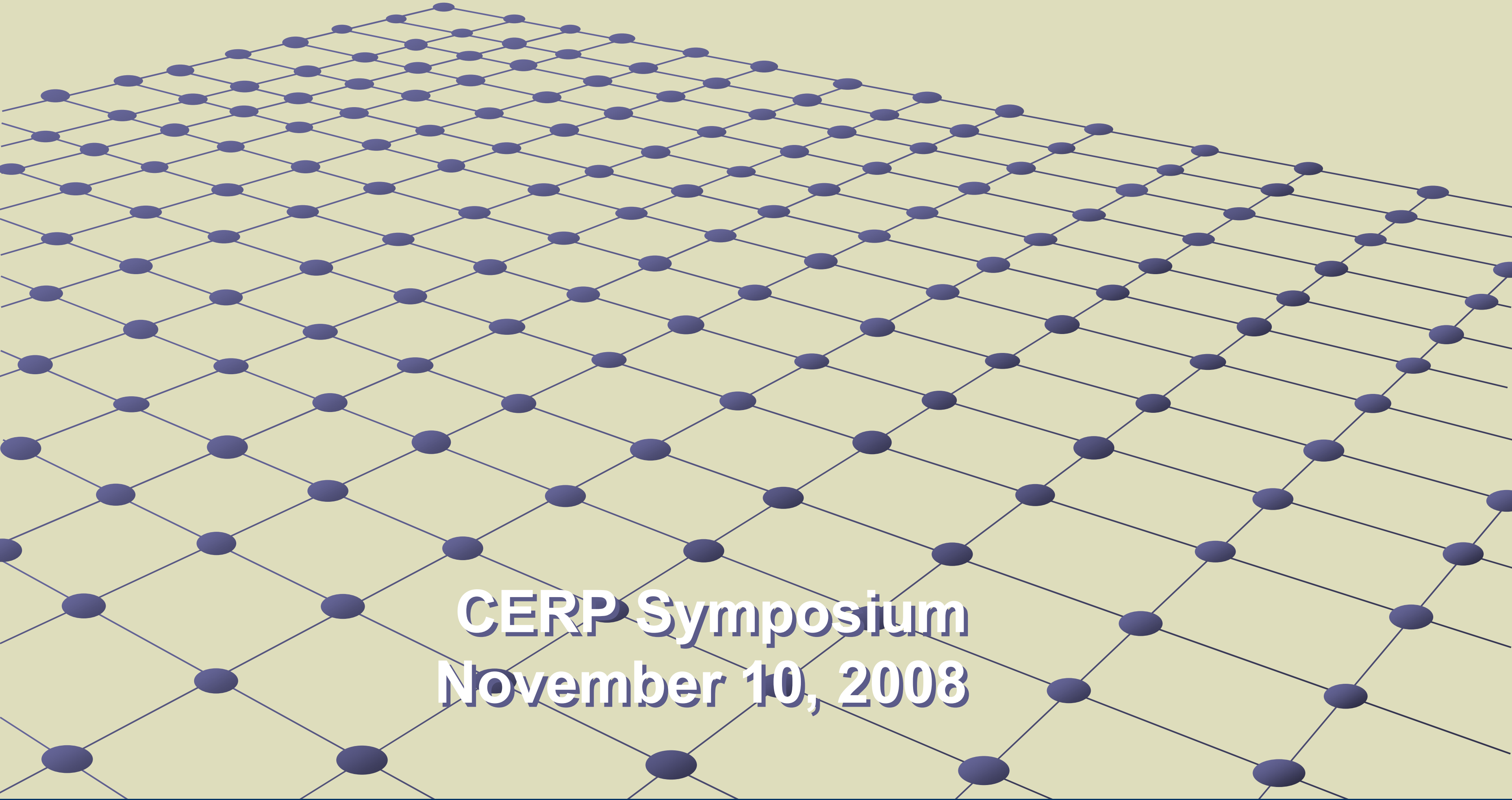
Parsing E-Mail: Lessons Learned



CERP Symposium
November 10, 2008

THE ROCKEFELLER ARCHIVE CENTER

Smithsonian Institution Archives

Overview of Topics

- Just what is email anyway?
- Email standards and conventions
- Diversity of native email formats
- Commercial tools vs. open source
- The co-evolution of the schema and the parser

What is Email?

