A Method for Acquisition and Preservation of Emails

Claus Jensen
The Royal Library
Søren Kierkegaards Plads 1
Copenhagen, Denmark
+45 91324448
cjen@kb.dk

Christen Hedegaard
The Royal Library
Søren Kierkegaards Plads 1
Copenhagen, Denmark
+45 91324525
chh@kb.dk

ABSTRACT
In this paper we describe new methods for the acquisition of emails from a broad range of people and organisations not directly connected with the organization responsible for the acquisition. Existing methods for acquisition of emails are based either on having easy access to an institution's email server, or a labour intensive process of transferring the emails from donors' email clients, involving for example visits to the individual donors.

Furthermore, we describe how different representations of the acquisitioned emails are ingested into our repository. The use of different representations makes it possible for us to perform a fine grained file characterisation, thereby facilitating the level of preservation watch we want in connection with the preservation of the acquisitioned emails.

Keywords
Email; Acquisition; Preservation; Repository: Linked data.

1. INTRODUCTION
The email project began with a request from the curators in The Royal Library Manuscript Collections to acquire emails from individuals in arts and sciences (scholars, authors, etc.). The request was based on the assumption that much of today's written memory is born digital and that there is a high risk of losing this material if the acquisition process does not start until after the donor has passed away. Some of the major threats to the material are deletion due to computer crash, and limited space on email servers.

The primary audience for the final service is Danish archive creators such as authors, researchers, artists, and persons and organizations active in the cultural domain taken in a broad sense. Their digital material is considered important for future research in particular in areas of science, the history of learning, cultural history, etc.

The curators in The Royal Library Manuscript Collections have analyzed their audience to fall within three major groups: The first group are employees in institutions which almost exclusively use their institutions' email system. The second group are also employees of institutions, but this group mostly use their own private email. The third group is not affiliated with an institution, and therefore only use their private email.

As most of the target group was in the latter two groups, it was not possible to use the acquisition method where access goes through an institution's email system. The method of acquiring email from the individual donors' email clients was considered far from optimal both from a labour resource perspective and from a technical perspective.

2. STATE OF THE ART
A survey of methods for acquisition and preservation of emails can be found in the DPC Technology Watch Report Preserving Email [7]. A series of articles concerning the acquisition and preservation of emails has been written [1], [2], [3], [4], [9], [10], [11], [12]. The articles do not always describe the exact method of acquisition, i.e. how the emails are transferred from the donors to the institution responsible for the acquisition. However, even when the method is not explicitly described, it is often possible implicitly to see what methods have been used. The two most widely used methods of acquisition are: To extract the emails from email servers from which the institution has easy access or to use a more labour intensive process involving acquisition of emails through the donors' email client.

Different methods on how to pre-process and ingest emails into a repository have been studied in a number of articles. In E-mails to an Editor [1] it is described how the project ingest emails into a repository in three different formats MSG, EML, and XML and the Aid4Mail program [6] is used for the pre-processing of the emails. In Reshaping the repository [2] the process of how the project converts emails into the RFC-282 Internet Message Format [8] using the Emailchemy program [13] is described. In Coming to TERM [3] it is described how emails are converted to the RFC-282 Internet Message Format, if the original format is a proprietary format. The email and its attachments are marked up in XML before they are ingested into the repository.

3. THE INITIAL REQUIREMENTS
The process of finding or building a system for the acquisition of emails was initiated by a phase of collecting requirements with input from both the curators and us. The curators had a series of mostly non-technical requirements for the new email acquisition system.

Table 1. Non-technical requirements

Maximum emulation of the traditional paper-based archiving criteria and procedures
High level of security against loss, degradation, falsification, and unauthorized access
A library record should exist, even if documents are not publicly available
Simple procedure for giving access to third-party by donor
Maximum degree of auto-archiving
Minimum degree of curator interference / involvement after agreement

Similarly, we had a number of technical requirements for the system.

