The International Journal of Digital Curation

Issue 1, Volume 1 | Autumn 2006

DCC Workshop Report: E-mail Curation: Practical Approaches for Long-term Preservation and Access, Newcastle-upon-Tyne, April 24 - 25, 2006

Maureen Pennock,

DCC, UKOLN, University of Bath

Summary

A report on the Digital Curation Centre workshop held in Newcastle-upon-Tyne in April 2006 to explore practical approaches for managing, preserving and re-using e-mail records.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

This two-day workshop on the curation of e-mail messages offered delegates and speakers the opportunity to investigate some of the organisational, cultural, and technical issues that must be addressed to ensure that e-mails can be located, retrieved, accessed, and reused over time. The number of attendees was relatively small (around forty), but was probably a contributory factor to the high level of interaction and discussion between delegates, speakers, and organisers.

Session 1: E-mails as Records

Seamus Ross, Associate Director of the DCC, chaired the first session. Maureen Pennock opened the session with an introduction to the issues that would be discussed over the course of the workshop. Her presentation, entitled *E-mails as Records: From Creation to Curation* can be accessed via http://www.ukoln.ac.uk/ukoln/staff/m.pennock/presentations/

Carys Thomas and Garry Booth of Loughborough University followed with a presentation on *Institutional Drivers and Barriers to Archiving E-mails*. This was derived in part from the study carried out at Loughborough in 2003 into Records Management and E-mail. The presentation was based on experiences gained during the project and on the work carried out since. Two surveys were undertaken, one in 2003 and one again recently in 2006, to identify universities with policies on archiving e-mail. The results revealed that the number had risen during the period between the two surveys. This may be due to increased awareness, although it is likely that the Freedom of Information Act (FOI) is also partially responsible. Despite this increase in policies, only one of the universities that responded in 2006 had actually implemented an archiving solution.

Garry noted that although many e-mail archiving solutions are being developed, they often have dependencies on other back-office systems. Implementing such an archiving solution could therefore require a back-office change, and possibly even front office if the system could only be integrated with a specific e-mail client. There was some discussion over the use of folders: one delegate claimed that it was less time-consuming and just simpler to use a good search machine than it was to assign messages to folders, although others replied that finding stored messages was not the only benefit of using folders – messages still had to be classified, it could help identify non-records, etc. A risk management approach was advocated to determine the risks or costs associated with keeping everything as opposed to applying selective retention.

Susan Graham, Records Manager at the University of Edinburgh (UoE) provided the last presentation of the session, again based on lessons learned from practical experience. Her presentation was entitled *Turning back the Tide: E-mail Management Without Specialist Software.* Legal requirements affecting e-mail management were discussed in detail, highlighting not only FoI and Data Protection, but also Health & Safety, Benefits & Taxation, and Prescription & Limitation legislation.

Authenticity and legal admissibility of records were cited as particular issues, for it can be difficult to tell the difference between an authentic record and representations that have been altered. An audit trail is a crucial element of proving authenticity in these circumstances. Susan identified some of the advice that the UoE issues to staff on managing e-mails, particularly the ‘do's’ and ‘don'ts’ on managing current e-mails and other advice on managing legacy mail. These issues and others are covered in

greater detail in the document *Managing Your E-mail*, available from the University records management website at

http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/ManagingEmail/ManagingEmail.htm

Risk management was again cited as the way to determine archiving requirements. More detailed information about the UoE approach followed on implementation, the Policy & Planning demonstrator project, evaluation, key learning points, and next steps. Of particular interest was the sheer number of e-mails that staff and the unit have been able to delete – up to 80%. The presentation finished on a high note, encouraging action despite the fact that most approaches are not yet complete. This encouragement surfaced more than once over the course of the workshop.

