JISC Final Report

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	Carcanet Press Email Preservation Project		
Project Hashtag			
Start Date	12 March 2012	**End Date**	1 May 2012
Lead Institution	The University of Manchester Library		
Project Director	Sandra Bracegirdle		
Project Manager	Fran Baker		
Contact email	fran.baker@manchester.ac.uk		
Partner Institutions			
Project Web URL	To follow		
Programme Name	Digital Preservation Strand – 12/11 Digital Infrastructure Programme		
Programme Manager	Neil Grindley		

Document Information			
Author(s)	Fran Baker; Phil Butler; Ben Green		
Project Role(s)	Project team		
Date	4 May 2012	**Filename**	Carcanet_Final_Report_JISC
URL			
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
01	4 May 2012	Final version

Table of Contents

1 Acknowledgements

The project was funded by JISC under the Digital Preservation strand – 12/11 Digital Infrastructure Programme. The project team would like to thank the following individuals and institutions for their involvement and/or advice during the planning and implementation of the project:

- The staff at Carcanet Press.
- Simon Wilson and the AIMS Project team at the University of Hull.
- Susan Thomas at the Bodleian Library, University of Oxford.
- Fookes Software/Aid4Mail helpdesk.
- James Peters, Stella Halkyard, Gareth Lloyd and Graham Johnson at the University of Manchester.

2 Project Summary

Collecting institutions are increasingly facing the challenge of preserving 'born digital' material when acquiring recent and contemporary archives. Interim solutions like printing important documents and correspondence to paper are clearly not feasible in the long-term.

Among the most important modern archives held by the University of Manchester Library (UML) is that of Carcanet Press, one of the UK's premier poetry publishing houses. Correspondence with famous poets, critics, editors, translators and artists forms one of the most important elements of this archive. Most of this correspondence is now conducted by email, with the result that the quantity of hard copy correspondence acquired in annual accruals to the archive has diminished significantly. It is therefore vital that libraries like the UML are able to preserve these emails in digital form. This will ensure that invaluable primary research material is not lost to the archival record.

Our project tackled the challenge of capturing and preserving the email archive of Carcanet Press. Basing our work on both traditional archival practice and digital preservation standards, we established workflows, mechanisms and documentation for acquiring, migrating and preserving this sizeable email archive, which (after appraisal) extends to 16 GB and approximately 170,000 emails.

Some further work on verifying the results of migration experiments remains to be done, but we have developed a data model and full metadata profile for the body of email material so it can be straightforwardly ingested into our institutional repository, Manchester eScholar, which is based on the Fedora Commons software.

Due to data protection, sensitivity and copyright issues, our focus for the project was on preservation rather than access. However, we also took into account and explored some of the ways in which researchers might access and use such a body of material in the future.

The project has enabled UML staff to take their first steps in practical digital preservation, and we now feel much better equipped to deal with similar large-scale email archives in future.

3 Main Body of Report

3.1 Project Outputs and Outcomes

Output / Outcome Type *(e.g. report, publication, software, knowledge built)*	**Brief Description and URLs (where applicable)**
New curatorial documentation	Essential documentation for acquiring and managing digital archives, including: a records survey questionnaire focusing on email; deposit agreement; transfer list template; processing plan template; and appraisal form.
Substantial digital archive	Representing the Library's first digital archival deposit, the email archive is 16 GB in size and is now almost at a stage where it can be ingested into eScholar, the University's digital repository.
Hardware and storage	Dedicated digital preservation hardware and a secure network drive for initial processing of digital archives.
Event-logging database	Used to create a full audit trail of work carried out on digital accessions, this will be converted to XML for ingest into eScholar.
Catalogue and cataloguing templates	A comprehensive set of EAD finding aids for the email archive, and EAD templates for reuse with other digital material. The EAD catalogues will be stored in eScholar as descriptive metadata for the archive, but have the potential in future to be made available more widely on ELGAR, the Library's archive catalogue database.
Full metadata profile	A full metadata profile for the email archive has been produced, including preservation metadata held in the PREMIS schemas.
Data model for Archival Information Packages	A model has been developed for creating eScholar objects (AIPs) at accession, mailbox, and individual email level, incorporating tool outputs, descriptive and preservation metadata; workflows have been established for populating the preservation and descriptive metadata schemas.
Software code for metadata extraction and for automatic verification of migration experiments	The project's software developer has written code and integrated a set of existing tools to process email archives in preparation for ingest – notably for extracting descriptive metadata about PST files; and for verifying migration experiments against the original PST files.
A set of email test data for verifying migration experiments	A set of emails representing technical and curatorial significant properties for more granular verification of email migration

	experiments.
Increased knowledge and experience of digital preservation issues among Library staff	This represented our first experience of putting digital preservation principles into practice.

3.2 How did you go about achieving your outputs / outcomes?

3.2.1 Project background

One of the most important archives held at the UML is that of Carcanet Press. A recent comprehensive collections analysis exercise ranked this archive tenth, in terms of its significance, out of a total 316 archive and manuscript collections surveyed.