Table 2. Technical-oriented requirements

No new software programs for the donor to learn
No installation of software on the donor's machine, and if programs had to be installed, it should be standard programs and not programs we would have to maintain

As much control over the complete system in our hands as possible
As much as possible of the workflows within the system should be automated
Independence from security restrictions on the donor system imposed by others (password secrecy, restrictions on installation of programs, etc.)

4. THE FIRST PROTOTYPE

The first prototype was implemented on The Royal Library email system for a limited number of donors, selected by the curators. Each donor was given an "archiving email account".

We allowed the donors to choose between different methods for archiving emails. One of the methods was adding their archiving account as a BCC recipient when sending or responding to an email. Another method was to forward received or sent emails to the archiving account. The use of forwarding would for example be necessary when donating the last received email in a thread.

The donors chose to employ two different processes: One group of donors donated their emails using a continuous process of sending and receiving emails by using BCC and forwarding. The other group used a periodic donation process. An example of the use of the periodic process was when donors donated on a monthly basis by forwarding the emails to their archiving account.

A major disadvantage of the forward method for archiving emails is that important information contained in the original email header is either lost or hidden inside unstructured email message text. For the curators the original date of the email was important.

In some cases it would be possible to extract the send date of the email from the email message, as a number of email clients use a semi-structured way of registering this information within the email message. However, the email clients used different methods to separate the send-date information from the rest of the email message. Therefore it was not possible to implement a general method to extract the original send-date information.

Other disadvantages of using the forward method for archiving emails that we encountered were:

- It was easy for the donor to forget to forward the last message in an email thread
- Periodical donation sometimes failed because the email "package" was too big due to the following reasons:
    - A timeout from the antivirus scanner because the scanning time of the email exceeded the maximum time period allowed
    - The email provider had a size limit on the emails

We had to conclude that the first prototype had some serious drawbacks. Thus we had to look for other solutions for the acquisition of the donors' emails.

5. THE SECOND PROTOTYPE

Using our experiences from the first prototype and combining them with new ideas for the acquisition process, a new series of requirements took form in the beginning of the second phase of the project. In formulating the new requirements, we drew on both the donor's and our own experiences with the first prototype.

The additional requirements were formulated in the following way.

Table 3. New requirements for the second prototype

The system should be based on standard email components
Easy to use for both curator and donors
No curators should have to visit the donors' residence for setup or email transfer (self-deposit)
The system should be based on voluntary/transparent deposit
It should be independent of technical platforms (PC, Mac, iOS and Android devices, etc.)
The donor should have the option to transfer emails to the deposit area at any time
The donor should always have access to their donated emails
Based on permission granted by the donor different levels of access for external use should be allowed at any time.
The donors must be able to organize and reorganize emails.
The donors must be allowed to delete emails in the system within a certain time-frame
The original email header metadata must be preserved
The donors must be able to deposit other digital materials along with their emails

During the new requirement process it became increasingly clear that it was necessary to create two areas for each donor. We named these areas, respectively, the *deposit area* and the *donation area*. The deposit area was defined as a temporary dynamic email area where the donor (also called the "archive creator") could transfer all of their emails from their different email accounts. Furthermore, the archive creator still has all rights to the materials in their deposit area and is able to edit the deposited emails (create new emails and folders, move emails and folders, delete emails and folders, copy emails and folders, etc.).

The desired time period for the deposit of emails is specified in the agreement between the curator and the donor. Typically a three year deposit period is chosen. When the archive creator is ready to donate, the curator moves the agreed emails from the deposit area to the donation area. The emails then become the property of The Royal Library. The donor (previously archive creator) will now only have read access to their emails. After this the donated emails are ready for ingest into the repository system as part of the long-term preservation process.

The new requirements initiated a major redesign of the system. We decided to continue the principle that every donor should have their own email account. The open question on how to transfer the donors' emails to their archiving accounts without losing important information remained.

We investigated the possibility of using the email clients' ability to handle more than one email account at a time. This ability does not only mean that it is possible to read and write emails in connection with many email accounts, but also support the process of moving and copying emails and folders between different email accounts. The moving or copying of emails from one email account to another within the email client itself does a much better job of preserving the important information we lost in the first prototype.

