entity', or conceptual object: essentially a description of a digital or hybrid archive which contains one or more accessions, multiple email sequence objects, each of which might contain many individual email objects. A collection object contains just 3 datastreams:
    - Descriptive metadata held in EAD.

[14] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284

    - Descriptive metadata held in Dublin Core (minimal; this datastream is required by Fedora software).
    - Structural metadata stored using Fedora's RELS-EXT relationship metadata. At collection-level this simply points to the 'Base' object (each object links to a 'base' record which defines the type of the object).

- Accession object: also an 'Intellectual Entity', representing a specific accrual of archive material with the same provenance, which might contain one or more Email Sequence objects. This type of digital object may not be used in all digital archives, and future guidance will take this into account; however, it reflects the way that the hard copy Carcanet Press Archive is arranged by accession. Currently an Accession object contains the same three datastreams as a Collection object. Its RELS-EXT metadata points to the Collection of which it is a part, and its Base object.

- Email Sequence object: we decided to define the content of each PST file as an 'email sequence' – essentially a snapshot of a mailbox taken at a particular point in time; this reflects the way we take in the hard copy correspondence of Carcanet Press in chronological sequences. Each email sequence object is an 'Intellectual Entity' (a sequence of email correspondence); a 'Representation' (a particular way of rendering that sequence of email correspondence); and a 'File' (i.e. an actual PST file). It contains 8 datastreams:

    - The PST file itself.
    - Technical metadata about it (output from the File Information Toolkit or FITS tool, in XML format).
    - Three datastreams containing different types of descriptive metadata: a basic Dublin Core record; an EAD record; and the HTML output of the PST Reporter metadata extraction tool.
    - An event log or audit trail, recorded using PREMIS. This is essential for demonstrating the authenticity of the object.
    - Preservation metadata about the PST, recorded using PREMIS.
    - Structural metadata stored using Fedora's RELS-EXT, pointing to the Accession, Collection and Base objects.

- Email folder (or Subfolder) object which is an 'Intellectual Entity' containing two datastreams:

    - Descriptive metadata held in a bespoke XML schema, based on metadata automatically extracted using Aid4Mail software; also a basic Dublin Core record.
    - Structural metadata stored using Fedora's RELS-EXT, pointing to its immediate parent folder, Sequence, Collection and Base objects.

- Email message object: an individual email, with one or more attachments where present: the email is an 'Intellectual Entity', containing several different 'Representations' (the email is preserved in four different formats, each providing a different way of rendering the message), and consists of several 'Files'. An email object consists of nine or more datastreams, depending on whether file attachments are present. These are as follows:

    - Four datastreams which each contain a version of the email in a different format: MSG, EML, XML, MHT.
    - A single datastream containing preservation metadata which relates to all four email files, stored using PREMIS.
    - One or more datastreams each containing a different attachment, where attachments are present.

- Where attachments are present, a datastream containing preservation metadata about each attachment, recorded in a FITS record which has been embedded into the PREMIS schema.
- A datastream containing descriptive metadata relating to the entire object (both email and attachment/s) stored in a bespoke XML schema based on the output of the Aid4Mail tool.
- A basic Dublin Core record.
- An event log or audit trail relating to the object, recorded using PREMIS.
- Structural metadata stored using Fedora's RELS-EXT, pointing to its immediate parent folder, Sequence, Collection and Base objects.

These content models were only finalised some way into the project, when we knew what was likely to be possible and had experimented with various different pieces of software.

See Appendix 2 for some screenshots of each type of digital object post-ingest, using the Fedora client software.

7.5.2 Descriptive metadata

Descriptive metadata is essential for resource discovery and we are using several different types of descriptive information. There is a Dublin Core datastream at every level, because this is required by Fedora software. However, this is a minimal record; more comprehensive metadata is stored using other schemas.

7.5.2.1 Collection, Accession and Sequence levels

We used the Encoded Archival Description (EAD) schema[15] to record descriptive metadata at all three of these levels, in its own dedicated datastream. This is the standard used to catalogue archives in traditional formats in the Library. Our way of implementing EAD as a separate datastream at each level was somewhat unorthodox; usually there is only a single EAD record for any archive, with lower-level descriptions nested in a hierarchical structure. Although EAD is not strictly designed to present 'standalone' metadata at specific levels in the way we have used it, it should be straightforward to automatically incorporate the standalone records into an overall archival catalogue in the future. Including an EAD datastream also enables us to conform to good practice in making digital objects self-describing.

We aimed to populate as much of the EAD as possible automatically. During the first phase of the project, we identified the free software, PST Reporter, as a reliable and accurate tool for extracting key descriptive metadata about a PST file, including: folder names; email and attachment counts; email and attachment sizes; number of read and unread items; sender names; email addresses; and number of messages/attachments. Its output was so rich that we decided to include the PST Reporter record as a datastream in its own right at email sequence level. However, we also used it to extract key pieces of metadata to populate an EAD record. Alongside this we used the commercial software Aid4Mail, which extracts descriptive metadata as well as undertaking format migration.

For the sequence-level EAD record, we used the output of both PST Reporter and Aid4Mail to automatically populate certain elements, including: covering dates; extent; material specific details (primarily used to record the mimetypes of attachments identified by Aid4Mail); and, in the <scopecontent> element, details of the folder structure of each PST file – including folder names and paths, size, covering dates, and number of messages and attachments in each folder.

[15] http://www.loc.gov/ead/

Part of the reason for using two different tools was for verification purposes. Due to the slightly differing ways in which PST Reporter and Aid4Mail interpret the content of a PST file, there are often slight variations in their output. We therefore produced some code which would record total numbers as calculated by each tool, and a list of any folders which appeared in the output of one tool but not the other.

The rest of the sequence-level EAD record was edited manually by the archivist.

At Accession level, some of the EAD elements are populated automatically using metadata extracted by Aid4Mail (primarily quantitative data like extent and covering dates); the rest are filled in manually. The Collection-level EAD record was created entirely manually (and is edited manually when new accessions are added to the archive).

We discovered that both PST Reporter and Aid4Mail can alter the checksum reading of a PST file simply by connecting to it, so we only ever run these tools on an exact working copy of the original PST file.

7.5.2.2 Folder and Email level

Folders and subfolders are stored as separate digital objects in eScholar. Their descriptive metadata is stored in a bespoke schema based on our email level schema (see below), alongside a minimal Dublin Core record.

For email-level descriptive metadata, we explored various different options. Initially we felt that only minimal information was necessary, and this could be stored in basic Dublin Core. However, in light of the very rich information captured by Aid4Mail, we decided that email-level descriptive metadata could potentially be much fuller, as long as it was based on data that could be extracted automatically by Aid4Mail.

Ideally, we hoped to find a recognized metadata standard or schema for email messages, which we could populate with Aid4Mail metadata (the output of Aid4Mail does not conform to any documented standard or have the status of a schema). We investigated some other projects and tools. We found the documentation of the Collaborative Electronic Records and the Electronic Mail Collection and Preservation Projects (CERP and EMCAP) [16] very useful, but their XML schema focuses on account-level preservation and description. Xena software has a defined XML schema for individual emails, but it does not allow for the detail of the Aid4Mail output and is also quite verbose. We discovered some other putative schemas for email, including one which has been released as a schema for exchanging email (see http://www.molengo.com/emailxml/title/email-xml-schema-xsd#schema). However, ultimately we chose to adapt and use Aid4Mail's native XML metadata output. It is rich, granular and structured – lending itself to indexing – and includes: message ID; display, sent, received and stored dates; names, display names, and email addresses for senders, recipients (including cc and bcc recipients), and anyone else associated with a message; subject line; attachment names and mimetypes; filepaths, including the folder to which a message belongs; any flagging; digital size; and priority level.

One problem arose from the fact that Aid4Mail's XML metadata was pulling its content from potentially non-XML-compatible sources, leading to issues around validating the XML. We therefore had to develop a separate process to ensure that all of the XML was encoded as UTF-8, and to strip out invalid XML characters; the latter were all control characters with no visible rendering.

[16] http://siarchives.si.edu/cerp/

7.5.3 Preservation metadata

7.5.3.1 PREMIS content models

Detailed metadata profiles were put together based on the PREMIS Data Dictionary, which is recognized as a key digital preservation standard. Although PREMIS does not dictate that content must be stored in any specific format, we were keen to use the dedicated PREMIS XML schemas, as the Fedora repository system is particularly suited to managing structured data in XML.

On experimenting, it became apparent that the schema rules did not allow us to use PREMIS in quite the way we had envisaged based on the Data Dictionary. We also had extensive discussions about the level/s at which PREMIS data should be stored, and whether it was remotely practical to include PREMIS records at individual email level.

Ultimately, we decided on the following approach, and produced PREMIS templates for each scenario:

At sequence (or PST) level:

- Each digital object has a PREMIS 'Object' record containing metadata about the email sequence, both as a 'File' (metadata specifically about the PST format), and as a 'Representation' (reflecting the fact that a PST is only one way of representing that sequence of email).

- Each object also has a separate PREMIS 'Event' record – i.e. an audit trail or event log recording all actions which have impacted on that object; this is key for ensuring authenticity and demonstrating provenance.

At individual email level:

- Each email object (which contains versions of the same email in MSG, EML, XML and MHT formats) has a single PREMIS 'Object' record relating to all four email formats. This contains 'Representation' metadata for the email – primarily a record of the significant properties to be retained. This is perhaps a slightly unorthodox interpretation of PREMIS, but we chose to record the significant properties we hoped to retain by preserving the email in *all four* different formats, whilst recognizing that in some cases migrated formats are unable to retain *all* desirable properties. Within the same PREMIS record, there is metadata about each of the four different formats as 'Files'.

- Each object also has a separate PREMIS 'Event' record – i.e. an audit trail or event log recording all actions which have impacted on that object as a whole.

- Each attachment also has a dedicated PREMIS 'Object' record; this is populated automatically by pulling data from the FITS output for each attachment (see below) and inserting it into PREMIS using the <objectCharacteristicsExtension> element as a container. Originally we planned simply to store the FITS output for each attachment as a separate datastream, but there was no straightforward way of directly associating the FITS record with the attachment to which it related. It is possible to link a PREMIS record directly to the relevant attachment using the datastream PID. We also produced a PREMIS template for migrated file attachments, in anticipation of future preservation actions such as migration.

7.5.3.2 Populating PREMIS records

Although we hoped to use the PREMIS schemas as they are, we ultimately chose to produce a slightly modified, local, version of the schema for use in two contexts:

a) For the event log at all levels: the schema is modified so that we can include drop-down menus for some elements containing locally-defined controlled terms.
b) For the attachments at email level, where we wanted to omit some elements which are defined in the PREMIS schema as mandatory but were not relevant in this context; the local schema overrides the mandatory rules.

