*Sergey GLUSHAKOV**

Public Email Archive

* Chief Information Officer, Open Society Archives at Central European University, Budapest; www.archivum.ws

GLUSHAKOV, Sergey, Public Email Archive. Atlanti, Vol. 17, N. 1-2, Trieste 2007, pp. 179-187.

Original in English, abstract in English, Italian and Slovenian, summary in English

Building an email archive raises issues, which still are not sufficiently studied. These include both archival and technological aspects: provenance, authenticity, integrity, data protection, privacy and etc. on the one hand, and variety of platforms and applications to create, transport, retrieve, store and preserve digital records on the other. The discussion is based on the experience acquired by Open Society Archives (OSA) while building The Hungarian Election Campaign Archive, which is available at www.kampanyarchivum.hu. The paper stresses the importance of understanding nature of email records and their lifecycle, relevant standards, infrastructure elements and technical metadata vital in the archival context. It also provides a comparison of basic approaches in archiving email records.

GLUSHAKOV, Sergey, Archivi pubblici di e-mail. Atlanti, Vol. 17, N. 1-2, Trieste 2007, pp. 179-187.

La costruzione di un archivio di e-mail solleva problemi non ancora sufficientemente studiati. Essi includono aspetti sia archivistici che tecnologici: da

Building an email archive raises several issues which still are not sufficiently studied. This concerns both the technological and archival aspects of the problem. In this paper we look at a public email archive as a *collection of records created and/or distributed by general public via electronic mail*. The discussion is based on the experience, which Open Society Archives (OSA) acquired while building *The Hungarian Election Campaign Archive*[1] subsequently in 2002 and 2006. Novelty of the project required tight cooperation of historians, researchers, archivists and IT specialists, not mentioning that it also brought close interest form the public, media and political parties as well.

The Hungarian Election Campaign Archive

The Election Campaign Archive began in 2002, right after the second round of elections has been announced and two political blocs started their fierce contest for power. At this point OSA started to collect election-related email, SMS[2] and MMS[3] messages circulating among public. To corroborate process of acquiring representative mass of electronic correspondence, OSA published an email account and a mobile phone number for people to send in messages they have received. Subsequently, these messages were anonymized to protect sender's identity and published online on a daily basis.

The project itself was aiming to catch up with the unique opportunity, when, according to Andras Mink, historian and editor, *"large number of people responded with forwarding messages supporting, criticizing, accusing or parodying the parties and candidates standing for election, along with messages that called for election rallies, thus contributing to a collection that, on the one hand, represents a peculiar field of application of new information and communication technologies, and, on the other hand, provides a unique snapshot of a post-communist country's election campaign."*

Challenges Dealing With Email

While dealing with the email messages[4], we are aware of the following factors:

- The message we have composed might look very differently on

1. The archive is permanently available for online access at http://www.kampanyarchivum.hu
2. SMS, short for Short Message Service, is a protocol which allows exchange of short text messages between mobile phones.
3. In addition to the ability to send text provided by SMS, Multimedia Messaging Service (MMS) allows exchange of multimedia objects (still and moving images, audio, rich text) between mobile phones.
4. Though The Election Campaign Archive has both email and SMS/MMS messages, we focus only on the former one as the issues related to dealing with the SMS/MMS messages do not pose substantial difficulties.

the recipient's screen.

- Email not always can be saved in a hard copy.
- The text of the original message being replied or forwarded could have been modified.
- The message we have received could have been sent from (and also to) someone else.
- Others could have read this email on its way here.
- Emails are not destroyed when we delete them.
- Emails (and especially attachments) can be dangerous to your computer.

The following example demonstrates only one of numerous problems which can happen to an email message, for instance when the message contains diacritical characters.


-----Original Message-----
From: Marc Elie [mailto:marcel@mail.com]
Sent: Sunday, July 02, 2006 9:33 PM
To: info@example.com
Subject: samizdat

Sehr geehrte Damen und Herren,

Das Russisch-französische Zentrum für Sozialwissenschaften in Moskau (geleitet von Alexis Berelowitch) plant ein wissenschaftliches Kolloquium zum Thema "Formen der alternativen Äußerung in der Sowjetunion der Fünfziger Jahre" für November 2006.

Figure 1a. Message text containing diacritic.