- Email is what is transmitted from the sender to the receiver.
- It is not simply what the receiver sees.
  - The email client software determines what you see
  - Multiple alternative bodies – you only see one of them
  - Child messages may or may not appear in-line, or at all
  - HTML rendering may differ on different machines
  - Headers may contain extra useful information
- We must archive all information that was transmitted and stored, not just what was visible

Weak Email “Standards”

- RFC2822 and other standards are less standard than they seem.
- Email continues to evolve and standards continue to lag.
- Lagging standards attempt to support all preexisting conventions … an impossible goal without compromises that are open to interpretation.
- Different email client vendors interpret the standards differently. Causes mismatches between interpretations (and inevitable bugs).

Variety is the Spice of Email

- Dozens of common email systems and hundreds of others
  - We have encountered mail from Eudora (multiple versions), Simeon for MacPPC, Outlook/Exchange (multiple versions), AppleMail, Lotus Notes, Groupwise, Mozilla/Firefox, Pegasus Mail, and various Internet mail services such as gmail, Hotmail, YahooMail, Juno, and AOL. Each has its peculiarities.
- Non-standard date and time-zone formats
- European and Asian mail may contain non-ASCII (actually, non UTF-8) characters
- Older email may have HTML in inappropriate places
- Treatment of nested forwarded and other “child” messages differs

Commercial vs. Open Source

- Weaknesses of Commercial Solutions
  - Most SARBOX solutions aim at the earliest possible legal destruction of email rather than long-term storage.
  - The storage formats are determined by the vendor, usually with an eye to supporting their own client software and advantaging their own business
  - Proprietary software suppliers may not even be in business 20 years hence.
- Benefits of Open Source
  - The software can be maintained by the archivist community at large,
  - Storage formats can be optimized for archival needs.

The Storage Format - XML

- Why not just use Native email format?
  - Which one? How well is it documented? How long will software exist to read it? Which companies (if any) have a real commitment to stability and longevity?
- Why e**X**tensible **M**arkup **L**anguage (XML)?
  - XML is open, human readable and “self describing”
  - A good descriptive schema supports validity checking
  - There are many open source tools to create, manipulate and read XML

The Importance of a Common Schema

- A Schema defines how the XML tags for the various parts of an email relate to each other.
  - <Account>, <Folder>, <Message>, <Header>, <Body>, <Attachment>, etc.
- It is the Rosetta stone that guides how raw email is converted to XML
- … and it defines the structure for subsequent search, display, provenance, preservation, etc.
- The **'Mail-Account'** XML schema serves the purposes of both CERP and EMCAP (thanks to David Minor of the NC State Archives)
- It is public, so you don't have to reinvent the wheel

Lessons Learned

- Email isn't easy and standards aren't very standard
- Child messages can be nested deeply -- complicates parsing, the schema, and search
- Recent email is reasonably well behaved
- Older email can contain all sorts of problematic email
- Email from overseas may have its own problems (dates and non-ascii characters)

Email Conversion Results

- We have converted and validated more than 70,000 in test sets to the XML Mail-Account schema
  - Smithsonian - 5,537 messages in 232 Mb of recent Outlook mail
    - 99.97% successfully parsed (4 could not be parsed),
  - Smithsonian - 20,000 messages in a 1.5 Gb Outlook account
    - 99.975% successfully parsed (5 could not be parsed)
  - Rockefeller Archives - 43,778 messages in 378 Mb of older eclectic mail
    - 99.85% successfully parsed (74 unparsed, but improvement is clearly possible)
- Parse speed: about a quarter gigabyte per hour on a Thinkpad T40

Assessment of the Schema

- The schema has passed the test of handling the complexities found in real-world email from institutions like RAC and SIA
- The schema provides for and encourages the capture at parse time of useful metadata (the main headers for instance).

Summary

- 100% success is an unrealistic goal
    - Some emails are just too broken to parse without manual intervention
- We ***can*** achieve at least 99.9% success (and save the few unparsed emails for human inspection and repair)
- This error rate is comparable to that of physical archives