This is not an ideal solution, and means that our local schema needs to be stored in a referenced location which will not change. However, all other PREMIS records were based on the standard PREMIS schemas.

We used a combination of methods to populate the PREMIS records, consisting of:

- The creation of PREMIS templates, which included some hard-coded content where information was common across a whole class of objects.
- Automated creation of some elements, notably the PIDs and post-ingest checksums.
- Pulling some content from the output of FITS using an XSLT.
- Manual entry, in some cases with the assistance of drop-down menus representing locally developed controlled vocabularies.

Currently, populating PREMIS records is largely a manual process, and one which could benefit from some degree of automation.

7.5.4 Technical metadata

Technical metadata was extracted using the File Information Tool Set (or FITS tool), which identifies, validates and extracts technical metadata for a range of file formats. It acts as a wrapper, containing the output from several other open source tools, including JHOVE and DROID. Outputs from these tools are converted into a common format, compared to one another and consolidated into a single XML output file, which can be stored as a Fedora datastream if desired.

At sequence level, we retained the FITS XML output as a datastream inside the Fedora sequence-level digital object.

At individual email level, we ran FITS only on attachments, and – using an XSLT – imported the whole of the FITS record into the PREMIS schema using the <objectCharacteristicsExtension> element.

It was not considered necessary to run FITS on the various formats of each email, as these had essentially been created in-house and we had full format information. Instead, we simply ran Jacksum software on each email format in order to obtain an MD5 checksum reading and filesize; this was imported into the PREMIS file.

7.5.5 Structural metadata

It is possible to represent structural relationships between digital objects in Fedora. This is done by including a datastream within each object containing RELS-EXT ('Relationships-External') metadata. Each object can have one RELS-EXT datastream, which is used exclusively for asserting object-to-object relationships; these are encoded in XML using RDF (Resource Description Framework).

For practical reasons, we used RELS-EXT to express upwards relationships only for each digital object type, e.g. an individual email points to its immediate parent folder; email sequence; and collection.

However, as we were dealing with quite complex hierarchies of folders, we used the powerful indexing offered by Apache SOLR[17] (see Section 7.6.2) to express relationships both upwards and downwards throughout the hierarchy, facilitating navigation through the entire collection.

7.6 Ingest and archival storage

7.6.1 Ingest

Throughout our processing, we ensured that the archival material, along with all supporting files (such as metadata outputs) were stored in a pre-defined folder structure, with careful naming and versioning conventions. This facilitated the preparation of files for ingest.

All of our digital objects are wrapped in Fedora's FOXML metadata schema, which associates all the datastreams making up an object.

XSLTs were written which created FOXML files for all five types of digital object in the collection.

To ingest the FOXML and datastreams for each object, these files were uploaded to a Fedora-enabled server. This is time-consuming, and is best carried out by compressing the files and uncompressing them after transfer. The actual upload was carried out using SFTP client software.

Batch ingest into Manchester eScholar was undertaken using the Fedora client software. This was also a time-consuming process. At email level, we divided the material into batches of 20,000 digital objects, which would each take approximately 48 hours to run. These batches were based on estimated ingest time only, and did not reflect any logical divisions in the content of the material.

Once the ingest had completed, the Fedora client software provided a summary of the number of successfully ingested objects, the number of failed objects, and the time taken. The software also produced a detailed log of ingested objects which was used to double-check the ingest and resolve any issues.

Various problems and issues arose, most of which we managed to resolve within the project's timeframe. These included the following:

- As with the XML metadata, there were problems with non UTF-8 characters occurring in emails which had been migrated into XML format. In virtually all cases, these were characters that are reserved in XML but have no visual representation (e.g. additional spaces or linebreaks). A process was developed to convert all XML to UTF-8 in order to make it validate.

- The incredibly time-consuming nature of some of the processes, including cleaning the XML, the transfer of all the files to the ingest server, and the ingest itself. The average time for ingesting a single email-level object was nine seconds. Our method of undertaking the ingest in batches was perhaps not the most efficient solution; as we discovered, any unexpected problems (like the failure of an object to ingest, or a network outage part-way through) could result in the process grinding to a halt without our knowledge, thus wasting time unnecessarily. In future, it might be possible to process everything to the point of creating the FOXML metadata, and then ingest one object at a time; this way, any failed objects would be logged as such but the job would still proceed.

[17] http://lucene.apache.org/solr/

- Objects which failed to ingest: initially it seemed as if several hundred objects had not ingested, but on investigation, most of these were due to indexing issues; the objects had in fact been ingested but had not been indexed so were not recorded in the SOLR index as being present. This problem was rectified by 'touching' the index for these failed objects – i.e. running an XSLT containing commands which reindexed the PIDS of the affected objects.

Ultimately, only 45 of over 282,000 digital objects that we tried to ingest failed, although the number of attachments which did not ingest was considerably larger, running to hundreds. We need to undertake some further investigation into this, but obvious explanations include Aid4Mail failing to extract certain attachments, and potentially rogue characters in file titles, e.g. double spaces or foreign characters which are not rendering properly; this might result in the system failing to locate the files during ingest.

We have developed a process for identifying and investigating failed objects and attachments, and subsequently re-ingesting them if necessary, although the actual work in this area remains to be done. It will involve weighing up the archival or research value of the failed files against the labour-intensive and time-consuming process of re-processing and re-ingesting them.

7.6.2 Indexing

We developed requirements for searching and interrogating the archive post-ingest, based both on typical 'traditional' research enquiries received about the Carcanet Press hard-copy archive, and on the opportunities for different types of access opened up by the digital environment. These ranged from basic queries to more complex faceted searches; a very small sample of the scenarios we identified includes searching for:

- an individual email message, attachment, folder or sequence;
- all email messages in a particular folder;
- all email messages with file attachments;
- all file attachments with a particular mimetype;
- all email messages received during a particular date range;
- all email messages sent to and received from a particular correspondent;
- all email messages with a particular word or phrase in the message body or subject line;
- the frequency of email messages sent by different individuals.

These requirements were taken into account when building the index for the email archive. Indexing was done using Apache SOLR, which enables both simple and faceted searches on specific metadata fields. All of the XML metadata produced by Aid4Mail was indexed, as well as plain-text versions of message content, meaning that Google-style full-text searching can be carried out on the content of messages.

While most of the indexed elements are intended for search and discovery purposes, some elements have been indexed for curatorial and preservation purposes, e.g. technical details about file attachments from PREMIS, and checksums from PREMIS and FITS.

We had debated whether attachments should be stored as separate digital objects in their own right – partly because there were potential issues about being able to search on attachment titles or count the total number of attachments post-ingest if they were stored as datastreams embedded within digital objects. Ultimately these problems were overcome through the index, which treats attachments as separate objects even though they are not stored as such in Fedora.

On ingest, all the fields identified to be searchable were pulled out using an XSLT and put into the SOLR index mechanism. Overall the index forms a highly powerful tool for interrogating the material, and carries out lightning-quick searches, because these searches are based on text that is present in the index itself in XML form, so searching inside the digital objects themselves is not necessary.

7.6.3 Management and access

Developing an access interface for researchers was not part of the project's remit – although enabling the archivist to interrogate and manage the material was identified as a priority.

Currently, direct access to the digital objects is only possible using the Fedora client. Fedora can display any datastreams which are in XML, but for viewing metadata, or content, in any other format, the user is directed to a secured URL from where the files can be downloaded, depending on whether the user has the appropriate software installed. However, currently access to the Fedora client is restricted to a very small number of system administrators.

The SOLR index mitigates this problem to an extent: its content is so comprehensive that in many instances, direct access to the datastreams is unnecessary. However, interrogating the SOLR index involves mastery of a specialised syntax, and search results are returned in raw XML form. This restricts its usefulness for the managing archivist.

We therefore developed an initial functional design for a curatorial tool, which would overcome many of these problems. The completion of this, and the actual development of the tool could not be carried out within the project's timescale but it has been identified as a priority for completion in 2014.[18]

The three figures on the following pages show a mock-up of how we envisage that this tool may look. It will enable sophisticated searching and filtering of search results. A single Google-type search would look across all indexed fields in the collection, but searches could also be limited to subject line or the body of emails only. Search results could be filtered on various parameters, e.g. date range, sender, recipient, object or datastream type. The curator could tag search results, and save the results of specific searches, which can be useful for dealing with common enquiry types.

There would also be a viewing pane, which could display an editable version of the whole datastream for all metadata stored in XML format; this could be used to amend individual elements or replace the whole datastream with an updated version. It could potentially include customised views of a single datastream, e.g. XML metadata to which a stylesheet has been applied so it renders in user-friendly HTML form (although this would be a read-only option). It could also give direct access to related objects, such as the parent sequence, accession and collection, or the child folder and email message objects – enabling navigation throughout the hierarchy.

The viewing pane could also offer the option of downloading 'managed' files, such as email messages (in MHT, MSG or EML format) and file attachments. In these instances, the viewing pane would present metadata about the files (name, mimetype, size, created and modified dates), and a hyperlink to a secure URL which would enable the curator to download and view a copy of the file, as long as the appropriate rendering software is installed on their PC (e.g. for a Word file, this would be MS Word, or any other software that can open and display a Word file). The downloaded file would only be a copy of the archived version.

[18] At the time of making this report available in Manchester eScholar (early June 2014), development work on the curatorial tool is in its initial stages.