From modest beginnings in the 1960s, Carcanet Press has grown to become one of the UK's premier poetry publishing houses, its list including some of the most well-known twentieth and twenty-first century poets writing in English. It is also recognised for publishing editions of work by earlier poets, poetry in translation, and to a lesser extent fiction and lives and letters. The press also produces the bi-monthly journal *PN Review*, which has been described as 'the premier British poetry journal'.

The UML purchases the Archive of Carcanet Press, taking in new accessions of material on an annual basis. The paper element of the archive now extends to over 1,300 boxes.

While Carcanet Press still prints off authorial manuscripts and does much of its proofing in hard copy, increasingly correspondence is carried out by email rather than letter. Current policy is to print off the most significant emails for the correspondence files which come to the UML, but most email correspondence remains in digital form. This means that much of the correspondence from recent years – and therefore much invaluable primary research material – is being lost to the archival record. The UML realised that this was a matter which needed to be addressed as an urgent priority, and made an undertaking to acquire the email archive of Carcanet Press.

This was the first practical digital preservation project undertaken by the UML, and aimed to tackle one of the most complex record types and formats in preservation terms. Like most of the archives taken in by HE special collections libraries, this is material over which the UML has no control during its active life, and it has not been subject to any kind of electronic records management programme.

The project's overarching objectives were:

- To acquire and preserve the Carcanet Press Email Archive, for a one-year period in the first instance, and subsequently for the long-term.
- To use this Archive as a test-bed to establish good practice, within institutional resources, for acquiring and preserving similar born-digital archive material in the future.
- To document every stage in this process and produce good practice guidelines and training materials in order to ensure knowledge and expertise are retained after the conclusion of the project.

- To share our experiences with the archival and digital preservation communities, and disseminate our findings.

Other objectives included:

- To complete a records survey in order to establish the scope and technical requirements of the project.
- To transfer the records to the UML's custody securely and safely, in such a way that their authenticity could be assured.
- To stabilize the transferred records, ensuring their integrity is maintained and there are no threats to preservation.
- To establish full intellectual and administrative control over the records.
- To explore issues relating to access and significant properties (including a curatorial profile for the material) which would inform preservation planning.
- To create metadata profiles for the body of emails, with the aim of extracting as much metadata as possible by automated means.
- To create a preservation version of the preserved emails along with all relevant metadata to form Archival Information Packages as defined by the Reference Model for an Open Archival Information System (OAIS), for ingest into the UML's institutional repository, Manchester eScholar (based on Fedora repository software).
- To identify appropriate preservation formats for the body of material, taking into account institutional resources and future researcher access requirements, and to test some existing migration tools.
- To establish an ongoing preservation plan and techwatch programme for the material.
- To produce documentation and training materials to ensure knowledge and expertise are retained after the conclusion of the project.
- To share experiences with the archival and digital preservation communities.

3.2.2 Project methodology

UML received £11,000 from JISC for the project, which funded the appointment of a software developer from an agency to work full-time on the project for seven weeks; and the purchase of two inexpensive software products to facilitate preservation work. Some preparatory work was undertaken by UML staff in advance of the formal project start date.

The Project team consisted of:

- Sandra Bracegirdle, Head of Collection Management (Project Director)
- Fran Baker, Assistant Archivist (Project Manager)
- Dr Phil Butler, eScholarship Manager
- Ben Green, Digitisation Infrastructure Manager
- Howard Smith, temporary Software Developer

As the UML's first hands-on digital preservation project, the approach taken by the project team was very practical in nature and the methodology consisted of:

- A small project team drawn from different departments within the UML (notably Special Collections, Digital and Technical Services and Collection Management) in recognition of the fact that digital preservation skill sets are likely to be dispersed across departments.

- An incremental and iterative approach, including frequent project meetings at which progress was assessed and actions modified where necessary.
- Two reports to UML's Digital Preservation Steering Group, under whose auspices the project was run.
- Developing digital preservation skills by using a body of 'real' digital archive material as a testbed, although all workflows and processes were first piloted on a set of carefully constructed dummy data reflecting the key characteristics of the digital archive.
- Applying and adapting traditional archival principles to deal with archives in new forms, whilst also applying standards specific to digital preservation like the OAIS Model.
- Developing new curatorial and technical documentation and workflows which can be re-used for processing similar bodies of digital archive material in future.
- Testing existing software tools to see how these could be integrated to create a preservation workflow.

3.2.3 Project overview:

3.2.3.1 Records survey and recordkeeping practice

Before the project began, a survey questionnaire was sent out to the two senior Carcanet staff whose email formed the focus of the project, with the aim of gathering data on:

- key series of email correspondence which are most likely to have long-term value for researchers;
- record-keeping behaviours of the two participants;
- sensitivity and confidentiality issues;
- the extent of the material in question;
- some of the technical practicalities involved.