-----Original Message-----
From: Marc Elie [mailto:marcel@mail.com]
Sent: Sunday, July 02, 2006 9:33 PM
To: info@example.com
Subject: samizdat

Sehr geehrte Damen und Herren,

Das Russisch-franzsische Zentrum fr Sozialwissenschaften in Moskau (geleitet von Alexis Berelowitch) plant ein wissenschaftliches Kolloquium zum Thema "Formen der alternativen uerung in der Sowjetunion der Fnfziger Jahre" fr November 2006.

Figure 1b. Diacritical characters distorted or lost due non-matching encoding used by recipient's MUA[5] or operating system.

Other types of presentation-related problems appear when messages are sent using national-specific encodings or such enhancements as text formatting[6]. What is less obvious, that it is the **combination of multiple factors** which might lead to a problem:

- different operating system installed on sender's and recipient's computers (could be Windows on the one and Macintosh on the other side),
- particular versions of software applications used by sender and recipient (could be web-mail accessed through the Internet browser

5. Mail user agent (MUA) is a software application used to compose and read email messages, it is also often called email client.
6. For instance, RTF (Rich Text Format) or HTML (Hypertext Markup Language).

un lato, la provenienza, l'autenticità, l'integrità, la protezione dei dati, la privacy, ecc., dall'altro la varietà di software e di applicativi utilizzati per creare, trasportare, ricercare, archiviare e conservare i documenti in formato digitale. La discussione si basa sull'esperienza acquisita dall'OSA nella costruzione del "Archivio ungherese della campagna elettorale", disponibile all'indirizzo: www.kampanyarchivum.hu. L'articolo sottolinea l'importanza di capire la natura dei documenti e-mail ed il loro ciclo vitale, gli standard rilevanti, gli elementi infrastrutturali ed i metadati tecnici vitali in un contesto archivistico. Fornisce inoltre un paragone fra approcci basilari nell'archiviazione di documenti e-mail.

GLUSHAKOV, Sergey, Javni elektronski arhiv. Atlanti, Zv. 17, Št. 1-2, Trst 2007, str. 179-1887

Za postavitev stavbe elektronskega arhiva in opreme za takšen arhiv, zahteva mnogo študij, ki jih zaenkrat ni v obilni meri. Študije morajo vsebovati tako arhivske kot tehnološke vidike: po eni strani izvor, avtentičnost, celovitost in povezanost, zaščito podatkov, zasebnost, po drugi strani pa veliko več osnov in podlog ter aplikacij, kako ustvariti, prenesti, hraniti in obraniti digitalne zapise ob drugih zapisih. Razprava upošteva izkušnje, ki jih ima Open Society Archives (OSA) z proučevanjem madžarskega arhiva za volilno politiko, ki je dostopen na spletni strani www.kampanyarchivum.hu. Zato moja razprava

or such proprietary software as MS Outlook),

- various localization settings set up on sender's and recipient's computer (default language, Western- and Central-European character set, with or without Unicode support),

- specific version of mail server software installed and configured by Internet Service Providers on both sides.

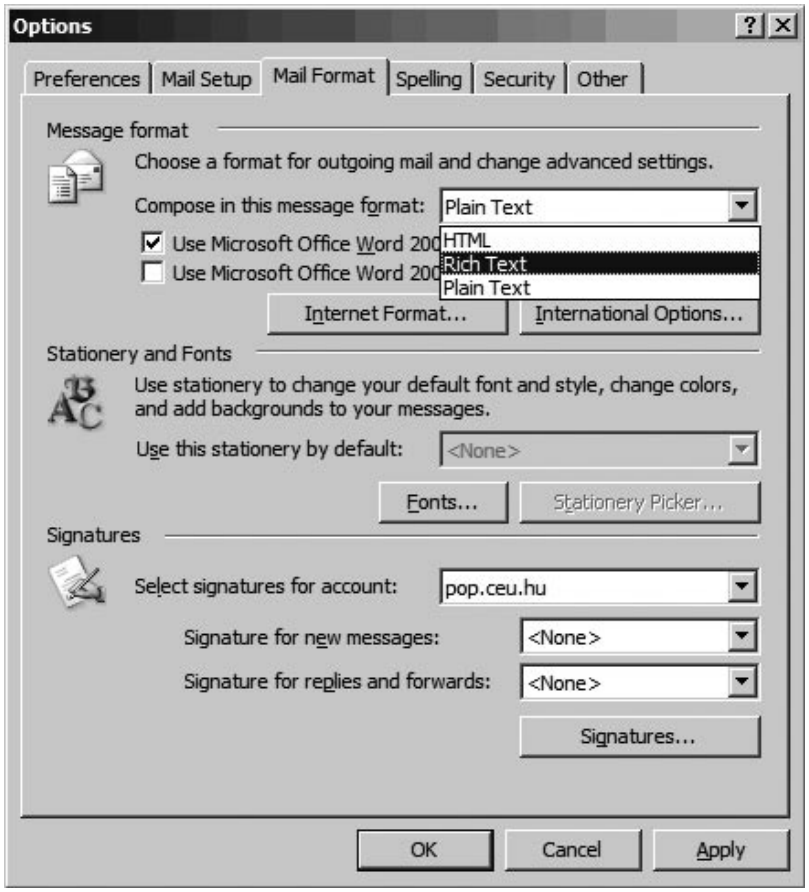
On the following figure we can see an example of how many options email sender has just for setting up his/her version of MUA, which in this example is MS Outlook:

- MS Word can be used for composing email message (and that results in the use of additional formatting data in the email message body, which in turn might or might not be interpreted correctly by the receiving party);

- messages can be sent as plain text, RTF, or HTML (with the plain text option problems might arise when text of the message was composed in a language other than English, or in case of RTF or HTML the receiving party might not be able to render the message correctly);

- furthermore, various proprietary MUA (like Novell GroupWise or IBM Lotus Notes) tend to use additional features which are not always compatible with the rest of the world - like Stationary option on the figure below.

Figure 2. MS Outlook configuration options



Email Message Structure

Email message format is defined by Internet Mail Protocol and RFC 2822[7] in particularly. It consists of three basic components: *envelope*, *header* and message *body*.

The Envelope is used internally by MTA[8] to route email and is not accessible by users, it is here where the real sender and recipient data is contained, not in the From and To fields, which are part of the email header[9].

The Header consists of multiple fields of which only some are visible to users: From, To, Date, CC, BCC, Subject, etc. Some of these fields are *mandatory* (Date, From, To) and some are *optional* (Subject, Reply-To, etc.) Some of the fields *are added by the system* (e.g. unique Message-Id) and some are *set by users* (e.g. Reply-To). Moreover, some of the fields despite being added by the system (like Date which is added by MUA) can be set up to *differ from the real.*

The Body contains the actual email message, which is always sent between mail servers encoded in plain ASCII text, even attached binary files[10] as it demonstrated on the next figure.

Figure 3. Fragment of an email message with an image file attached.

In the above example a binary file, JPEG image file *'200km.jpg'*, has been encoded using base64 algorithm which uses only printable ASCII characters and is supported by any version of SMTP servers[11].

Archiving Email Message
Email Message as Archival Record

It is still quite common practice when only visible text part of email messages is being archived: either by Copy/Paste or by exporting message to a text file. Export to rich text, HTML or XML format saves more data, including metadata for some common header fields like From, To, Date, etc. However, each email message still has im-

poudarja pomembnost razumevanja bistva elektronskih zapisov, njihovo obstojnost, upoštevajoč potrebne standarde, infrastrukturne elemente in tehnično velikost podatkov v okviru arhivistike. Po drugi strani pa raziskava zagotavlja primerjanje temeljnih postopkov in dostopov do arhiviranja elektronskih zapisov.

SUMMARY

The presentation will cover essential aspects of archiving electronic mail, including such issues as various types of content, metadata capture, privacy, long-term preservation, as well as the heterogeneous environment in which emails are created and distributed. The latter refers to diverse end-user hardware and operating systems, variety of applications used to create and send email (either simple text messages, sometimes with complex digital objects attached,

7. RFC, or Request For Comments, is a term used for series of documents adopted by the Internet Engineering Task Force (IETF) for Internet standards.
8. Mail transfer agent (MTA), mail server maintained by Internet Service Providers (ISP)
9. Most obvious case is when BCC (or blind copy) is used to specify recipients.
10. Email message body format is governed by the Multipurpose Internet Mail Extensions (MIME) set of standards based on RFC 2045.
11. Simple Mail Transfer Protocol (SMTP) is used to send a message from MUA to MTA

or richly formatted messages with embedded multimedia content), chain of intermediate servers and nodes to access the Internet and to deliver an email to its destination point. The premises for this paper are two projects implemented by the Open Society Archives (OSA), a private archive located in Budapest, Hungary. In 2002 and 2006 OSA created an online archive of Email, SMS and MMS related to the parliamentary election campaign. Taking into account the sensitive nature of the content, the need to preserve both the authenticity and the author's privacy, have posed significant challenge for the successful implementation of the email archive. Trustworthiness and accountability became the criteria for the selection of formats, technologies and supporting infrastructure, as well as the processes and procedures within the project. The presentation will focus on identifying key technological issues

portant data in its header fields, which might be essential for establishing integrity and authenticity of the message. These fields usually contain information about software environment and context in which email was created (e.g. software application used for composing email, regional settings, MUA, etc.) Another important header field is the Received field, which contains trace information generated by mail servers that have handled email message on its way from originating to destination point.