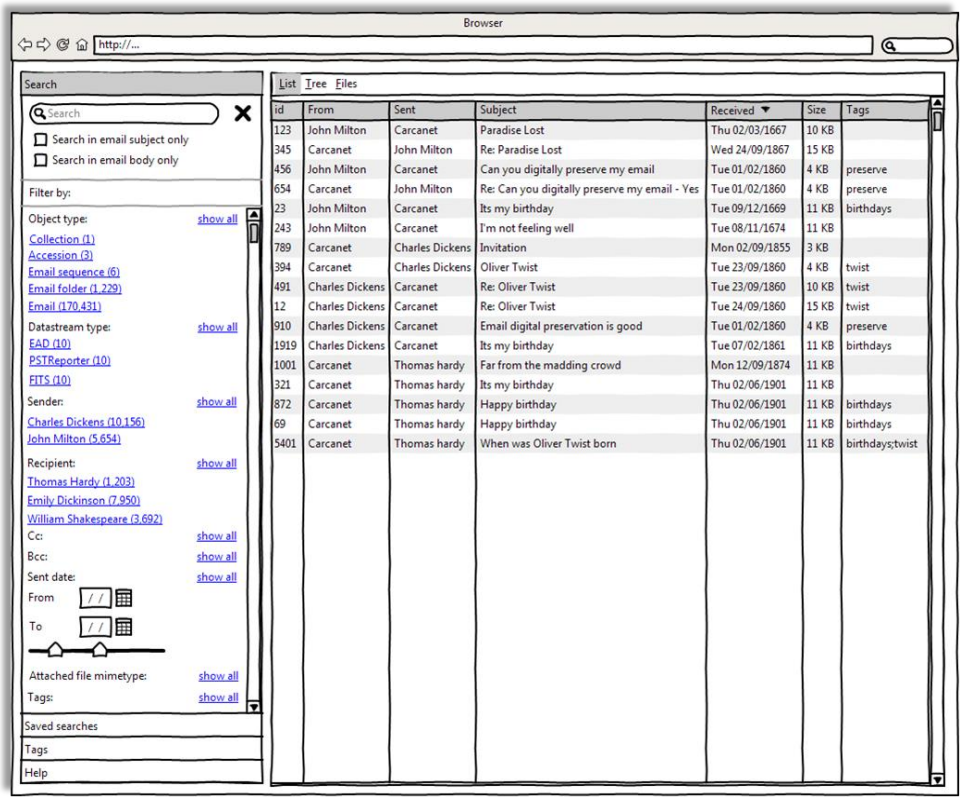


Figure 1: This shows the results of a search in 'List' view.

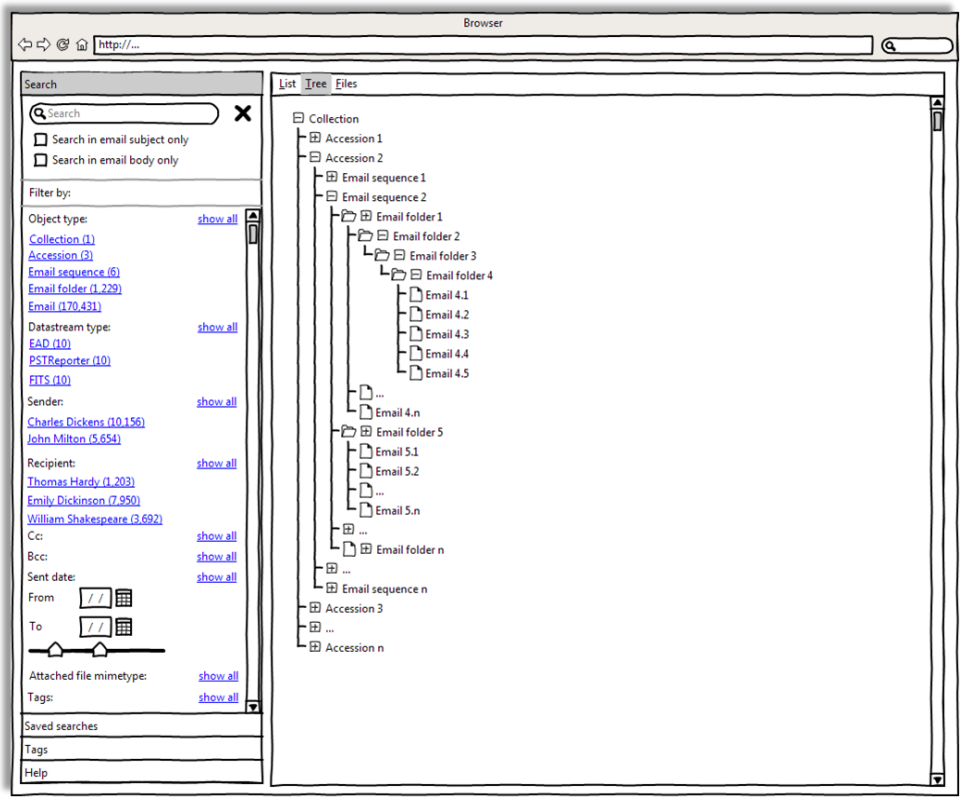


Figure 2: Here, search results are displayed in Tree view, reflecting where they sit in the overall hierarchy of the archive.

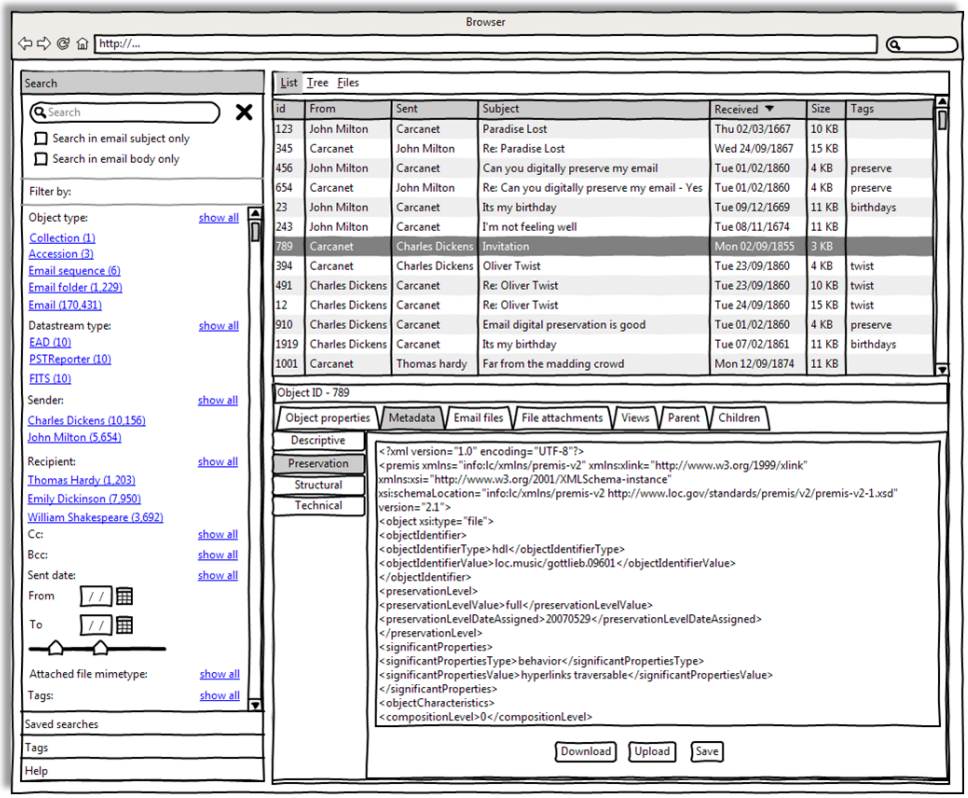


Figure 3: This displays both search results in List view, and the viewing pane, in which certain XML datastreams associated with any single digital object could be viewed, and which would also offer the option of downloading non-XML versions of files (e.g. email files and attachments) via a secured URL.

7.7 Manual and workflow documentation

Due to its highly specialised nature, we recognized the importance of fully documenting every stage of our workflow. We have produced an in-house *Manual for Processing and Ingesting Archival Email* which extends to 120 pages and is based around the workflow steps shown in Appendix 1. This has been refined throughout the project, and was road-tested when the archivist took responsibility for processing the third accession of email in 2013 up to the point of ingest itself, which can only be carried out by a system administrator.

The full suite of code, including XSL transformations and batch files as well as templates, has been submitted to our central access-restricted repository of master code for re-using in future projects.

7.8 Access and interpretation

7.8.1 Researcher consultation

Although the project's focus was on preservation rather than access, we were aware that decisions taken at an early stage can make a difference to how people are able to access and use digital archive material at a much later stage. All of our work therefore tried to take account of how our end users (or 'designated community' as it is termed by OAIS) might wish to interrogate and access the collection in future. Our planned curatorial tool could potentially also form a prototype from which we could develop an access interface.

Although the scope of our project was very limited, we were keen to gauge opinions from researchers on how they might envisage making use of email archives in their field, and to this end, carried out some one-to-one interviews. We interviewed nine people in total, as follows:

- Two academics from the University of Manchester's English Language and Linguistics discipline who have carried out corpus studies based on archival correspondence in hard copy form.
- An academic from the University's Drama department who has made use of paper-based and sound archives in research.
- An academic from the University's Sociology department who has carried out research based on social media.
- A poet (published by Carcanet Press), former academic, now independent researcher and writer, who is a scholarly editor, literary researcher and biographer, and has written on textual criticism.
- An academic in literary studies, literary biographer, and director of an author-based research institute and archive.
- An academic from the National Centre for Text Mining (NaCTeM).
- A PhD student from Italian Studies, whose doctorate focuses on the Carcanet Press hard copy archive.
- A former member of staff from Carcanet Press who has edited a collection of literary correspondence.

Participants were sent a brief questionnaire in advance of the interview, and this guided the discussions; they were also shown slides indicating some of the issues, and showing ways in which archival email might be used or represented, including some basic visualisation experiments. The interviews were otherwise relatively unstructured in nature.

Some of the key opportunities identified by participants for working with email archives included:

- Full-text (Google-style) searching.
- Searching based on keywords, topics, or issues.
- Searches based on senders, recipients, those appearing in the cc field and the bcc field of emails (although the last of these options raises issues of privacy and confidentiality).
- Searching for the names of individuals, not just as senders/recipients, but also as 'entities' (i.e. individuals being mentioned in emails between other people).
- The possibility of determining dates and times at a much more granular level (this was identified as particularly important by those working on literary biography or specific writers).

The participants were shown some different ways of representing data in email archives, such as network graphs. There was particular enthusiasm for:

- Visual representations of networks around particular projects, publications or collaborations.

- Quantitative representations of emails exchanged between particular correspondents over time.

When asked about how they would like to discover and access archival email, around half of the researchers (probably those who had used archives in the past) still valued the traditional online archival catalogue which offers a *context* for email correspondence, rather than being taken to that correspondence directly via a separate interface.

Most participants had no strong opinions on ways of viewing and navigating around an email archive, being more interested in the content of individual messages. There was no great enthusiasm for experiencing emails in an interface which aimed to represent a particular email client at a particular time; in fact, at least two participants expressed a preference for an interface that was deliberately neutral, and not attempting to artificially reproduce something that looked 'authentic'.

Some of the examples shown to participants aroused suspicion, notably some visualisation tools based on text mining which depict levels of different 'emotions' expressed in a particular body of email.

Perhaps surprisingly, only two participants brought up the issue of ethics, data protection, privacy and sensitivity associated with email correspondence – subjects which loom large for the curators of such collections.

We are bearing the researchers' comments in mind when thinking about how to develop access tools.