To support as many email clients as possible we decided to use the IMAP (Internet Message Access Protocol) and SMTP (Simple Mail Transfer Protocol) between email clients and email servers. The IMAP protocol is implemented in all widely used email servers and email clients and it is platform independent. Furthermore, the IMAP protocol is both supported by the email clients of modern smart devices and by the many

free email clients for computers. Even though it is not possible to transfer emails directly from web-based email systems such as Gmail and Yahoo Mail, it is possible to transfer these emails using an email client supporting the IMAP protocol.

The process of moving and copying email and creating new folders within a single email account are well-known tasks for most donors. Therefore it was expected that these processes would be easy to perform for the donors even though they now had to perform these tasks between two email accounts instead of only a single account.

The second prototype allows the donor group that prefers a continuous donation process the ability to copy and paste (drag and drop) single emails to their archiving account immediately after they either send or receive new emails. The other group of donors who prefer using a more periodic donation process would have the ability to copy and paste multiple files or folders to their archiving account using larger time intervals.

Our second prototype was implemented as an independent email server, in our case an Exchange server [14], totally separated from The Royal Librarys' email system. Furthermore, the deposit area was separated from the donation area

The service options of the Exchange email server were limited as much as possible. Available service options were

- Full access via IMAP
- Webmail, but limited to read access and for changing the password for the email account.

The email accounts were set up so they could not receive emails. This was done to avoid unauthorized email messages like spam emails getting into the deposit area.

The new method of acquisition gave the donors the following benefits:

- They were able to use their own email client (Outlook, iOS mail, Thunderbird, etc.)
- They could deposit via different devices (Windows, Linux, iOS devices, Android devices, etc.)
- They could use several devices for depositing emails.

There were now only the following requirements for donors to deposit their emails:

- The donor must have access to an email client
- They must be able to setup an IMAP account in their email client on their own device.

The configuration of the IMAP and SMTP connections was, due to internal IT-polices at our institution, non-standard. The non-standard configuration resulted in the need to use a more complicated configuration for most of the used email clients. However, the latest developments in modern email clients has resulted in, that much of the complicated configuration can be done in an automated way, where only basic information like email address, user name, and email-server name need to be inserted by the user.

6. FROM DEPOSIT TO DONATION

At a given time (based on the agreement between the donor and the receiving institution) the deposited material becomes the property of the institution and is transferred to the donation area. In our setup the donation area is another email server where the curators can work with the donated emails. This means that the curators can process the emails in a familiar environment using the tools they normally use for handling their own emails. When the curators have finished processing the donated emails, the complete email account is exported to an email account file container (we currently use the PST file format) and this file is then ready for further processing and ingest into our repository system.

7. EXPERIENCES WITH THE SECOND PROTOTYPE

The experiences with the second prototype, which has become the current production system, were much better for everyone involved: donors, curators, and system managers. The curators could work with the donated emails in the same way that they work with their own email, and the work process was easy and well-know. Similarly the donors had the same experience in their donation process which they also found easy and familiar.

The configuration of their email account on their own devices caused problems for many donors. Even though the configuration of the email account only had to be carried out once, we had to put a lot of effort into the user manual. This part of the system was not completely standard as we for security reasons was using other ports and encryptions than the ones most email clients employ as defaults.

Many of the donors did not want to read the user manual, particularly when it came to setting up port numbers and encryption standards. Furthermore, given the many different email clients in different versions it was not possible to write documentation for every single one, and this complicated the configuration process for some donors.

In most cases the curators were able to help the donors with the email-client configuration. When a donor's email client was properly set up, no further problems were observed in the depositing process itself.

The new method of depositing emails provided a more intuitive way of depositing for those donors who prefer a periodical process. At the same time the difficulty of depositing emails for the donors who prefer a continuous deposition process was not increased when comparing with the first prototype where depositing was done using BCC or forward.

Furthermore, the new method of depositing emails has the advantage that the donor can easily organize their emails into folders or upload entire folders if they prefer. In addition to this the donor has full access to the email account and can also delete emails if they want.