Typically an email message is passed between at *least* four computers: sender's desktop (MUA), senders' ISP mail server (MTA), recipient's ISP mail server and desktop. Each MTA adds a new Received entry to the header of the message on its way from server to server. On the following example, reading bottom-up, we can establish exact timing and routing points from the moment email message was received by the sender's Internet Service Provider (ISP) mail server *(mail.invitel.hu)* and until three minutes later it was delivered to the recipient's mail server *(osamx)*.

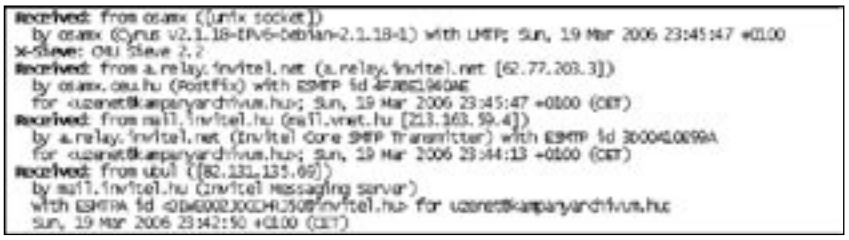


Figure 4. Fragment of an email message Header.

From the archival point of view this is an important task to capture and to preserve this type of information in addition to the message body itself. Only data contained in heard fields can be used to establish authenticity of an email message and its true origins.

Warning

In case of Intranet, or internal mail system used by many organizations, sending and receiving email message can be just a one step long. Proprietary corporate mail systems (like Novell GroupWise, Microsoft Outlook or IMB Lotus Notes) use vendor-specific formats and protocols for internal communication. However, for the external communication Internet mail gateways are used to ensure that outbound mail will reach recipients over Internet by using standard protocols.

Getting Header Data

The ability to interpret headers might not be as important for the archivists - this can always be accomplished with the help of IT staff and by referring to respective standards - as long as the source data has been properly archived. However, getting this data from the message already in the MUA mailbox might be a challenge. Depend-

ing on the particular MUA and its configuration, message header can be only partly visible to the user. For instance, header information shown on Fig.3-4 above with MS Outlook can be accessed through the following context menu only:

```
View-Options-Internet_headers.
```

Also depending on the particular MUA and its configuration, message header might only partly be retrieved from the recipient's mail server. This situation usually can be fixed by configuring MUA software appropriately. In given case, to enable MS Outlook to download entire message source from the server, the following tweaking has to be done[12] [4].

In the MS Windwos registry the following key hast to be modified:

```
HKEY_CURRENT_USER\Software\Microsoft\
Office\11.0\Outlook\Options\Mail
```

A new DWORD has to be created:

```
DWORD: SaveAllMIMENotJustHeaders
Value: 1
```

This will enable MS Outlook MUA to download entire source of email message including its header data.

Best Practice in Email Archiving

As now we looked at various aspects of retrieving complete data of email messages, it is time to compare at least most important implications of options available for archiving.

Approach	Benefits	Drawbacks
Saving individual messages.		Content of the message is separated from its context: valuable metadata (like real date stamp or email address) is lost or substituted with a text record. Binary attachments are no longer associated with the respective message. Labour-intensive, can be done only manually.

12. **Warning**: Serious problems might occur if you modify the registry incorrectly**.**

Savingmessage body along with its header.	Contains detailed metadata about context, origins, encoded original content including attachments, formatting, delivery data.	Proprietary vendor- or third-party solutions might be required. Can be labour-intensive.
Saving entire envelope.	Highest level of integrity: contains most complete and accurate set of metadata, including email account configuration. Can be done in entirely automated way.	Technical expertise and access to MTA is required.

Warning

Use of corporate communication systems (like Novell GroupWise, Microsoft Outlook or IMB Lotus Notes) comparing to their Open Source alternatives might give you better functionality and support, however it also leads to so called lock-in situation, when customers become more and more dependant on a particular proprietary system and cannot migrate to another system because of the high cost associated with such migration or simply inability to migrate data which is already in the proprietary format. Certain solutions are available from the vendors themselves or from third parties, however their focus is removing retired correspondence from the mailing system and relocating it to a separate location or system where it can still be reached as reference data, rather than long-term preservation.