7.8.2 Visualisation and text mining

Our SOLR index constitutes a huge body of structured email data, in XML format, which can be used as the basis for interrogating, interpreting and representing the archive in different ways, lending itself in particular to visualisation and different types of text-mining.

The project's remit did not extend to investigating new and innovative forms of access like this, but we did undertake some basic visualisation experiments.

Based on researchers' enthusiasm for graphical representations of incoming and outgoing messages, we produced some basic graphs using Excel, showing Michael Schmidt's email correspondence with four different correspondents over a specific time period (see Figure 4). The bars above the line represent his outgoing messages and those below the number of messages received from the same correspondent. These provide a useful visual summary of an email correspondence, which is more immediately understandable than a simple list of figures. They reveal obvious peaks and troughs which may be immediately meaningful to a researcher working on a specific writer or publication. They also reveal degrees of mutuality in correspondence (which can sometimes be lacking, as in one of the examples below).

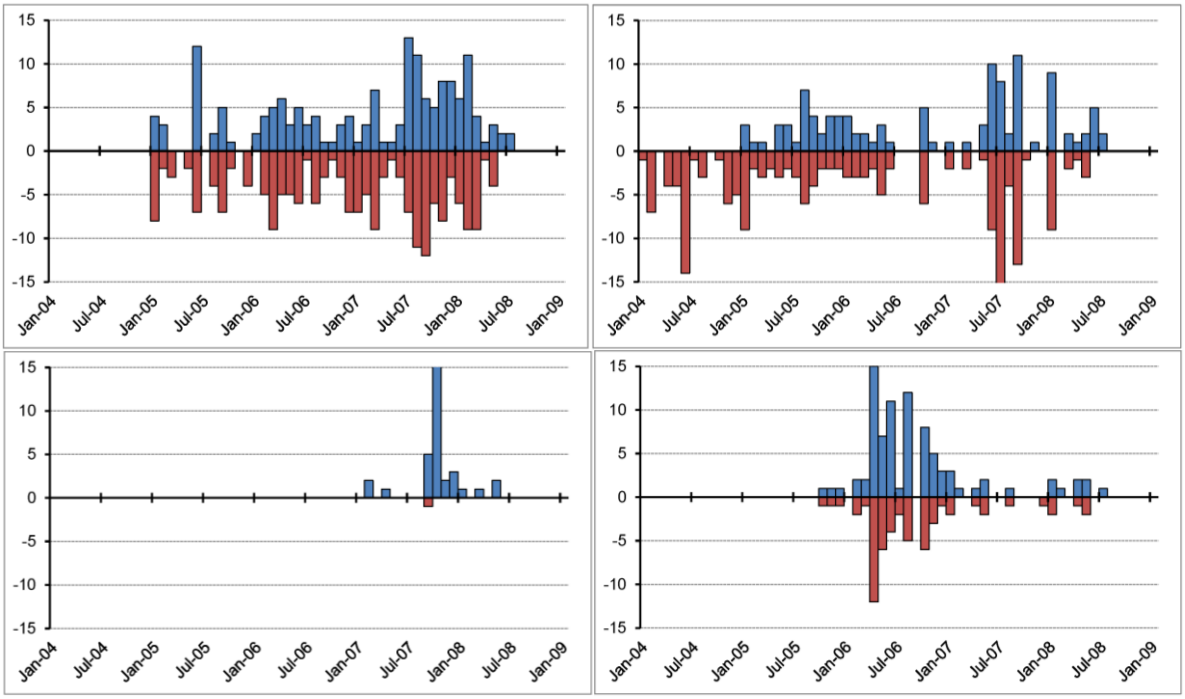


Figure 4: quantitative representation of Michael Schmidt's in/out emails with four different correspondents.

We also experimented with network graphs using free Gephi software.[19] It runs various different algorithms, all of which can produce multiple representations of the source data.

In the examples below (Figures 5-8), the nodes are individual email addresses, with the lines (or 'edges') representing both direct and indirect relationships between them. Although not visible in the figures reproduced here, the edges have arrows at the end indicating the direction of the relationship. A direct one-way relationship is formed when a correspondent appears in the direct 'To' field of an email; a direct two-way relationship is formed when the recipient replies directly. Indirect relationships are formed between correspondents who are recipients of the same email, or who appear in the cc field of an email. The network graphs reveal many simple one-to-one relationships, but there are also numerous small groupings where two or three individuals participate in the same ongoing thread of correspondence. Larger groupings tend to represent distribution lists, and in some cases a single individual links two otherwise distinct groups.

In theory, there is no limit to what can be represented by these diagrams: a node might be a person, a keyword, a date, a phrase, or a book title. Network representations could be datestamped in order to obtain snapshots of a particular network at different points in time, showing entries into and exits from a network. It may also be possible to filter diagrams after their creation, much like filtering search results using faceted searching, e.g. a graph showing connections between certain individuals could be filtered to show only the emails within that diagram which contain a particular book title. Graphs could focus on as small or as large a body of data as desired; one of our examples alone contained 10,104 separate email addresses and 1.3 million relationships between those email addresses.

[19] https://gephi.org/

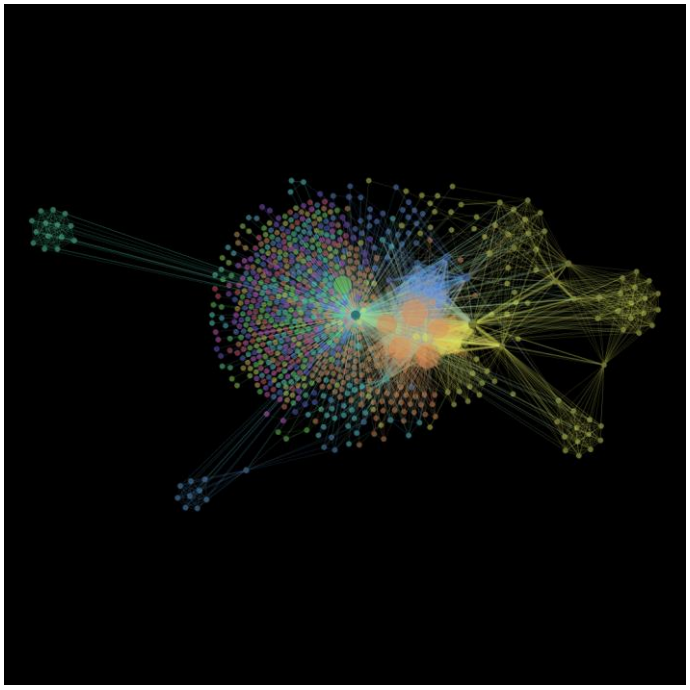


Figure 5: Network graph based on a PST file which had been created by the Editorial and Managing Director of Carcanet Press; it contains 3,536 messages, dates from 2001-2003, and represents the 'Sent Items' folder only.

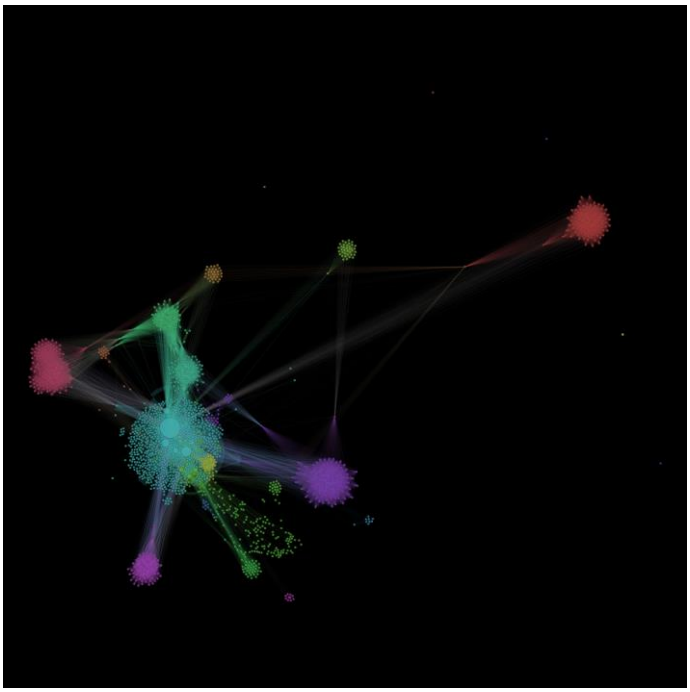


Figure 6: Network graph based on a PST file created by the Editorial and Managing Director; this contains 8,275 messages, dates from 2001-2004, and lacks the 'Sent Items' folder.

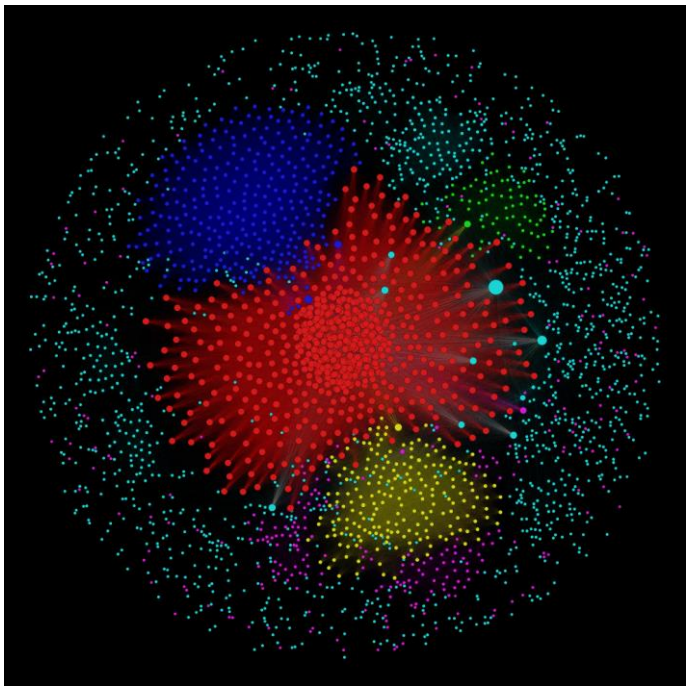


Figure 7: Network graph based on the same PST file as Figure 6, but created using a different algorithm.