8. INGESTING EMAILS INTO OUR REPOSITORY

When we ingest the received emails into our repository, we employ some of the same tools used by institutions having similar ingest workflows, e.g. The University of Manchester Library [1]. However, the way we use these tools and particularly the way our repository is structured is very different. We ingest the donated emails into our repository system (which is based on Hydra [15] and Fedora Commons [16] version 4). Different representations of the email account are ingested. The email container file is one representation and this representation is ingested manually by our curators using the repository's web interface for upload of files and addition of metadata. We also ingest another representation of the email account where the account has been "unfolded" into its parts (folders, emails, attachments, and their relations). See the sketch in Figure 1 for an example case. The transformation from the container representation to the multi-parted representation is done using the Aid4Mail program [6]. A specialized script has been produced that bundle the different Aid4Mail processes and extract additional metadata.

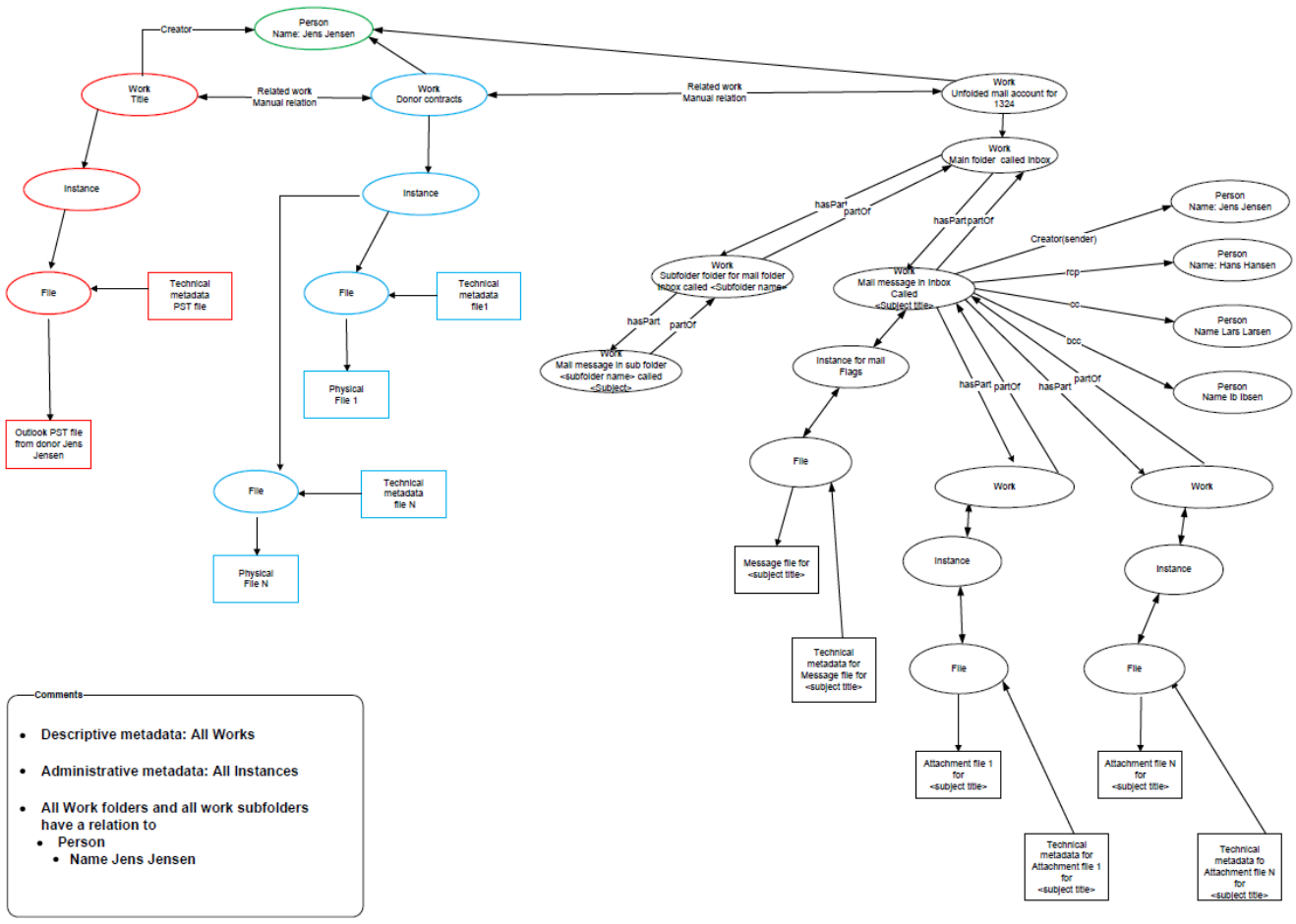

Figure 1. Handling of email objects and metadata

Another product of the transformation is a XML representation of the email container which contains structural information, email-header information, the email body in Unicode text format, and information about the individual emails attachments. We use this XML representation to generate metadata for the individual objects (folders, emails, and attachments) and their relations when ingesting them into our repository.

9. LINKED DATA AND EMAILS

Our repository supports linked data and uses RDF within its data model. We use this feature to create relations between the objects. For example: hasPart and its reverse partOf holds the relationship between folders and emails and between emails and attachments. Furthermore, we use RDF relations to connect emails with agents where the agents act as sender or recipient of the emails.

In the long-term perspective this use of linked data can connect not only our donors internally within our system, but in principle also our donors to other email collections in other institutions. This means that networks with the correspondence of, for example, a group of researchers can be formed.

In a preservation context ingesting the different email representations into our repository system provides the possibility to perform file characterisation on all the parts of the email collection; the email container files, individual emails, and attachments. The ability to do this characterisation on the whole content allows us to perform preservation watch. If we only ingested the container file we would not be able to perform a complete preservation as currently no characterization tools are able to unpack the container file and perform a characterization on its individual objects. The cost of this approach is obviously an increase in the amount of storage (roughly doubling it). However, we can still decide not to long-term preserve every representation so there is not necessarily an increase in the storage cost for long-term preservation.