A more generic version of this tailored questionnaire will be produced for reuse by other curators involved in assessing email archives in future. The results revealed numerous interesting and/or challenging issues around recordkeeping, including:

- The bulk of the Press's author correspondence is now carried out by email, including key exchanges like formal acceptance letters, correspondence relating to the publishing project, author questionnaires (sent as attachments), and correspondence relating to proofs, text and production.
- The two Carcanet staff surveyed had dramatically contrasting recordkeeping habits: one email directory was highly structured, with unfiled messages being deleted after two months, and mail relating to all finished projects being archived to a single PST file; the other contained a filing structure for incoming mail, but only a single, vast, 'Sent Items' folder, with ad hoc 'archived' files, much non-Carcanet-related correspondence, and only very trivial material being deleted.
- Lack of understanding about Microsoft Outlook's 'archiving' facilities and the location of archived files – which in most cases resided on the creator's PC.
- The extent of duplication: multiple staff members were frequently copied into long correspondence threads; attachments were both downloaded, saved elsewhere, printed off, and retained with their covering messages.

- The extent of the material involved: it was clear from the survey that there were hundreds of email folders and an initial visit confirmed that the material came to over 10 GB in size.
- Significant problems with SPAM and junk mail; as well as vast quantities of junk mail received, hundreds of genuine messages received from bona fide correspondents were identified as SPAM in the subject line.
- Use of idiosyncratic folder naming (involving combinations of letters and numbers) to prioritise and sort folders by staff.

3.2.3.2 Data transfer and security

The email archive forming the focus of the project dates from 2001 to the present and contains both personal and sensitive personal data as defined by the Data Protection Act. For both DPA and copyright reasons, it was stipulated in the deposit agreement that the email archive would be subject to embargo and would not be made accessible to any third parties other than the UML's project team. The focus of the project was therefore primarily on preservation rather than access.

The original plan was to target individual email folders for transfer, but this proved too complex, time-consuming and disruptive for Carcanet staff. Both staff members therefore agreed that the UML could harvest their entire email directories, with appraisal and deletion to take place at a later stage.

Both participants were using Microsoft Outlook 2007; the project team created two PST files (using the 'Export' function) which essentially formed live snapshots of the participants' email directories on the date of transfer, and then went about locating six further 'archived' PST files. These were not straightforward to locate and Carcanet staff were unclear about the circumstances of their creation; some of them were probably created using the 'Auto-Archive' function of MS Outlook. They were inconsistent in content: one turned out to be completely empty; they all had significant overlap in date ranges; one had no sent items, and one contained only sent items.

On locating them all, it also became clear that their total size was significantly larger than anticipated, and the initial transfer involved approximately 25 GB of material. Copying the files also took far longer than anticipated, and the process had to be left running overnight.

The email archive was transferred from Carcanet's staff PCs onto a portable storage device. The device used was an Iomega Prestige 1TB USB Hard Drive. This device connects to a PC without any need to install software or drivers. As it doesn't require an external power source, it is easily portable and is relatively unobtrusive. A software application called TrueCrypt was used to encrypt this portable hard drive. TrueCrypt can be run from a USB memory stick, therefore no software installation is required to decrypt the portable hard drive when it is connected to a PC. To verify the integrity of the data transferred to the portable hard drive, a fixity check was performed on each file using the software application 'Jacksum'. This application is also run from a USB memory stick. Each time Jacksum processed a file and generated a 'checksum output', this information was recorded in the event log. By grouping each checksum process output together in an event log, it is visually easy to compare the algorithm outputs and spot any differences. The Jacksum outputs also provided some useful metadata such as original file paths and filenames of archived PSTs where they sat on the creator's PC.

Transfer lists were also completed and signed by both parties; these captured some additional technical metadata for use at later stages of preservation.

Jacksum was also used to ensure the integrity of the PST files once transferred to the quarantine PC. The same fixity checks were subsequently performed following antivirus checks.

Two dedicated digital preservation workstations were set up – a quarantine machine for initial processing (virus checking and appraisal) which was isolated from the network, and a 'workbench' PC for subsequent processing. Both PCs were stored in a physically secure area of the Library. Immediately prior to the transfer of the email archive data, the quarantine PC was temporarily connected to the network so that the anti-virus software could obtain the latest updates. A secure access-restricted folder was created on a standard shared network drive for storing metadata, software tools and other documentation; and a secure folder on the Institution Storage Area Network of 100GB was obtained for the storage of the original and migrated archival material, with access limited to team members.

The project technician created an 'event log' for audit-trail data – in the form of a multi-dimensional spreadsheet, which allowed the team to create a full audit trail of every action undertaken on the archived files, and to trace each action to an agent. The fields for this were broadly based on the PREMIS Event entity, with a view to this data being extracted and used to populate PREMIS records after ingest into eScholar. An extensible list of anticipated events was created at the outset.

The digital archive material was also accessioned in the UML's accessions database; the material was treated as two separate digital accessions based on provenance (i.e. their individual creator within the Carcanet office).

3.2.3.3 Appraisal and disposal

The limited timescale of the project and the extent of email captured precluded carrying out detailed appraisal; this issue remains to be resolved, particularly in relation to the less structured PST files. It was recognized that tools like MUSE (which can scan mail directories for duplicate messages, and has the potential to be configured to search for sensitive content) may be useful for more granular appraisal at a later date, but for the purposes of the project, appraisal was limited to folder/sub-folder level.