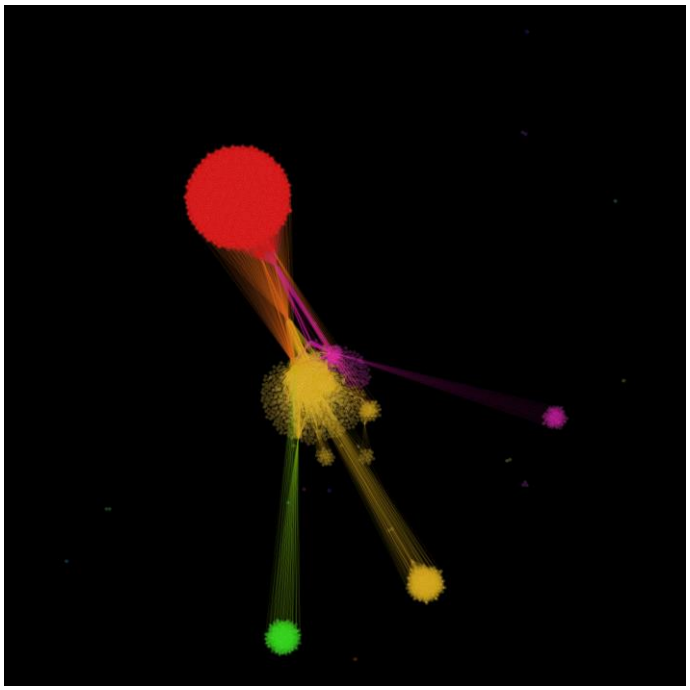


Figure 8: Network graph based on a PST file created by the Managing Editor at Carcanet Press; this includes both sent and received items, dates from 2002-2012, and contains 9,909 messages.

In addition to visualisations like this, we identified text mining as having great potential both for curators (e.g. to assist in the creation of descriptive metadata about a collection) and for researchers (offering new ways into a collection, e.g. by topics). Our own efforts did not extend beyond some basic Wordle diagrams like the following:

This was based on a relatively small body of test data, so is perhaps not the most useful of tools ('poems' being a somewhat predictable term), but already certain names and book titles are being identified, and there is scope for producing much more sophisticated word clouds than this.

On speaking to staff at the National Centre for Text Mining, it became clear that with the right privacy controls in place, our structured email data could provide an interesting source for text-mining; it would be a contrast to much of the work done by NaCTeM, which has focused on published sources, often of a scientific nature. Text mining also opens up whole new ways of drawing meaning from an archive; its algorithms can highlight concepts based on dictionaries which are actually derived from the body of data itself.

Overall, we felt that visualisations and text-mining techniques could offer several opportunities and advantages, e.g.:

- They can summarise quantitative information in a much more immediate and understandable way than would ever be possible manually.
- Issues of data protection, confidentiality and copyright mean that the content of email archives cannot be made generally accessible to researchers without prior assessment by an archivist. However, these new techniques potentially offer different ways of providing meaningful access without releasing the full content of messages.
- They have potential for opening up new avenues of research for those working in literary studies and the humanities more generally, e.g. constructing and mapping literary networks or the evolution of literary movements; revealing behavioural trends and changes in the way people communicate in the digital environment.
- They can provide archivists with new ways of managing and curating email archives.
- They can assist with administering Data Protection by highlighting particular types of sensitive or personal data in a collection.

7.9 Enhancing knowledge and understanding

One of our objectives for Phase 3 of the project was to update our knowledge and understanding of the current landscape in the field of email preservation. This was partly done via personal contacts and visits – to the British Library (where Wendy Cope's email archive is being preserved) and the Bodleian Library (where staff are dealing with email in other formats).

Our Digital Preservation Co-ordinator was also able to undertake a short literature review, starting with the bibliography of Chris Prom's 2011 DPC Technology Watch Report, *Preserving Email*[20], and then undertaking a review of her own.

This uncovered some useful articles and projects which we are following up. Some of these we had already drawn on (such as the CERP/EMCAP work, and the InSPECT Project), but others were new to us. We were particularly excited by the work going on at Stanford University as part of their ePADD Project,[21] which is looking at innovative ways of representing and managing the content of email archives.

7.10 Publicity

Several papers and talks were given in different contexts during the lifetime of the project. As well as in-house events, these included:

- Fran Baker, 'Carcanet Case Study', JISC SPRUCE Project event, London School of Economics, 19 January 2012.
- Fran Baker, 'Emails to an Editor: Preserving the Digital Correspondence of Carcanet Press': Archives and Records Association Section for New Professionals Summer Seminar, 17 August 2012.
- Fran Baker, 'Carcanet Press Email Preservation Project', for the Advisory Board of the Bodleian Library's futureArch Project, 7 September 2012.
- Fran Baker, 'Emails to an Editor: Preserving the Digital Correspondence of Carcanet Press': Group for Literary Archives and Manuscripts, 12 September 2012.
- Fran Baker, 'Emails to an Editor: Preserving the Digital Correspondence of Carcanet Press', Beyond the Text: Literary Archives in the 21st Century, 26-27 April 2013, Beinecke Rare Book and Manuscript Library, Yale University.
- Fran Baker and Phil Butler, 'Email Archives: New challenges and opportunities for Digital Humanities research', Digital Humanities @ Manchester conference, 8 November 2013.
- Fran Baker, 'Update on the Carcanet Press Email Preservation Project', for the Steering Group of the Modern Literary Archives Programme, John Rylands Library, 6 December 2013.
- Fran Baker, 'The Email Explosion: preserving the digital correspondence of Carcanet Press', article published in *PN Review* 216, Vol. 40, No. 4 (March-April 2014).

[20] Christopher J. Prom, *Preserving Email*. DPC Technology Watch Report 11-01 (December 2011) http://www.dpconline.org/component/docman/doc_download/739-dpctw_11-01.pdf

[21] http://library.stanford.edu/spc/more-about-us/projects-and-initiatives/epadd-project

8. Lessons Learned

Lessons learned as a result of these two phases of work include the following:

- The importance of knowledge transfer and detailed documention of every process cannot be overestimated: this is such specialised work that the loss of a key staff member could jeopardise progress (and in fact, the key technical member of the team has subsequently moved to another role elsewhere in the University). We hope that our detailed manual, awareness-raising work, and training sessions will help to mitigate this issue. Work is also overseen by the Library's Digital Preservation Steering Group to ensure continuity.

- The work has given us a more realistic idea of the time and resources required to implement digital preservation, and we are encouraged by the fact that so much has been achieved by existing staff with no additional resource.

- Some of the technical processes are extremely time-consuming. We have identified certain areas for improvement, including further automation of some currently manual processes.

- While preserving to the level of granularity that we have done is the right thing for the Carcanet Press Archive, in other contexts we may choose to preserve archival emails differently. There is no 'one size fits all' approach – and the modular processes we have developed will hopefully provide a useful basis for adaptation and development.

- Our literature review and work with researchers has given us an insight into how much research potential there is in an archive like this one – we are now aware of research uses that we had not necessarily considered at the outset of our work.

9. Opportunities and Next Steps

We have achieved a great deal during a relatively short timescale, working with limited resources, and are keen to build on the work we have completed to date. Some of the essential next steps and potential opportunities we have identified are as follows:

- An immediate priority is the design and building of a curatorial interface, which will enable the archivist to manage the Carcanet Press Email Archive.
- Our work provides an important building block for one of the University of Manchester Library's proposed strategy projects, which focuses on building expertise in the long-term preservation of email.
- Other key areas of work we are keen to pursue in the field of email archiving include:
    - fine-tuning and automating more of the preservation processes;
    - refining our ingest process to mitigate against problems like unexpected network outages or reboots;
    - further exploring techniques like visualisation;
    - working with email in different formats from different creators;
    - developing the curatorial tool for potential use as a user interface;
    - creating additional header information for emails in MHT format, which would enable them to be viewed in a standard web browser;
    - extending indexing and searchability to file attachments;

    - enabling the tracking of emails within particular correspondence threads;
    - exploring how researchers should encounter descriptive metadata, e.g. look into linking descriptions in our archive catalogue database with metadata records and actual content held in our Fedora repository.
- The large body of file attachments we have captured provides a basis for extending our digital preservation activities beyond email; this will be key as our digital collections continue to grow. The processes we have developed are modular, so can be adapted to meet digital preservation requirements for other born-digital content.
- We hope to undertake more formal preservation planning using the Plato or other appropriate tools to ensure the attachments remain accessible over time.
- We will also be developing workflows and processes for managing hybrid archives – i.e. harmonising and joining up how we deal with hard copy and digital components of the same archive.
- We hope to contribute to the rapidly developing area of digital humanities – both by opening up new avenues for academic research, and through our expertise in issues of preservation and sustainability.

10. References

Reports and papers

- Fran Baker, Phil Butler and Ben Green, *Carcanet Press Email Preservation Project Report* (University of Manchester, 2012): the report of the JISC-funded initial phase of the project. Available at: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:165096 [Accessed: 3/4/14].

- Jeremy Leighton John, *Digital Forensics and Preservation*. DPC Technology Watch Report 12-03 (November 2012). Available at: www.dpconline.org/component/docman/doc_download/810-dpctw12-03pdf [Accessed: 3/4/14].

- Gareth Knight, *Significant Properties Testing Report: Electronic Mail* (30 March 2009). Available at: http://www.significantproperties.org.uk/email-testingreport.html [Accessed: 3/4/14].

- Christopher J. Prom, *Preserving Email*. DPC Technology Watch Report 11-01 (December 2011). Available at: http://www.dpconline.org/component/docman/doc_download/739-dpctw_11-01.pdf [Accessed: 3/4/14].

Projects

- Collaborative Electronic Records Project and Electronic Mail Collection and Preservation Project (CERP and EMCAP): http://siarchives.si.edu/cerp/

- ePADD Project: http://library.stanford.edu/spc/more-about-us/projects-and-initiatives/epadd-project

Metadata and other standards

- Dublin Core: http://dublincore.org/
- Encoded Archival Description: http://www.loc.gov/ead/
- ISO 14721:2012 – the Open Archival Information System (OAIS) Reference Model: http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284
- Preservation Metadata: Implementation Strategies (PREMIS): http://www.loc.gov/standards/premis/

Software and registries

- Aid4Mail by Fookes Software: http://www.aid4mail.com/
- EMu software by KE Software: http://emu.kesoftware.com/about-emu/overview
- Fedora Commons: http://www.fedora-commons.org/software
- File Information Tool Set (FITS): http://projects.iq.harvard.edu/fits
- Jacksum: http://www.jonelo.de/java/jacksum/
- Kernel Outlook PST Reporter: http://www.nucleustechnologies.com/outlook-pst-reporter.html
- PeDALS Email Extractor: http://sourceforge.net/projects/pedalsemailextr/
- PRONOM: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
- TrueCrypt: http://www.truecrypt.org/
- Xena: http://xena.sourceforge.net/

Appendix 1: workflow and object profile diagrams

These diagrams illustrate the workflows we have developed and the structure of each type of digital object or AIP in the Carcanet Press Email Archive.