Having a multi-parted representation in our repository also allows us to preserve individual emails or attachments, or groups of these, at different preservation levels. The different preservation levels could for example consist of giving a particular selection of objects a higher bit safety. Furthermore, in a dissemination context where there are restrictions on the email container, the restrictions on individual emails or attachments or groups of these can be lowered, making it possible to disseminate them to a much broader audience.

10. FUTURE WORK

The email project is still active, and there is still time to explore alternative or supplementing methods for the acquisition of emails. Also the task of finding good ways of disseminating the email collections has not yet begun.

10.1 Alternative Acquisition Methods

An alternative or supplementary way of acquiring our donors' emails could be to harvest them. This could be done in a similar way to the one we employ in our web harvests. This process would require the use of the IMAP protocol and therefore the use of other tools than the ones used in a standard web harvesting would be necessary. Challenges concerning authentication in connection with harvesting of a donors' email account would also have to be solved. A simple proof of concept has been made and the method is worthy of further investigation.

We are also interested in allowing the deposit of other digital materials. These could be video and audio files which in general

are large in size. Even though the IMAP protocol supports transfer of large (in size) attachments, our experience is that it is not the best protocol for the task, as the performance in general is poor.

Instead a possibility could be to use a "Dropbox like" solution; another could be the use of sneakernet (physically moving media like external hard drives or similar devices).

10.2 Dissemination of Emails

At the current phase in the project we have only just begun considering the possibilities for a dissemination of the acquired emails. We considering two tools for this purpose: a standard email client (like Outlook [17]) and ePadd (formerly known as MUSE) [4], [5], [18].

The use of Outlook or similar email clients will give the end-user a well-know experience in which the search and reading of emails would be done in the same way as when the user handles their own email. The use of ePadd gives a greater series of possibilities for the users such as entity extraction, easy browsing and thematic searching. However with new possibilities come new features to be learned by the user, so this option would most likely mean more work both for the users and the curators.