Even this was hampered by the unavailability of a quick and reliable tool to provide directory and folder listings of PST file content. The team understands that some institutions are using Forensic Toolkit software for providing listings of email in other formats, and have identified this for further investigation. However, for this project, appraisal at folder level was undertaken by a laborious process of creating screenshots from a working copy of each PST file using a dummy Outlook account on the quarantine machine – involving the creation of over 60 individual shots. The archivist scanned the folder titles to make initial decisions about appraisal targets, and undertook sampling of messages in these and other ambiguously-titled folders. All suggested targets for appraisal were submitted to Carcanet staff and after discussion a list of folders to be deleted was compiled.

Following research into proprietary and free tools which claimed to facilitate the secure deletion of individual folders within a PST file, and following some testing on dummy data, a deletion *process* was instead decided upon as the most effective option. Folders targeted for

disposal were deleted from within Outlook on the quarantine PC; the recycle bin was emptied; and each PST file was then compacted. Once compacted, new fixity checks were run and the files moved to the Workbench PC – and subsequently to the secure network storage drive. The hard disk free space on the quarantine PC will be completely wiped using a secure deletion tool after the conclusion of all the project's work.

Further appraisal at a later date will be essential – notably where junk mail slipped through the initial net, and particularly in relation to the enormous 'Sent items' folders retained in several of the PSTs, which are likely to preserve large quantities of correspondence from other deleted folders in the form of message threads. Certain information about calendars and contacts was also inadvertently acquired when the PST files were exported, and it will also be necessary to discuss the retention of these with the depositors.

3.2.3.4 Archival arrangement

The hard copy archive of Carcanet Press is vast, and only a small proportion of it (dating from the late 1960s to c. 1980) has so far been subject to archival arrangement and cataloguing in Encoded Archival Description (EAD). Whilst an overall arrangement for the archive into subfonds has been established, the majority of the archive is simply described in basic box lists, its order reflecting the way in which it arrived at the Library.

Accessions of the hard copy archive are taken in regularly – currently on a sixth-monthly basis – and each of these represents a 'snapshot' of the company's records as they stood at the point of transfer. Correspondence sequences are arranged alphabetically by author name and cover a specific chronological period.

In this sense they are very similar to the email 'snapshots' forming the focus of this project, and despite the inconsistencies between the content and structure of each PST, it was decided that they should be treated in the same way as the hard copy correspondence and left in their original order. Imposing any kind of archival arrangement on the material would also be extremely labour intensive, and at levels lower than folder or sub-folder almost impossible for such a large body of material.

It was therefore decided to treat each of the six PST files as an 'email correspondence sequence', for lack of a better term. The term 'file' was deemed inappropriate, as the team aimed to break down the PSTs to individual email level as part of the preservation process – and recognized that the PST file is only one 'Representation' (in PREMIS terminology) of the email correspondence.

There was also so much chronological overlap between the PST files that the order in which they were numbered simply reflected the order in which they were exported from the creator's PCs rather than any kind of chronological ordering system.

3.2.3.5 Descriptive metadata

Although the email forming the focus of the project was to be embargoed, and full-scale cataloguing was beyond the project's scope, it was recognised as being crucial to create some level of descriptive metadata about the material:

- to enable the archivist to search and locate material for possible appraisal;

- to answer any queries from Carcanet Press staff: the email archive is due to be transferred to the UML's ownership in early 2013, and Carcanet staff hope to dispose of some of the archived files on their own PCs after that date; they have occasional need to consult material in the archive for ongoing business reasons;
- to lay the groundwork for future, more detailed, cataloguing, and facilitate researcher access;
- and in conformance with the OAIS information model, in which Information Packages are accompanied by Descriptive Information to support their discovery and retrieval.

The minimum level of preservation envisaged by the project was to ingest the PST files in their native format into the eScholar repository, and the next level involved similar mailbox-level preservation but in a neutral format like MBOX.

Viewing the content of the files would still require an email client of some kind, so in order to pinpoint relevant material within each file before viewing, it was considered important to extract and create some level of metadata.

Typically, the key types of enquiry for the hard copy Carcanet Archive are either:

- author-based: enquirer needs information about material relating to a specific Carcanet author – sometimes limited to a specific publication by that author.
- subject based (sometimes delimited by date range), e.g. women writers, poetry in translation from specific languages, literary movements.
- occasionally more specialised, e.g. editorial decisions and policies; cover artwork.

Currently, most of the archive is described in the form of very basic box lists which include information on:

- provenance/creator;
- covering dates for correspondence;
- correspondent names;
- book titles for which files are included;
- extent/size;
- access restrictions.

Subject-based enquiries rely on the expertise of the archivist to identify relevant material as the collection is not currently indexed by subject.

It was therefore decided to try and achieve a similar level of basic metadata for each sequence of email correspondence, but in a structured XML format (using EAD) which would facilitate ingest and interrogation in eScholar. The wish list included:

- overall covering dates for each accession;
- the digital extent of each accession, correspondence sequence/PST file, and folder/subfolder;
- the logical extent of each sequence/PST: number of folders, subfolders, messages and attachments;
- folder structure, including folder names and folder/subfolder hierarchies;
- covering dates for each folder and subfolder;
- number of messages and attachments in each folder/subfolder;
- a list of file formats (or at least file extensions) represented as attachments.

After some research, the team located a piece of free software called Kernel PST Reporter [http://www.nucleustechnologies.com/outlook-pst-reporter.html] which can generate exactly this kind of key information about Microsoft PST files, i.e.