Figure 1: Email Sequence

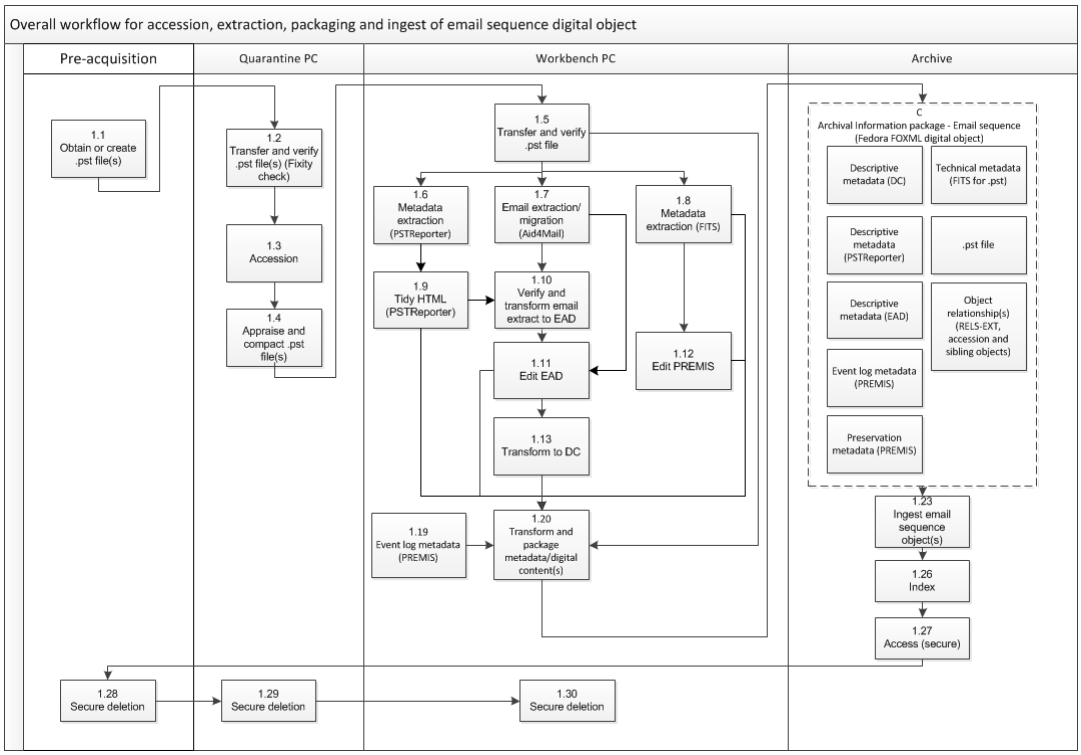


Figure 2: Accession

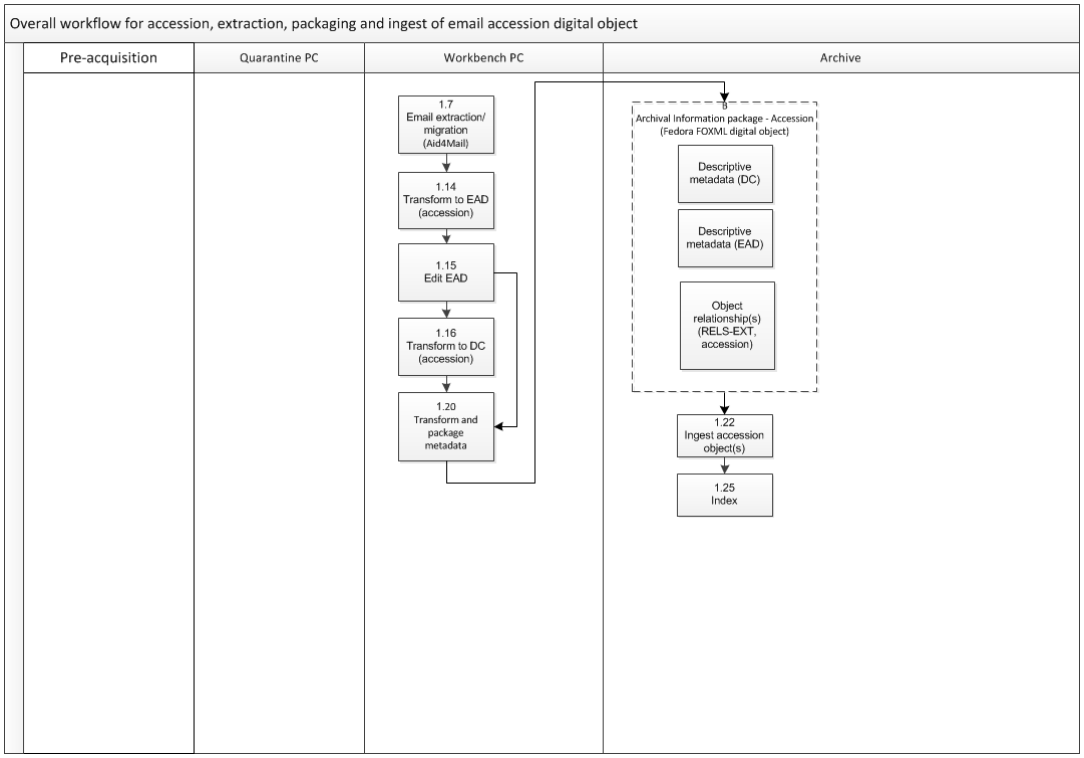


Figure 3: Collection

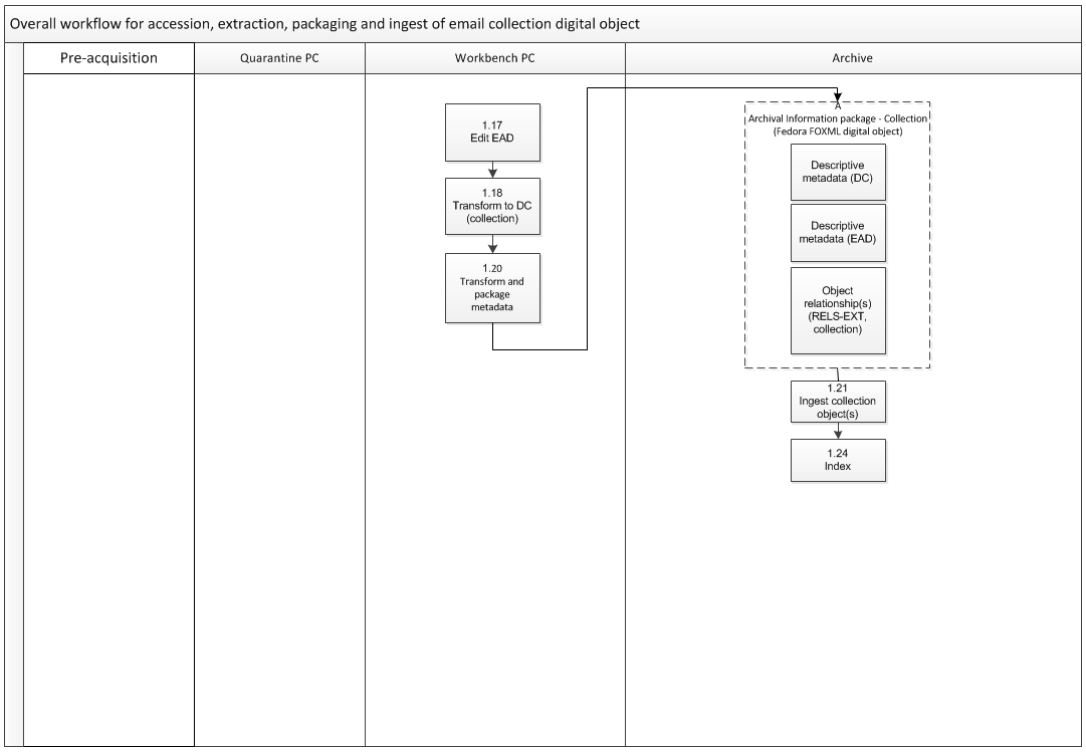
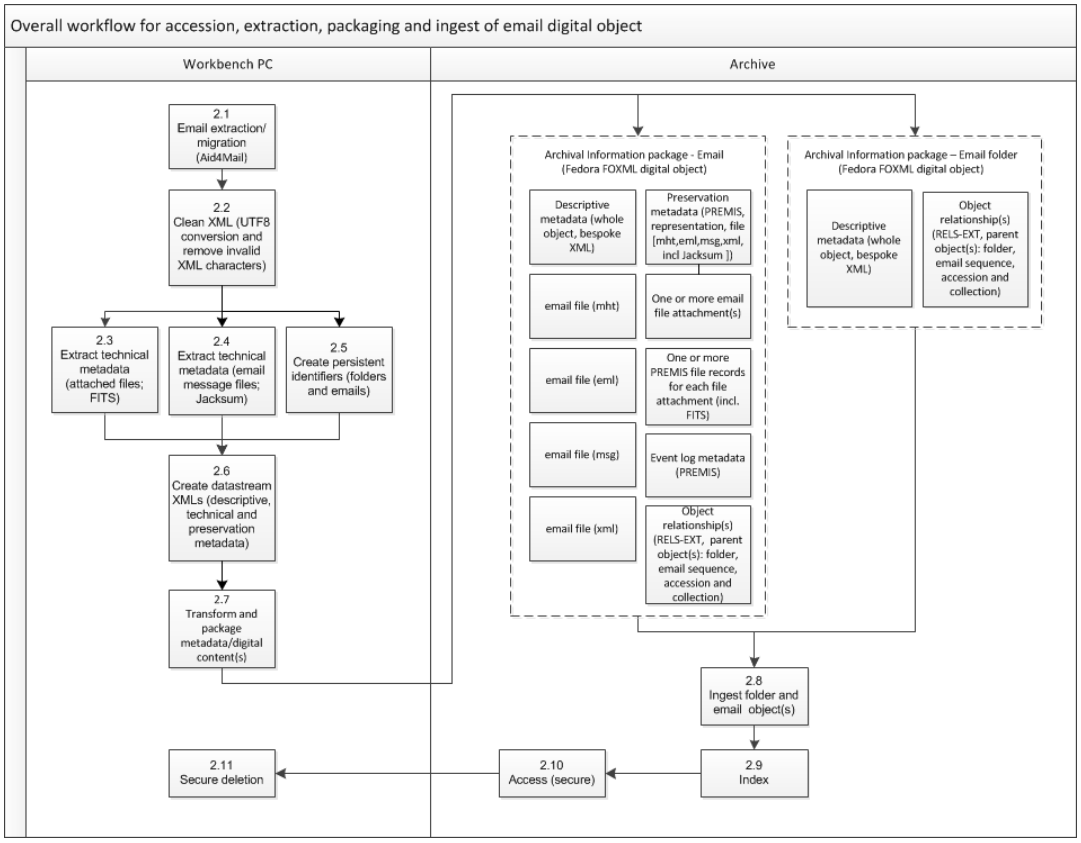


