



Digital Preservation Testbed

XML for Preservation

Maureen Potter

Urbino, October 10th 2002



Testbed Background

- Established October 2000 by
 - Ministry of the Interior
 - Ministry of Education, Culture and Sciences
 - National Archives
- Will finish in October 2003
- Objective:
 - To secure sustained accessibility to reliable government information in the digital era



Research Questions

- Advantages of different preservation approaches?
- Factors and effectiveness of each approach?
- Basic Requirements for preservation?
- Which metadata are essential for preservation?
- Options for Attribute preservation?

Scope

- 4 Archival Record Types:
 - Text documents
 - Spreadsheets
 - Emails
 - Databases
- 3 Preservation Approaches:
 - Migration
 - Emulation
 - XML

Advantages of XML (1)

- Multiple Uses
 - Metadata storage and exchange
 - File format
 - Object Linking and referencing
 - Encapsulation
- Can represent different record attributes:
 - Content and Context (Basic ability of XML)
 - Structure (DTD or Schema)
 - Appearance (dictated by Style Sheet)

Advantages of XML (2)

- A reliably standard Standard
 - Well controlled through W3C
 - Planned for interoperability
- Self describing and human readable
- Good for generating indexes or searching aids
- Supported by most software and tools

Disadvantages of XML

- User Scepticism
- Software must still be paid for
 - For conversion
 - For processing
- Superseded in 10 years?
- Verbose – makes files bigger

XML for Emails

- Email as a standardised record type
 - MIME format controlled by IETF
 - Well defined, well structured, text based
 - Interoperable on different platforms
- XML as a standardised format
 - XML format controlled by W3C
 - Well defined, well structured, text based
 - Interoperable on different platforms
- Conversion is therefore relatively straightforward



Implementation Options

- Post-Use Conversion
 - The ‘All In One’ Option
 - All message content wrapped in one XML file
 - The ‘Split Files’ Option
 - Headers, Body, Attachments and metadata all stored separately in appropriate formats
- Pre-Use Provision
 - The ‘Forward Facing’ Option
 - Email to XML Demonstrator
 - stores XML version of message at time of transmission



All-In-One (1)

All contents in one file

- Step 1: Obtain Transmission File

```
Received: from gw01.dh01.ictu.nl (unverified) by ms01.dh02.ictu.nl
(Content Technologies SMTPRS 4.2.5) with ESMTTP id <T59961a58390a1305020b3@ms01.dh02.ictu.nl>
Tue, 12 Mar 2002 08:39:58 +0100
X-MimeOLE: Produced By Microsoft Exchange V6.0.5762.3
content-class: urn:content-classes:message
MIME-Version: 1.0
Content-Type: multipart/mixed;
    boundary="-----=_NextPart_001_01C1C99A.A6A4DE16"
Subject: FW: mr-KMail-msg-24
Date: Tue, 12 Mar 2002 08:51:00 +0100
Message-ID: <C3719FE945884C4EBEFA3C4C72EEFE983306889@gw01.dh01.ictu.nl>
X-MS-Has-Attach: yes
X-MS-TNEF-Correlator:
Thread-Topic: mr-KMail-msg-24
Thread-Index: AchJMPDEdVxIrIXzQ7+EV6bGW3W32wAaauKw
From: "Maureen Potter" <Maureen.Potter@ictu.nl>
To: <Testbed@10.18.1.241>

This is a multi-part message in MIME format.

-----=_NextPart_001_01C1C99A.A6A4DE16
Content-Type: text/plain;
    charset="iso-8859-15"
Content-Transfer-Encoding: quoted-printable
```



All-In-One (2)