- 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 sent.

This was a significant find, as exhaustive manual comparison between working copies of the PST files viewed in a dummy Outlook account and the PST Reporter output revealed the tool to be 100% accurate as far as it was possible to tell.

Some of its outputs (like the sender log) were recognized to be valuable but too extensive to import into EAD files, so it was decided to retain the full output of the tool for each file in HTML format and to ingest this as a datastream which would form part of the Archival Information Package for each correspondence sequence.

The only drawbacks of the tool were:

- its lack of dating information;
- and its failure to replicate the folder structure of a PST file: while it is evident from the recorded folder path where each folder sits in the structure, the folders are not listed in the order in which the creator kept them in their Outlook account.

The project technician therefore produced some code which would use the PST Reporter output to:

- populate the <scopecontent> element of an EAD record with a list of folders in their original hierarchical structure;
- record the number of messages and attachments in each folder, as well as their digital extent;
- provide a summary of file extensions represented as attachments;
- produce aggregated information for higher levels of description.

Date ranges were also imported into the EAD template using the output of MessageSave and Aid4Mail software (see p. 14 for details).

The descriptive output produced by this method was enhanced and augmented manually by the archivist to include: further detail about scope and content based on scanning a working copy of the material in Outlook; appraisal information; provenance; and access conditions.

The EAD records were treated in a somewhat unorthodox way in that separate records were created for each accession and each sequence of email correspondence; each EAD record would therefore form one of the datastreams within a specific digital object in the eScholar repository – and thus conform to the OAIS information model.

The PST Reporter software was retained for characterization purposes: as its output appeared to be reliable, it was used as a benchmarking tool for evaluating the success of subsequent migration and extraction processes.

3.2.3.6 Access

While the focus of the project was preservation rather than access, it was recognized that decisions taken at this early stage would potentially affect future methods of access to the material, and that the needs of the 'designated community' (in OAIS terminology, i.e. both curators and researchers) for the material should be taken into account.

The team envisaged two principal access scenarios once the material had been ingested into eScholar:

Preserving at email directory or mailbox level:

Each sequence of correspondence would be stored as a single file (initially in native PST form, but also in a neutral format such as MBOX), with some basic descriptive metadata to facilitate the location of relevant material (as above). In order to access the material directly the whole file would need to be viewed in an email client, and the search and discovery tools of the client used to pinpoint material.

Preserving at individual email level:

The high-level files would be migrated and broken down to individual message level – either in the neutral EML form, and/or migrated to XML, with each email (along with associated attachments and metadata) being stored as a separate digital object in eScholar.

This would mean that key header fields could be indexed so as to be searchable across records, and in the case of XML records, full-text searching could be facilitated.

Indexing of digital objects would be done using Apache SOLR. This enables simple and faceted search on specific email metadata fields (e.g. the recipient or sender) and a full-word search (e.g. on all words and phrases in the email body). Apache SOLR delivers search results as XML or JSON and these are then rendered to HTML in a web browser. Our search API is secured using the institution's central authentication system and repository authorisation functions.

It was with these access models in mind that the team addressed the challenge of migration to preservation-friendly formats.

3.2.3.7 Migration and preservation planning

In order to assess the effectiveness of any format migration, the team established a set of significant properties considered crucial to maintain at both top level, and individual message level. These included:

- High-level properties that could be measured by comparison against PST Reporter output, such as numbers of top and sub-level folders; number of messages; and number of attachments.
- A body of just under 100 individual messages, selected to represent: a range of message header properties based on the categories identified by the InSPECT Project [http://www.significantproperties.org.uk/email-testingreport.html]; and a curatorial profile of message body properties, including formatting and its significance for this particular body of email specific to a literary publishing archive. Other curators

in the UML's Archives and Manuscripts team were also consulted on this issue, and their responses fed into the selection process.

In order to generate covering dates for the descriptive metadata, the project software developer had already experimented with extracting emails from PST files using various tools – which were also identified as tools for migrating the material to neutral formats. These were the PeDALS Email Extractor [http://sourceforge.net/projects/pedalsemailextr/]; MessagSave from TechHit [http://www.techhit.com/messagesave/reg.html]; and Aid4Mail by Fookes Software [http://www.aid4mail.com/].

Based on the experience of other projects and accepted best practice, the project team identified several potential preservation formats as migration targets for the PST files: at mailbox level, MBOX or one of its variants; and at individual email level, EML and/or XML.

Aid4Mail offered MBOX as an output at mailbox level. All the PST files successfully converted to the single MBOX format, but this failed to preserve one of the key high-level significant properties identified, i.e. folder structure. The MBOXrd alternative retained folder structure, but does not output as a single file, and so for preservation at mailbox level, further compression would be necessary. Further work is still needed to assess the successful retention of other significant properties by the MBOXrd file format.

At individual email level, XML was the preferred format. However, initial experiences with PeDALS (which converts to XML) suggested that this would not scale to the large email collection forming the focus of the project. Correspondence with the PeDALS developers identified a data typing error in the PeDALS code that pertained to certain binary fields which can occur in PST files.