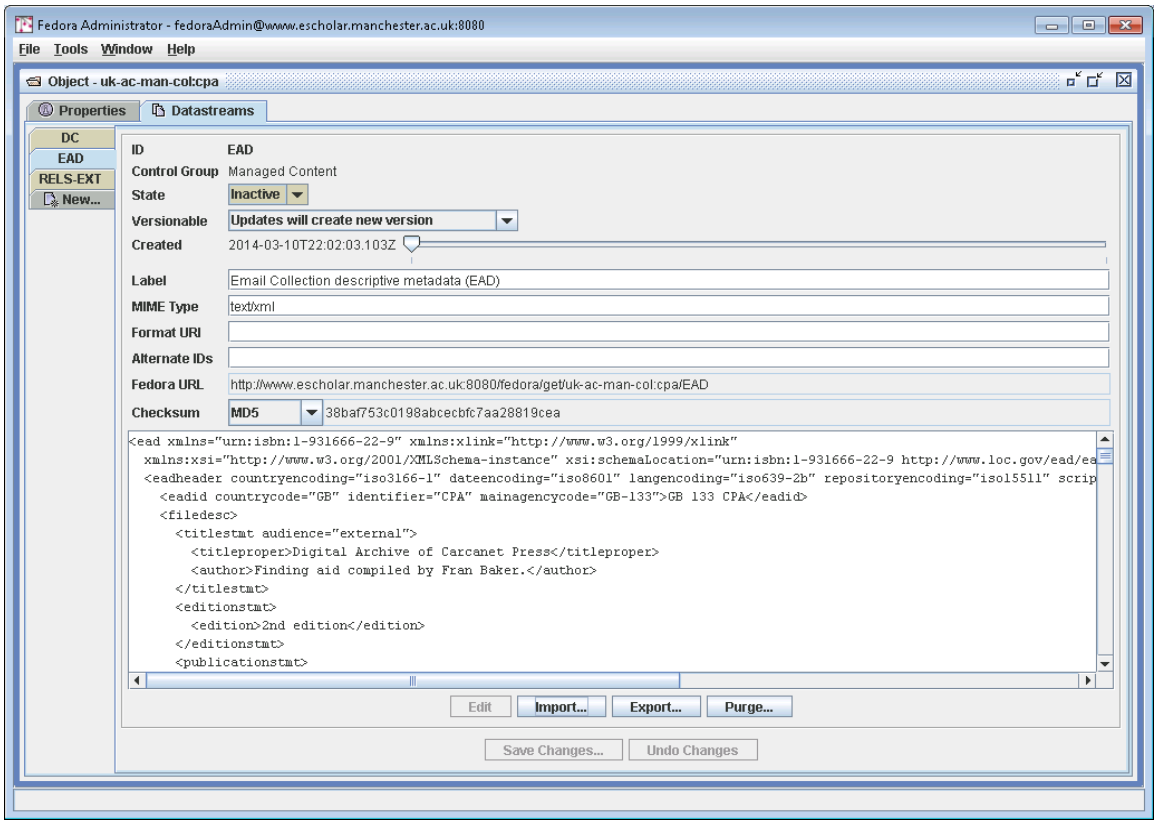
Figure 4: Email folder and email message



Appendix 2: Object profiles

This appendix contains screenshots of each type of digital object after ingest, showing different datastreams.

Figure 1: Collection object



The various datastreams are shown in tabs on the left, and the EAD record is open in the viewing pane.

Figure 2: Accession object

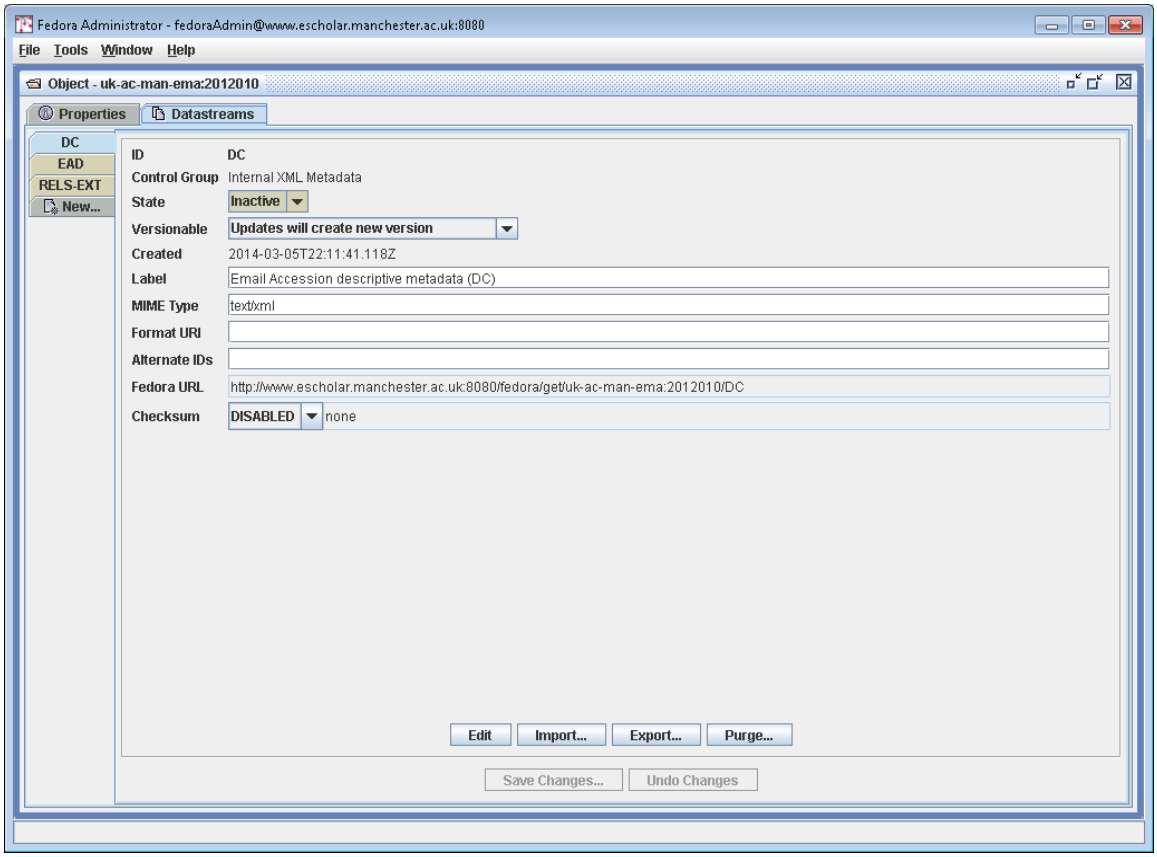
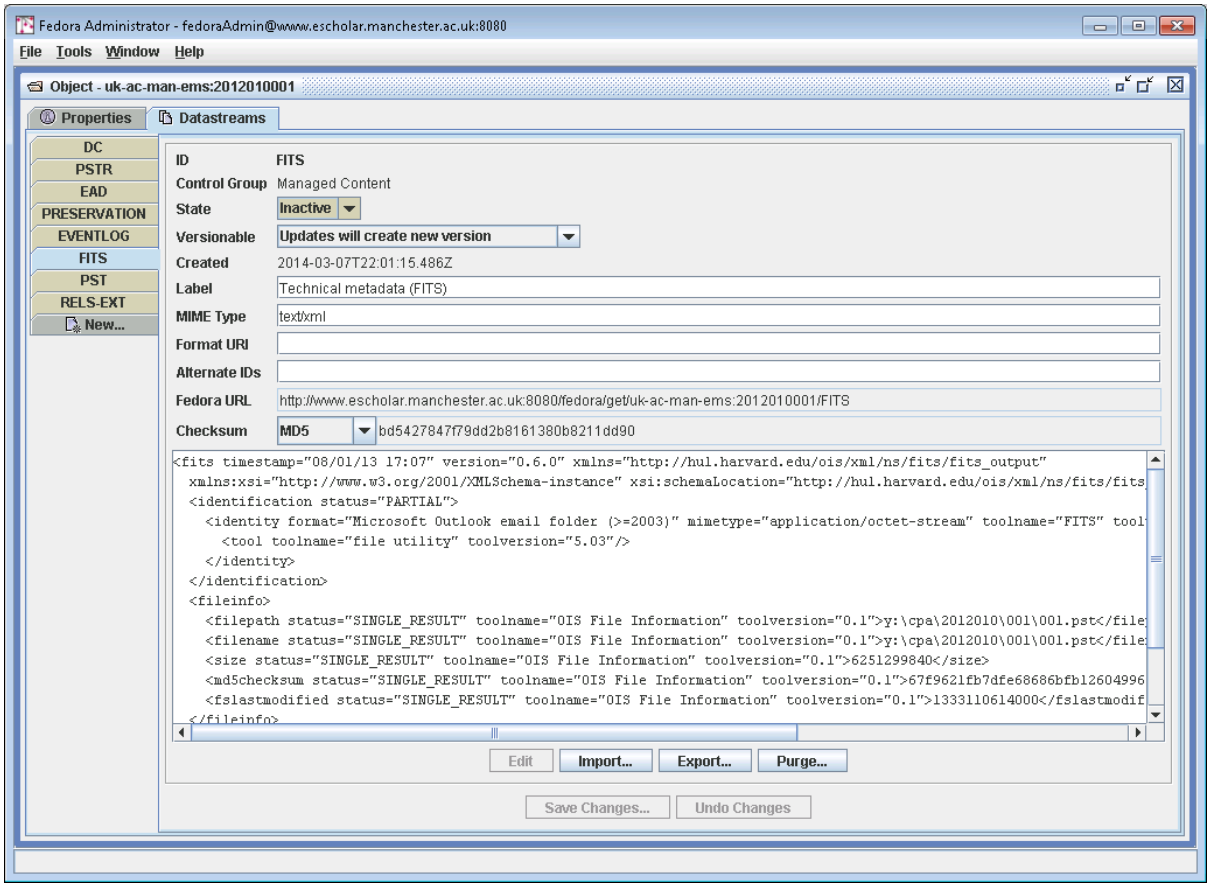
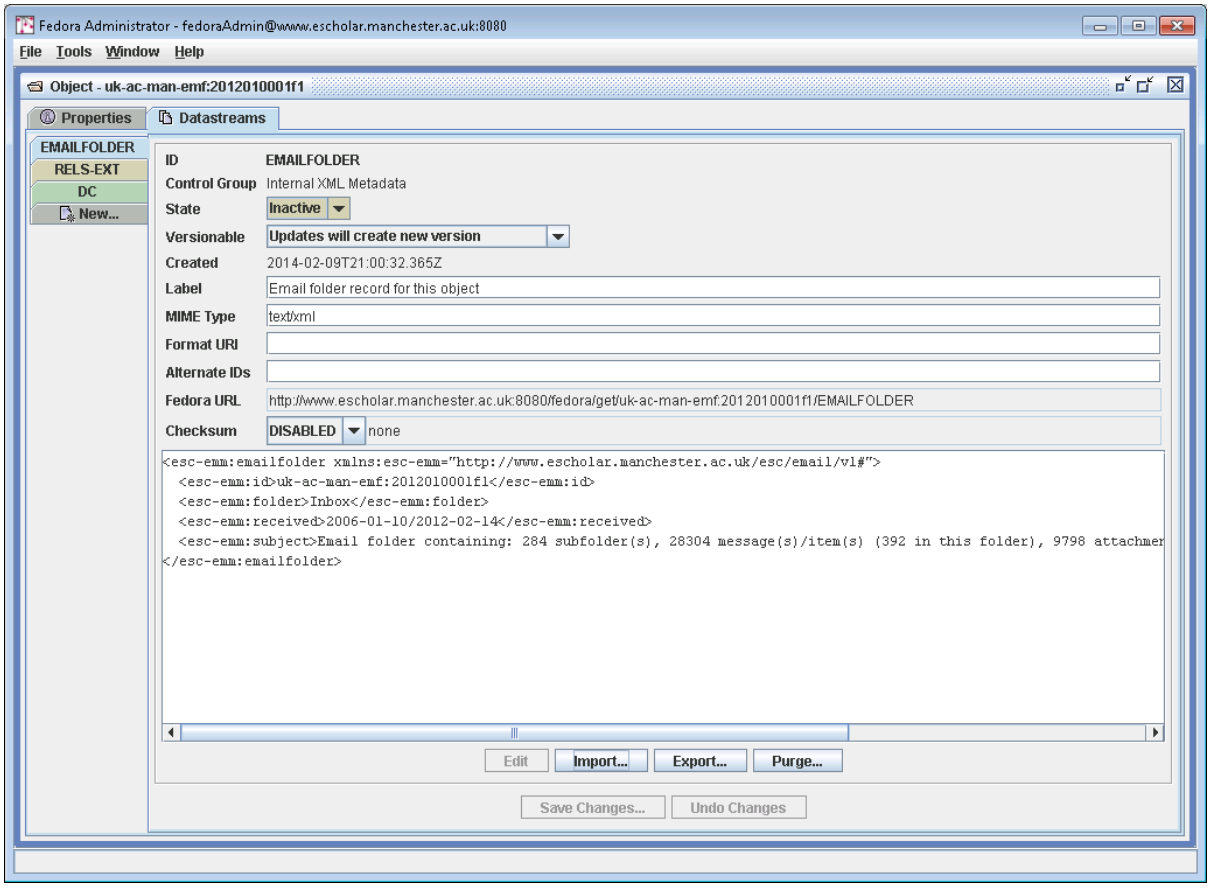