- Step 2: Testbed uses Java tool to convert transmission file into XML file

```
- <record>
+ <metadata>
- <email>
- <headers>
- <header>
  <headerName>Received</headerName>
  <headerValue>from ms01.dh02.ictu.nl ([10.19.5.2]) by tb1 (Build 101 8.9.3/NT-8.9.3) with ESMTMP id IAA
  <Testbed@10.18.1.241>; Tue, 12 Mar 2002 08:55:04 +0100</headerValue>
</header>
- <header>
  <headerName>Received</headerName>
  <headerValue>from gw01.dh01.ictu.nl (unverified) by ms01.dh02.ictu.nl (Content Technologies SMTPRS
  ESMTMP id <T59961a58390a1305020b3@ms01.dh02.ictu.nl> for <Testbed@10.18.1.241>; Tue, 12 Mar
  08:39:58 +0100</headerValue>
</header>
- <header>
  <headerName>X-MimeOLE</headerName>
  <headerValue>Produced By Microsoft Exchange V6.0.5762.3</headerValue>
</header>
- <header>
  <headerName>content-class</headerName>
  <headerValue>urn:content-classes:message</headerValue>
</header>
- <header>
  <headerName>MIME-Version</headerName>
```

All-In-One (3)

- Step 3: Examine Contents
- Record.xml
 - Metadata (additional)
 - Headers (To, From etc)
 - Body (Message Content.txt)
 - Headers (Content type definition)
 - Body (eg Attachment.html)
- /Record.xml

```
- <record>
+ <metadata>
- <email>
+ <headers>
- <multipart-mixed>
- <part>
+ <headers>
- <body>
+ <![CDATA[  ]>
</body>
</part>
- <part>
- <headers>
+ <header>
+ <header>
+ <header>
+ <header>
</headers>
- <body>
+ <![CDATA[  ]>
</body>
</part>
</multipart-mixed>
</email>
</record>
```



Split-Files (1)

All components in separate files

- Step 1: Obtain Transmission File

```
Received: from gw01.dh01.ictu.nl (unverified) by ms01.dh02.ictu.nl
(Content Technologies SMTPRS 4.2.5) with ESMTMP id <T59961a58390a1305020b3@ms01.dh02.ictu.nl>
Tue, 12 Mar 2002 08:39:58 +0100
X-MimeOLE: Produced By Microsoft Exchange V6.0.5762.3
content-class: urn:content-classes:message
MIME-Version: 1.0
Content-Type: multipart/mixed;
    boundary="-----=_NextPart_001_01C1C99A.A6A4DE16"
Subject: FW: mr-KMail-msg-24
Date: Tue, 12 Mar 2002 08:51:00 +0100
Message-ID: <C3719FE945884C4EBEFA3C4C72EEFE98330689@gw01.dh01.ictu.nl>
X-MS-Has-Attach: yes
X-MS-TNEF-Correlator:
Thread-Topic: mr-KMail-msg-24
Thread-Index: AchJMPDEdVxIrIXzQ7+EV6bGW3W32wAaauKw
From: "Maureen Potter" <Maureen.Potter@ictu.nl>
To: <Testbed@10.18.1.241>

This is a multi-part message in MIME format.

-----=_NextPart_001_01C1C99A.A6A4DE16
Content-Type: text/plain;
    charset="iso-8859-15"
Content-Transfer-Encoding: quoted-printable
```