As an alternative to PeDALS, the team investigated Aid4Mail. This successfully produced XML-structured emails with full formatting, at scale. However, in two instances (the two largest PST files), Aid4Mail failed to produce any output. This is currently under further investigation, but we believe the XML output has significant potential and flexibility for both preservation and access.

For producing individual EML records, the team initially tested MessageSave. Verification against the output of PST Reporter revealed that a significant number of both folders and messages were not successfully migrated. We were unable to identify the root cause of this problem. Aid4Mail produced much more reliable results, with only very minor discrepancies in the number of individual emails migrated. Further work needs to be undertaken at a more granular level to determine whether the technical and curatorial significant properties have been retained in the EML records; detailed analysis of a small number of messages suggests some discrepancies in certain elements of formatting and date/time display.

Overall, none of the tools tested met all the project's requirements, although Aid4Mail appears to be the most promising. Further immediate work will focus on Aid4Mail's outputs in MBOXrd, EML and XML formats. The MHT format, which generates a viewable (browser) mail viewer for the emails without the need of a client is also a possible consideration for access purposes.

3.2.3.8 Ingest and archival storage

Manchester eScholar uses the Fedora Commons Repository software [http://www.fedora-commons.org/]. Fedora (Flexible Extensible Digital Object Repository Architecture) is a

flexible software architecture for storing, managing and accessing digital content in the form of digital objects. This software gave the Project significant flexibility when considering how to ingest and archive the Carcanet Press email collection.

Our investigations resulted in two data models for creating Archival Information Packages – one for preservation at the PST or mailbox level and one for preservation at individual email level. The following objects and datastreams made up these data models.

For archiving email collections at the PST/mailbox level the core data model consisted of one or more accession objects and one mailbox object per PST/mailbox file. For archiving at email level the core data model consisted of one or more accession objects, one 'email sequence' object (equating to the content of an original PST file) for each 'sequence' in the collection; one folder/subfolder object for each email folder/subfolder in the collection; and one email object for each email/attachment in the collection.

An accession object contains metadata relating to each archival accession. As an 'Intellectual Entity' (as per the PREMIS data model), this contains no direct digital archival content. Datastreams are:

- Dublin core (descriptive metadata used by Fedora for internal management)
- EAD (more detailed descriptive metadata about the accession)
- RELS-EXT metadata (used by Fedora to manage digital object relationships) pointing to the PST/mailbox objects making up the accession

A PST/mailbox object contains all the essential metadata (e.g. preservation and descriptive) pertaining to the PST file and its migrated MBOX equivalent, combined with actual .pst file and the .mbox file. Datastreams are:

- Dublin core
- RELS-EXT, pointing to the accession object and sibling objects
- PREMIS including Representation and File metadata for .pst and .mbox files (preservation metadata)
- PST Reporter output (descriptive metadata)
- File Information Tool Set (FITS) output (technical metadata)
- EAD (descriptive metadata)
- .pst file
- .mbox file

An email object contains metadata which is automatically extracted and files pertaining to an individual email. Datastreams are:

- Dublin core
- RELS-EXT, pointing to parent folder object
- FITS (including technical metadata for .eml and .xml email files and file attachments)
- .eml file
- .xml file
- none, one or more file attachment(s)

Further digital objects would be created for folders and subfolders as Intellectual Entities.

In practice, archiving at the PST/mailbox-level created large and complex objects, but only a relatively small number for ingest. In contrast, archiving at the email-level created small and simple objects but many 10,000's for ingest.

Figure 1 illustrates the overall workflow developed to process and preserve the Carcanet Press email archive at the higher, mailbox level.