Figure 3: Sequence object



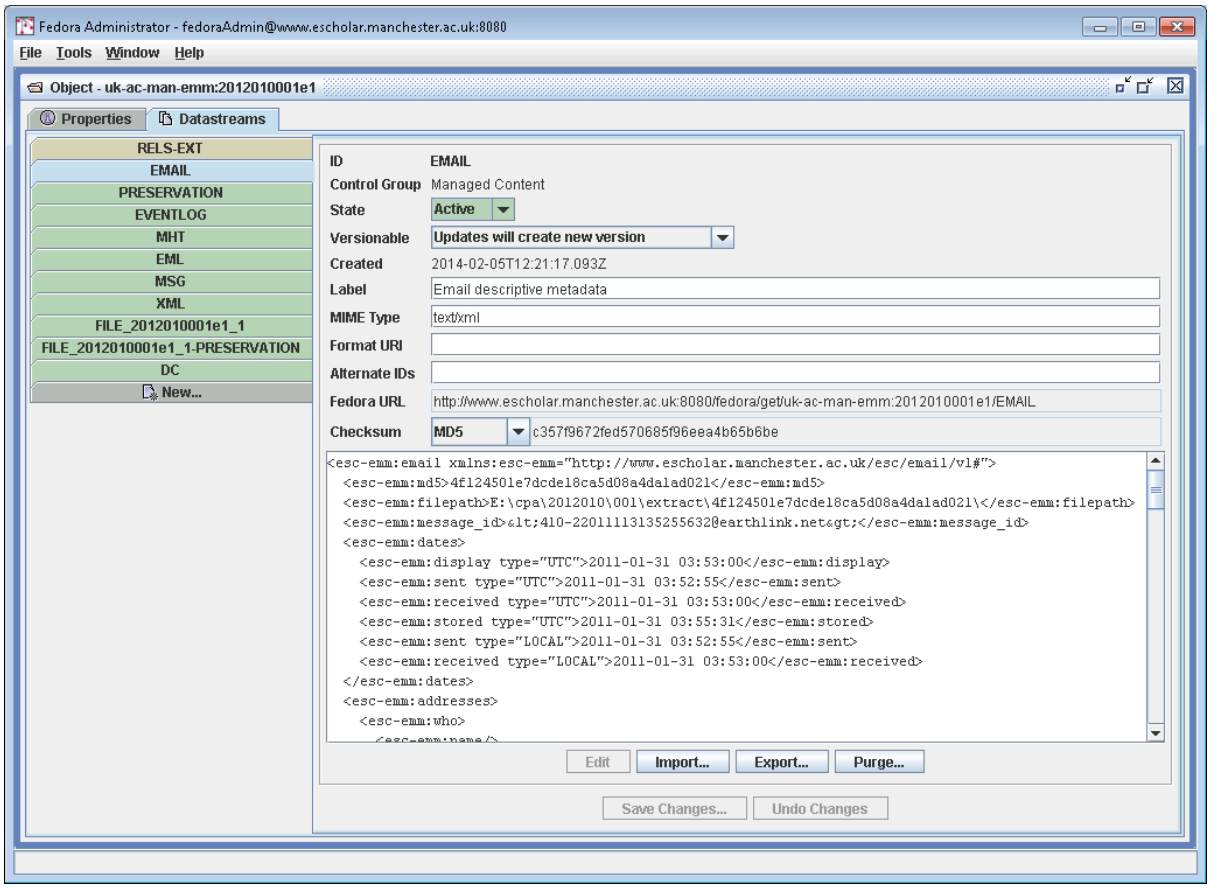
This shows the FITS record open in the viewing pane.

Figure 4: Folder object



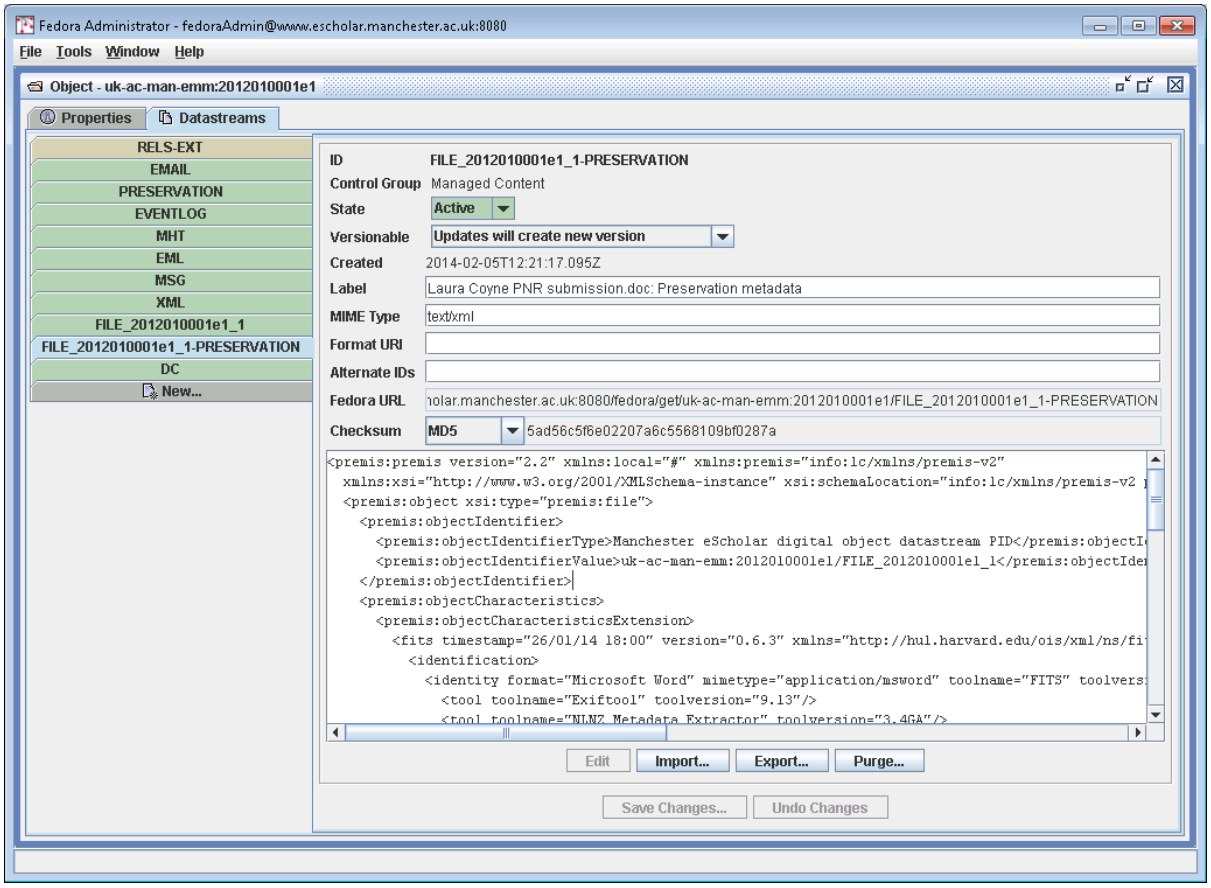
Example of a folder object, which shows the bespoke descriptive metadata open in the display pane.

Figure 5: Email object (1)



Example of an email object with an attachment, which shows the bespoke descriptive metadata open in the display pane.

Figure 6: Email object (2)



The same email object, in this case displaying the PREMIS record for the file which is attached to the message.