Building Public Email Archive

When OSA initiated The Election Campaign Archive project, the following issues were addressed first: privacy, access, integrity, authenticity and preservation. To encourage wide public participation in collecting representative electronic archive of the election campaign, OSA promised senders to **protect their identity**. On the other hand, from the researcher perspective it would be important to know how many unique senders contributed to the creation of the archive, what is the average number of submissions per sender, etc. Subsequently, while sanitizing published messages became one of the important tasks in the overall project workflow, each sender's unique address got a unique identifier, produced by the database ingest system.

Another aspect of keeping personal data protected, which in

real life is too often overlooked, is removing this data from all the temporary copies of mail server logs, database entries, backup copies, etc. This task can never be fully automated, especially as there are numerous operations on various stages performed by various people: network and database administrators, editors, web designer. This requires professional responsibility and constant control over the whole lifecycle of the email message being archived.

Dealing with such sensitive issue as an election campaign, required our best effort to ensure data integrity for every document acquired. Even though we don't hold responsibility for the content and source of messages archived, from the moment they reach our server they become our responsibility. Also, malicious attacks on the mail server could not be entirely excluded. That is why a standalone installation of *Cyrus IMAP* server[13] had been chosen as MTA: it can run on sealed servers, where normal users are not permitted to log in. Cyrus IMAP supports *mbox* format for holding email messages as plain text and thus is an open, platform-independent solution suitable for long term preservation.

It was expected that number of submissions to the archive will be constantly growing, so most of the acquisition and processing operations were automated: ingest of the newly arriving email messages from the campaign mailbox to the processing database, metadata capture and, after semi-automatic anonymization by the editors, online publishing of newly arriving messages.

Conclusion

It has been estimated that in 2007 only in North America alone eight million email messages will be sent [5]. How many of them will be archived and how many will be lost immediately or over some time? This very brief summary gives only rough outline of the spectrum of possible issues of which archivists have to be aware when dealing with email archiving and preservation. Obviously, concrete solutions will be different for each workflow/preservation scenario. For instance, higher level of standartisation and automation can be achieved when email messages are created within an institution which operates according to predefined workflow in homogenous environment. Also, best scenario always depends on resources and expertise available. Nevertheless, all the issues above to certain extent concern any project or activity concerned with the email archiving.

Glossary

Domain Name System (DNS) server stores listing of mail exchange servers which can relay email message to the destination MTA.

Fully Qualified Domain Address (FQDA), or email address, consists of the local part (often the user name) and a domain name, e.g. *username@mailservername.com*.

14. The Cyrus IMAP server development started at Carnegie Mellon University in 1994.

Mail transfer agent (MTA), mail exchange server maintained by Internet Service Providers (ISP).

Mail user agent (MUA) is a software application used to compose and read email messages, also often called email client.

Multimedia Messaging Service (MMS) as further development of SMS is a protocol which allows exchange of multimedia objects (still and moving images, audio, rich text) between mobile phones.

Multipurpose Internet Mail Extensions (MIME) is set of standards based on RFC 2045 which specifies email message body format.

Post Office Protocol (POP3) based on RFC 1939 is used to retrieve messages from MTA to MUA.

Internet Message Access Protocol (IMAP) defined by RFC 3501 is one of two prevalent protocols (the other being POP3) for retrieving email messages from mail server (MTA).

Protocol is a set of rules which enables data exchange between hardware devices and software applications.

Request For Comments (RFC) is a term used for series of documents adopted by the Internet Engineering Task Force (IETF) for Internet standards.

Short Message Service (SMS) is a protocol which allows exchange of short text messages between mobile phones.

Simple Mail Transfer Protocol (SMTP) is used to send a message from MUA to MTA.

Bibliography

"*The archive of electronic campaign letters*", background paper by Andras Mink; 2002, revised 2006.

Registry of Message Header Fields, http://www.iana.org/assignments/message-headers/perm-headers.html

RFC Index, Internet Engineering Task Force, http://tools.ietf.org/rfc

"How Outlook applies encoding to plain text e-mail messages", Microsoft Knowledge Base http://support.microsoft.com/kb/278134

"Worldwide Email Usage 2005-2009 Forecast", Legal Technology Resource Center Survey Report, 2006.

Free Online Dictionary of Computing, http://foldoc.org/

Wikipedia, the free encyclopedia, http://en.wikipedia.org