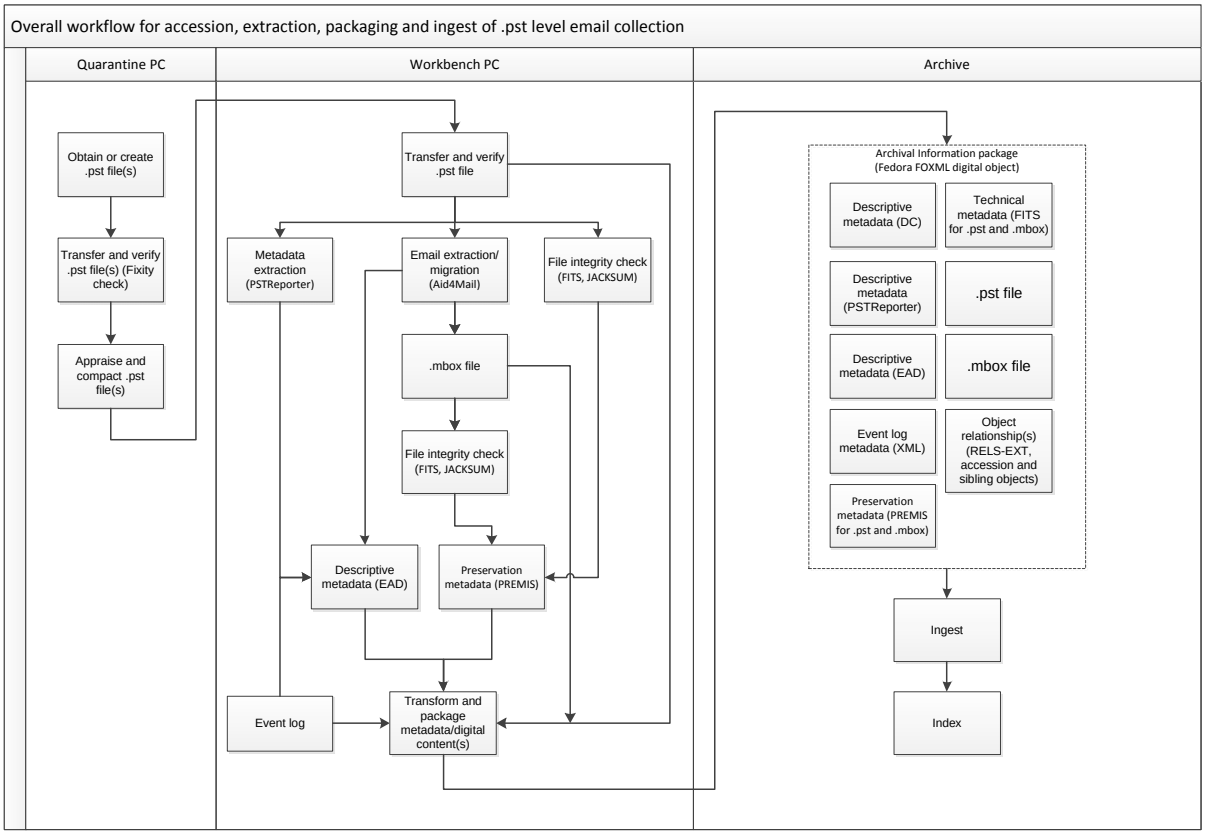

Figure 1

3.3 What did you learn?

The project enabled the UML to take a valuable first step in practical digital preservation; it required a concentrated effort over a short time period, which allowed the team to focus in a way which had not previously been possible. It proved to be a good team-building exercise which enabled an exchange of knowledge and expertise between staff working in different departments of the Library.

However, as the UML's first practical digital preservation project, one of the aims was to 'learn by doing', and the team encountered numerous challenges, problems and unanticipated outcomes. Not all of these could be overcome in the short timeframe of the project, and a considerable amount of future work has been identified.

Some of the issues and problems confronted were:

- The length of time needed to export such large PST files, a process it was necessary to leave running overnight – causing some inconvenience to staff at Carcanet and narrowly missing an automatic reboot which took place two hours after the files finished transferring. This will be taken into account when making arrangements to transfer digital archives in future.

- Similarly, the unforeseen length of time it took to carry out other procedures involving the larger PST files, including compaction, fixity checks and migration.

- Although it was chosen as a secure and sustainable working platform, difficulties were experienced as a result of using the University's 'managed desktop' system. Issues included: automatic reboots; Windows limitations on filepath lengths which compromised early migration and extraction experiments; problems with installing certain types of software because of conflicts with the University's preferred programs – in particular, this hampered the team's attempts to comprehensively virus check *within* a PST file; the institutional anti-virus software can only scan a folder at a time, which was painstakingly slow.

- Storage space: the 100 GB secure network ultimately did not provide enough space for carrying out processing and migration experiments at scale; work had to be divided across two network drives and the hard drive of the 'workbench' PC. Staff will take this space requirement into account when undertaking future digital preservation activities.

- It proved impossible to proceed through the project's workpackages in the chronologically structured way which was originally envisaged; certain elements of the work had to be undertaken earlier than expected and others spanned a number of workpackages.

- The lack of readily available tools which are able to work within a PST file, e.g. for generating folder structures and overall dates, and for securely deleting individual folders or messages.

- The difficulty of balancing extensive pre-acquisition assessment of the material with the disruption this caused to staff at Carcanet Press who needed ongoing access to their email. It is intended to create some record-keeping advice for the staff at Carcanet to address issues like: using the 'Archive' process rather than the 'Export' process to create PST files which should avoid so much overlapping content in future; and labelling archived folders with meaningful filenames (MS Outlook automatically assigns the title 'Archive' to archived PST files, and the project team acquired three PSTs with the same original name).

- The team's lack of specialised technical knowledge about email in different formats.

- The timescale of the project prevented the testing of a wider selection of tools for migration, metadata extraction and characterization; it also precluded the full development of customised tools to fill any gaps encountered. In retrospect, the team would have tested a wider number of tools on a body of test data at the outset of the project as some of the tools trialled had reliability issues when used at scale.

- The short timeframe of the project also meant that the work has not been completed to the degree the team had envisaged; in particular there was no time to carry out a detailed evaluation of the results of migration experiments, and it was decided to defer ingesting any material until this can be achieved.

- From its outset it was envisaged the project would require additional technical staff resource. Sourcing this proved more difficult than first imagined. Our ideal was to appoint an individual with a combination of technical and digital preservation or digital archiving experience. We approached a number of consultants with digital

preservation knowledge but unfortunately none were available to take on the work. We then sought staff from IT employment agencies. Again it proved difficult to find suitable candidates. In the end we compromised our requirements and appointed someone with a high degree of technical experience but little digital preservation knowledge. These difficulties resulted in some delays in initiating the project and also impacted on what was deliverable within the agreed timescales.