Other alternatives or supplements to these tools should also be considered and tested, but our starting point will be the testing of the two above mentioned tools in collaboration with our curators and users.

11. ACKNOWLEDGMENTS

We would like to thank all that have helped with the development of the service, especially the members of the email project team.

12. REFERENCES

1. Fran Baker. 2015. E-mails to an Editor: Safeguarding the Literary Correspondence of the Twenty-First Century at The University of Manchester Library. *New Review of Academic Librarianship* 21, 2: 216–224. http://doi.org/10.1080/13614533.2015.1040925
2. Andrea Goethals and Wendy Gogel. 2010. Reshaping the repository: The challenge of email archiving. *iPRES 2010*: 71.
3. Marlan Green, Sue Soy, Stan Gunn, and Patricia Galloway. 2002. Coming to TERM: Designing the Texas Email Repository Model. *D-Lib Magazine* 8, 9.
4. Sudheendra Hangal, Peter Chan, Monica S. Lam, and Jeffrey Heer. 2012. Processing email archives in special collections. *Digital Humanities*.
5. Sudheendra Hangal, Monica S. Lam, and Jeffrey Heer. 2011. MUSE: Reviving Memories Using Email Archives. *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, ACM, 75–84. http://doi.org/10.1145/2047196.2047206
6. Fookes Software Ltd. Aid4Mail : Reliable Email Migration, Conversion, Forensics and More. Retrieved April 12, 2016 from http://www.aid4mail.com/
7. Christopher J. Prom. 2011. *Preserving email*. Digital Preservation Coalition.
8. P. Resnick. 2001. *RFC-2822 Internet Message Format*. The Internet Society.
9. B. Srinija, R. Lakshmi Tulasi, and Joseph Ramesh. 2012. EMAIL ARCHIVING WITH EFFECTIVE USAGE OF STORAGE SPACE. *International Journal of Emerging Technology and Advanced Engineering* 2, 10.
10. Arvind Srinivasan and Gaurav Baone. 2008. Classification Challenges in Email Archiving. In *Rough Sets and Current Trends in Computing*, Chien-Chung Chan, Jerzy W. Grzymala-Busse and Wojciech P. Ziarko (eds.). Springer Berlin Heidelberg, 508–519.
11. Frank Wagner, Kathleen Krebs, Cataldo Mega, Bernhard Mitschang, and Norbert Ritter. 2008. Email Archiving and Discovery as a Service. In *Intelligent Distributed Computing, Systems and Applications*, Costin Badica, Giuseppe Mangioni, Vincenza Carchiolo and Dumitru Dan Burdescu (eds.). Springer Berlin Heidelberg, 197–206.
12. Frank Wagner, Kathleen Krebs, Cataldo Mega, Bernhard Mitschang, and Norbert Ritter. 2008. Towards the Design of a Scalable Email Archiving and Discovery Solution. In *Advances in Databases and Information Systems*, Paolo Atzeni, Albertas Caplinskas and Hannu Jaakkola (eds.). Springer Berlin Heidelberg, 305–320.
13. Emailchemy - Convert, Export, Import, Migrate, Manage and Archive all your Email. Retrieved April 19, 2016 from http://www.weirdkid.com/products/emailchemy/
14. Secure Enterprise Email Solutions for Business | Exchange. Retrieved April 13, 2016 from https://products.office.com/en-us/exchange
15. Hydra Project. *Hydra Project*. Retrieved February 23, 2016 from http://projecthydra.org/
16. Fedora Repository | Fedora is a general-purpose, open-source digital object repository system. Retrieved February 23, 2016 from http://fedoracommons.org/
17. Email and Calendar Software | Microsoft Outlook. Retrieved April 5, 2016 from https://products.office.com/en-US/outlook/email-and-calendar-software-microsoft-outlook?omkt=en-US
18. ePADD | Stanford University Libraries. Retrieved April 5, 2016 from https://library.stanford.edu/projects/epadd