After coffee, delegates reassembled to review the DCC curation manual instalment on e-mail curation produced by Maureen Pennock. This met with approval from the audience and several delegates contributed comments and further suggestions for content. The published instalment is now available on the DCC website at:

http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-e-mails/

Session Two: Practical Tools and Approaches

The first session on day two looked at practical approaches and technical tools for e-mail curation. The first two speakers both approached e-mail preservation from a public body records management/archival perspective using an XML-based approach.

Jacqueline Slats of the Nationaal Archief van Nederland introduced the Digital Preservation Testbed Project, which developed an approach for the long-term preservation of several record types including e-mails, and presented the Testbed e-mail preservation object. The Testbed Project developed a plug-in tool to work with Microsoft Outlook that captures additional e-mail metadata and an XML representation of messages formatted to a particular house style. The tool is currently in use as part of a pilot project in the Netherlands Ministry of Defence and in several municipalities. Jacqueline also presented the Testbed cost model, paying particular attention to the costs of preserving e-mail. The model is a spreadsheet tool that can be used to anticipate the costs of preserving different types of records using XML, based on several variables that are filled in according to the needs of a given institution.

Filip Boudrez of the Stadsarchief Antwerpen (Antwerp City Archives) followed with a presentation on the David Project and the eDavid Expertise Centre. After outlining the organisational and legal context in Belgium, Filip talked about the experiences of the Stadsarchief in a pilot project using another e-mail/XML plug-in converter that was developed as part of the David Project. This tool has a different storage, formatting, and metadata approach to the Dutch tool. The tool was reviewed in the pilot project and adapted according to the results and feedback. It now incorporates very few additional metadata fields, requiring only that users identify a 'category' for the e-mail, and provides them with an 'Archive' option built into the email client toolbar. Filip also spoke about the training that is provided for users, and how the e-mail XML files are incorporated into the overall storage infrastructure as part of Archival Information Package (AIP) model corresponding with the Open Archival Information System (OAIS). Both Filip and Jacqueline and highlighted the importance of addressing digital preservation from the creation stage of the object life cycle.

After a short break for coffee, Jason Baron, Director of Litigation for the Office of General Counsel at the U.S. National Archives and Records Administration (NARA)

spoke about NARA's experiences in applying best practice guidelines for managing desktop records. The presentation had a distinctly legal perspective, arising from Jason's position as director of litigation and his experiences with electronic records cases such as the Bush and Clinton Administration e-mails. There have been several commercial and high-profile legal cases in the US with examples of poor e-mail records management practice and Jason used a few of these to illustrate why good e-mail management is so important, particularly given the ubiquity of email, the volume of messages, and the range of attachment types. The problem is compounded by the range of messaging/ communicative software that is now available, such as IM, blogs, webcasting, blackberrys, etc.

NARA and government agencies are required to ensure that records, including e-mails, are retained and the information they hold remains accessible for as long as their retention period demands. U.S. Records laws and electronic discovery intersect as follows:

- As a baseline the U.S. Federal Records Act requires appropriate preservation of all electronically stored information that falls within the definition of a 'federal record'.
- The existence of a valid record retention policy is a factor used by courts in considering whether to impose sanctions when hearing allegations of destruction of evidence.

Retention schedules are thus a vital part of a justified disposal/retention schedule not just for archives, but also for legal accountability in discovery cases.

NARA's current e-mail policy has taken into account the results of legal cases involving e-mail retention, particularly regarding the format in which e-mails are stored. The current stance is that agencies must manage the unique 'electronic' e-mail record as it is only a 'kissing cousin' of a hard-copy printout. Hard copies are only acceptable if they include all of the information, including headers, contained in the electronic version of the e-mail. Jason also briefly introduced some of the Sedona principles from the Sedona Conference Working Group on Best Practices for Electronic Document Retention and Production and the Best Practice Guidelines for Managing Information and Records in the Electronic Age.