3.4 Immediate Impact

Although this was only a seven-week project, the following immediate significant impacts have been identified:

- For the staff at Carcanet Press, significant progress has been made towards preserving their digital archive, and they are likely to have increased confidence in the Library's ability to deal with further digital accessions in future.

- The project has ensured the 'rescue' for the archival record of a large body of research-rich material which was formerly at risk of loss; several of the files acquired resided only on the hard drives of Carcanet staff.

- The project marked the Library's first acquisition of a substantial 'born-digital' archive, thus enhancing its ability to continue collecting modern archives in the digital age.

- It has improved the confidence of Library staff in their ability to acquire and preserve born-digital archive material, marking a vital first step in practical digital preservation.

- The project has contributed to the advancement of strategic goals, both for the Special Collections Division and for the Library-wide Digital Preservation Steering Group.

3.5 Future Impact

The project has resulted in a large digital archive which, although currently embargoed, will form a key resource for future researchers – not just those working in the field of literary studies, but also for historians, sociologists and others.

The ability to deal with born-digital archives has the potential to enhance the UML's reputation as a collecting institution at an international level.

The University's central IT Services have expressed interest in preserving institutional emails, and may draw on the experiences and outcomes of the project.

The project will contribute to the growing dialogue about issues involved in the long-term preservation of email.

4 Conclusions

The project has succeeded in its objective to acquire and preserve the Carcanet Press Email Archive, and envisages that ownership of the archive will be transferred to the UML following the initial one-year deposit period.

The archive has provided an invaluable test-bed for establishing good practice in acquiring and managing this type of born-digital archive material. Although we have not yet achieved all our aspirations, we feel the project has enabled us to make significant progress in this area.

The project benefited from in-house technical and archival experience. In particular, our existing experience with the Fedora Commons repository system greatly facilitated technical development.

The project has created all the necessary components to produce in-house guidelines and training materials, e.g. the project's software developer has produced comprehensive documentation on the technical workflows he developed. The development of support materials will continue beyond the formal end of the project.

We also intend to disseminate the project's experiences and findings within the archival and digital preservation communities. The project manager has already been invited to speak at two events – for the Archives and Records Association, and for GLAM (the Group for Literary Archives and Manuscripts).

5 Recommendations

This project recommends:

- That JISC makes available funds for further practical projects to address the issue of preserving email archives – in particular to work towards standardisation of approaches in this area.

- That further work is done on testing suitable characterization software tools; we support Chris Prom's recommendation [http://www.dpconline.org/component/docman/doc_download/739-dpctw11-01pdf, p. 31] that further characterization tools are tested and developed in order to complete audit verification.

- PST Reporter as a reliable characterization tool for PST files.

- Aid4Mail as a promising tool for migrating email to numerous preservation formats, and for producing highly structured metadata. However, our project has highlighted potential problems with its effectiveness when working on very large PST files.

- That further, more detailed, work is undertaken on assessing the effectiveness of format migration tools in preserving the significant properties of email, as outlined in the InSPECT Project.

- That further work is done on creating curatorial profiles of significant properties based around bodies of real-world email archives from different sectors/subject areas.

- That the community considers creating large test datasets of non-sensitive email in different formats and with different properties, in order to facilitate further work on characterization and verification.

6 Implications for the future

From the outset the project took into consideration the sustainability of the Carcanet Press Email Archive.

The Library's Digital Preservation Steering Group (DPSG) will henceforth oversee the ongoing preservation of this collection.

The project chose existing repository infrastructure to host the email archive. This infrastructure is embedded within the University's IT strategy. The archive will be stored on the University's Storage Area Network which is dual sited and enterprise level.

The archive material is managed by the open-source Fedora Commons repository software, which has a well-developed support community. The Manchester eScholar support team has significant technical experience in working with this software and is an integral part of the Library's future IT strategy.

The project has produced a workflow that builds on existing software tools and a data model for the effective ingest and storage of archived emails in a Fedora-based repository. This may be of value to the wider digital preservation community.

Following the formal completion of the project, the team has identified the following outstanding areas of work which the Library will take forward:

- Verifying that migration to different formats has been successful in preserving specified significant properties, prior to ingest.
- Investigation of other tools for migrating email to preservation formats – including CERP Email Parser, ReadPST, Emailchemy, and Xena.
- Implement a search and discovery platform for email messages stored in a Fedora-based repository.
- Produce preservation plans for all formats represented in the email archive – including attachments.
- Complete work on a project website.
- Further dissemination of the project's work – both in-house and externally.
- Determine how future accruals to the Carcanet Press digital archive will be managed, and provide some record-keeping advice for the staff at Carcanet.

7 References

- Christopher J. Prom, *Preserving Email*. DPC Technology Watch Report 11-01 December 2011 [http://www.dpconline.org/component/docman/doc_download/739-dpctw11-01pdf]

- Dublin Core [http://dublincore.org/]

- Encoded Archival Description [http://www.loc.gov/ead/]

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

- ISO OAIS Reference Model, ISO 14721:2003 [http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683]

- PREMIS: Preservation Metadata Maintenance Activity, Library of Congress [http://www.loc.gov/standards/premis/].