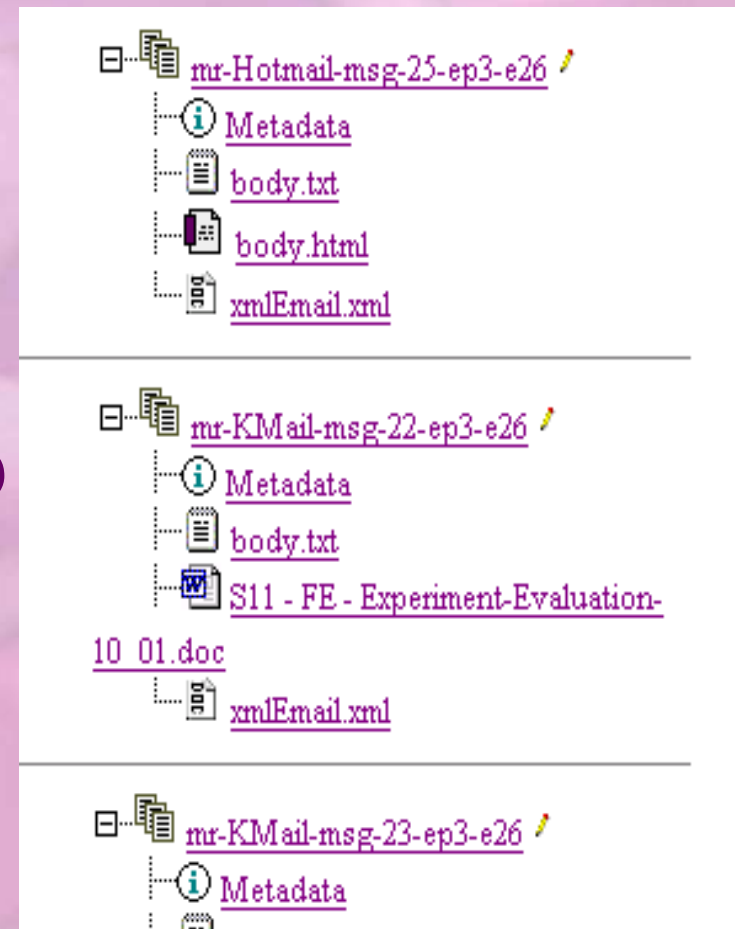
Split-Files (2)

- Step 2: Testbed uses Java tool to separate digital components of transmission file and transform headers into XML file

```
<?xml version="1.0" ?>
- <record>
- <email>
  - <headers>
    <Received>from ms01.dh02.ictu.nl ([10.19.5.2]) by tb1 (Build 101 8.9.3/NT-8.9.3) with ESMT
      <Testbed@10.18.1.241>; Tue, 12 Mar 2002 08:55:04 +0100</Received>
    <Received>from gw01.dh01.ictu.nl (unverified) by ms01.dh02.ictu.nl (Content Technologies S
      id <T59961a58390a1305020b3@ms01.dh02.ictu.nl> for <Testbed@10.18.1.241>; Tue, 12
      +0100</Received>
    <X-MimeOLE>Produced By Microsoft Exchange V6.0.5762.3</X-MimeOLE>
    <content-class>urn:content-classes:message</content-class>
    <MIME-Version>1.0</MIME-Version>
    <Content-Type>multipart/mixed; boundary="-----_=_NextPart_001_01C1C99A.A6A4DE16"</
    <Subject>FW: mr-KMail-msg-24</Subject>
    <Date>Tue, 12 Mar 2002 08:51:00 +0100</Date>
    <Message-ID><C3719FE945884C4EBEFA3C4C72EEFE98330689@gw01.dh01.ictu.nl></Message
    <X-MS-Has-Attach>yes</X-MS-Has-Attach>
    <X-MS-TNEF-Correlator />
    <Thread-Topic>mr-KMail-msg-24</Thread-Topic>
    <Thread-Index>AChJMPDEdvxIrIXzQ7+EV6bGW3W32wAaauKw</Thread-Index>
```

Split-Files (3)

- Step 3: Examine Contents
- Record Object
 - Metadata (Recordkeeping)
 - Body (Message Content.txt)
 - Body (Message Content.html)
 - Attachment (any format)
 - Headers (xmlEmail.xml)
- /Record Object





Forward-Facing (1)

- Two applications:
 - Add-In for Outlook; the email is converted to XML behind-the-scenes
 - Web service; validating the XML, transforming to HTML, and storing separately
- Two template options: formal and informal
- Compulsory metadata completion
- Preview of message in HTML
- Email sent in HTML; XML version stored centrally

Forward-Facing (2)

XML version AND regular version

- Step 1: Develop demonstrator
- Step 2: Integrate with existing email application
- Step 3: Implement throughout organisation
- Step 4: Harvest and Store XML representations of messages via Web Server



Conclusions

- XML is highly suited towards Emails
- Implementation depends on the Institution
- Pick the approach to suit your needs
- Incorporate additional metadata
- Store for the long term

<http://www.digitaleduurzaamheid.nl>