The final speaker of the morning was Jeremy John of the British Library (BL), who presented his work on the Digital Manuscripts Project. This project aims to extract, identify, catalogue and preserve digital records of scientists deposited with the BL. In many cases, deposits consist of badly labelled discs of varying type, hard drives, and sometimes even old paper hole-punch tape. The challenge of reconstructing a suitable environment that can be used to extract information from these types of carriers and formats represents a significant challenge. The data must first be captured and transferred onto a more modern medium, after which it can be sorted and identified. The whole process takes place according to a life cycle model. E-mails are, of course, included in the types of information deposited. The approach is a collection-level one, but the project has not yet reached the stage whereby conversion of e-mails has been necessary.

Session Three: Reuse of Preserved E-mails

Dave Thompson of the Wellcome Library chaired the final workshop session on the use and reuse of e-mails. This session was intended to provide a perspective on how e-mails can be reused and the types of issues that will arise in e-mail reuse. It was extremely valuable as there are as yet few instances of e-mails publicly available for reuse outside their originating environment.

Susan Davis of the University of Maryland presented the work of colleagues Adam Perer, Ben Shneiderman and Douglas Oard on *Understanding the Rhythms of Relationships in E-mail Archives*. This presentation was based on work carried out using the e-mail collection of Ben Shneiderman, Professor of Computing Science at the University of Maryland. Messages collected over a fourteen-year period were examined using visual methods for identifying and portraying social networks. Susan showed several ways in which relationships between Shneiderman and his e-mail correspondents could be portrayed, including rhythms, distributions, growth rates, comparisons, and queries.

Rhythms can be developed according to organisations, organisation type, country codes, and users, and different visual representations. Analysis is based not on message body or content, but on message header – to, from, subject etc. This reuse of e-mail messages is significantly different to the typical reuse scenario, which usually focusses on the content of the message in the message body, and is a prime example of messages being reused for a completely different purpose to their original function.

Susan Thomas of the Paradigm Project at the University of Oxford then presented on the barriers to reusing e-mails over time. She introduced the Paradigm Project – Personal Archives Accessible in Digital Media – which is exploring the issues involved in preserving digital private papers by gaining practical experience in accessioning and ingesting the papers into digital repositories and processing these in line with archival and digital preservation requirements. Identifying several reasons why one would want to preserve e-mails for future use, such as evidence of social interactions and networks, activities, and correspondence, Susan identified a number of barriers to reuse of e-mail. These include:

- Initial acquisition – problems can be caused by import/export, selection, and even users themselves
- Legal issues – privacy, defamation, intellectual property rights, copyright, evidence of illegal activity
- Metadata – additional metadata is needed to reflect the legal issues involved
- Access restrictions/rights – these need to be compatible with legal issues, which may be slightly different for individual e-mails within a collection

As with several other speakers, Susan was of the opinion that early curation is a prerequisite to reuse. Both presentations together gave a fascinating insight into the challenges we can expect in future access to archived e-mails and on the possibilities for reusing e-mail collections in different ways.

Jason Baron led the final interactive workshop session to survey and provide insights into the current state of affairs in e-mail management. Several participants shared details of their organisation's e-mail management, including archiving, preservation and curation activities. This not only provided a valuable overview of the current state of affairs, but also highlighted the different types of organisational infrastructures represented and the challenges and problems being faced in different

areas. This gave delegates the opportunity to identify others facing similar issues and to discuss as a group some of the ways these could be addressed. The outlook for the future was distinctly positive, with most delegates of the opinion that although few adequate e-mail archiving solutions are currently available, the situation should significantly improve in the near future. The Twenty Questions that Jason used to stimulate this discussion can be found in his presentation at http://www.dcc.ac.uk/events/ec-2006/

Conclusion

It would be fair to sum up by saying most delegates seemed to find the event a valuable experience, not only because of the speakers but also because of the very high levels of participation and interaction between delegates. Feedback since the event has been very positive and all presentations are available on the DCC website - http://www.dcc.ac.uk/events/ec-2006/