Archiving and Accessing Email from Failed Companies

David Kirsch
Douglas Oard
Jyothi Vinjumur
University of Maryland
http://ediscovery.umiacs.umd.edu/

Library of Congress
Archiving Email Symposium

June 2, 2015
Outline

- Introduction & Background
- Collection Development
- Sample Collections
- Ongoing & Next Steps
Background: User-driven inquiry

- How will we learn about what really goes on inside modern firms?
- Surviving firms unlikely to grant researchers access to large, unvetted email collections
- Contra the Whig view that “history is written by the winners,” records of failures may be all that survives
- Failed firms are important, understudied historical agents

Source: Kirsch, “Record of Business at Risk” (2009)
Partners

- **Intermediaries**
  - Sherwood Partners (Workout specialist)
  - FTI Consulting (Bankruptcy Trustee)
  - Gallivan, Gallivan & O’Melia (Custodian)

- **Sponsors**
  - Sloan Foundation (BPA)
  - LoC / NDIIPP (DABODCE)
  - NSF (eDiscovery)

- **Scholars**
  - RRCHNM at GMU
  - UMD i-School / UMIACS
  - Advisory Council

- **Repositories**
  - Hagley Museum & Library (NDP)
  - Linguistic Data Consortium at UPenn
  - LC
  - Others?
Collection Development

Question: Will we ever see it?

Ongoing Firms Generate Email

Researchers Access Email
Collection Development

Question: Will we ever see it?

Researchers Access Email

Ongoing Firms Generate Email

Firms Fail & Lose Control of Records, incl. Email

Researchers Access Email

UMD LDC LoC

Repositories Process Collections for Distribution

3rd Parties Aggregate & Stabilize Collections

Sherwood Bankruptcy Tte Gallivan

Ongoing Firms Generate Email
Sample Collections: AvocadoIT

- Conducted risk assessment
- Balance technical & institutional mechanisms to mitigate risks

**Technical**
- Limited redaction (PII & images)
- Anonymization (AvocadoIT)

**Institutional**
- Limited access via LDC @ UPenn
- All users and institutions must sign license agreements
**Avocado Research Email Collection**

- **Item Name:** Avocado Research Email Collection
- **Author(s):** Douglas Oard, William Webber, David Kirsch, Sergey Golitsynslyi
- **LDC Catalog No.:** LDC2015T03
- **ISBN:** 1-58563-704-1
- **ISLRN:** 102-408-869-995-0
- **Release Date:** February 16, 2015
- **Member Year(s):** 2015
- **DCMI Type(s):** Text
- **Data Source(s):** email
- **Application(s):** information retrieval, discourse analysis, topic detection and tracking
- **Language(s):** English
- **Language ID(s):** eng
- **License(s):**
  - Avocado Collection - Organization Agreement
  - Avocado Collection - Individual Agreement
- **Online Documentation:** LDC2015T03 Documents
- **Licensing Instructions:** Subscription & Standard Members, and Non-Members
- **Citation:** Oard, Douglas, et al. Avocado Research Email Collection LDC2015T03. DVD. Philadelphia: Linguistic Data Consortium, 2015.

Source: https://catalog.ldc.upenn.edu/LDC2015T03
Avocado Research Email Collection
Version 1.0.3

Douglas W. Oard <oard@umd.edu>
William Webber <wew@umiacs.umd.edu>
David A. Kirsch <dkirsch@umd.edu>
Sergey Golitsynskiy <sergeig@umd.edu>

README

Date: Mon Nov 24 10:18:44 EST 2014

1. INTRODUCTION
=============

The Avocado Research Email Collection ("the Avocado collection") is a corpus of emails and attachments, distributed for use in research and development in e-discovery, social network analysis, and related fields. Please carefully read and adhere to the usage agreements before beginning work on the collection. All users must personally sign the Avocado Collection End User Agreement before working with the collection, and the collection must be stored on a password protected computer in a way prevents access to the collection by anyone who has not personally signed the Acocado Collection End User Agreement.
There are a total of 2,033,740 items in the collection, broken up into the following categories:

<table>
<thead>
<tr>
<th>Type</th>
<th>Total</th>
<th>Redact</th>
<th>Dup</th>
<th>Final</th>
<th>No Text</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>938035</td>
<td>0</td>
<td>323574</td>
<td>614461</td>
<td>0</td>
<td>614461</td>
</tr>
<tr>
<td>Attachment</td>
<td>325506</td>
<td>110023</td>
<td>126411</td>
<td>89072</td>
<td>23731</td>
<td>65341</td>
</tr>
<tr>
<td>Extracted file</td>
<td>556286</td>
<td>298022</td>
<td>172870</td>
<td>85394</td>
<td>14775</td>
<td>70619</td>
</tr>
<tr>
<td>Contact</td>
<td>59455</td>
<td>0</td>
<td>0</td>
<td>59455</td>
<td>59455</td>
<td>0</td>
</tr>
<tr>
<td>Appointment</td>
<td>76902</td>
<td>0</td>
<td>0</td>
<td>76902</td>
<td>0</td>
<td>76902</td>
</tr>
<tr>
<td>Task</td>
<td>15474</td>
<td>0</td>
<td>0</td>
<td>15474</td>
<td>0</td>
<td>15474</td>
</tr>
<tr>
<td>Schedule</td>
<td>26980</td>
<td>0</td>
<td>0</td>
<td>26980</td>
<td>0</td>
<td>26980</td>
</tr>
<tr>
<td>Report</td>
<td>7206</td>
<td>0</td>
<td>0</td>
<td>7206</td>
<td>7206</td>
<td>0</td>
</tr>
<tr>
<td>Journal</td>
<td>4204</td>
<td>4204</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Stickynote</td>
<td>2232</td>
<td>2232</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>21460</td>
<td>21460</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Total: 2033740  435941  622855  974944  105167  869777
### Sample Collections: Brobeck

#### [Document Categories]

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Count</th>
<th>Percent Off Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Processor [.DOC, .WPF]</td>
<td>3,195,672</td>
<td>57%</td>
</tr>
<tr>
<td>Image Format [.TIF, .GIF]</td>
<td>831,289</td>
<td>15%</td>
</tr>
<tr>
<td>System Configuration Files [.DLL, .INI, ...]</td>
<td>215,274</td>
<td>5%</td>
</tr>
<tr>
<td>Email Message [.EML]</td>
<td>178,503</td>
<td>3%</td>
</tr>
<tr>
<td>Spreadsheet [.XLS]</td>
<td>168,206</td>
<td>3%</td>
</tr>
<tr>
<td>Text Format [.TXT, .RTF]</td>
<td>160,438</td>
<td>3%</td>
</tr>
<tr>
<td>Database Files [.SQL,.MDB, ...]</td>
<td>65,709</td>
<td>1%</td>
</tr>
<tr>
<td>Error Log Files [.BAD]</td>
<td>66,380</td>
<td>1%</td>
</tr>
<tr>
<td>Web Page [.HTM]</td>
<td>22,343</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>Portable Document [.PDF]</td>
<td>22,287</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>Data file [.DAT]</td>
<td>21,614</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>Miscellaneous [.LNK,.PST,.ZIP...]</td>
<td>-</td>
<td>Approx. 10 -11%</td>
</tr>
</tbody>
</table>
### SAMPLE COLLECTIONS: BROBECK
### DESCRIPTIVE STATISTICS OF TOP 100 EXTRACTED .PST FILES
[ DOCUMENT CATEGORIES ]

<table>
<thead>
<tr>
<th>DOCUMENT TYPE</th>
<th>COUNT</th>
<th>PERCENT OFF TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Messages [No Extension]</td>
<td>935234</td>
<td>45%</td>
</tr>
<tr>
<td>Text Document [.TXT, .RTF]</td>
<td>859769</td>
<td>42%</td>
</tr>
<tr>
<td>Word Processor [.DOC, .WPF]</td>
<td>191725</td>
<td>9%</td>
</tr>
<tr>
<td>Portable Document [.PDF]</td>
<td>35345</td>
<td>2%</td>
</tr>
<tr>
<td>Spreadsheet [.XLS]</td>
<td>28634</td>
<td>1%</td>
</tr>
<tr>
<td>Image Format [.JPG, .GIF, .TIF, …]</td>
<td>12580</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>Web Page [.HTM]</td>
<td>3895</td>
<td>&lt;&lt; 1%</td>
</tr>
<tr>
<td>Power Point Presentation [.PPT]</td>
<td>3690</td>
<td>&lt;&lt; 1%</td>
</tr>
<tr>
<td>Electronic Business Card [.VCF]</td>
<td>3590</td>
<td>&lt;&lt; 1%</td>
</tr>
<tr>
<td>Link File iManage Format [.NRL]</td>
<td>2314</td>
<td>&lt;&lt; 1%</td>
</tr>
<tr>
<td>Compressed file [.ZIP]</td>
<td>1252</td>
<td>&lt;&lt; 1%</td>
</tr>
</tbody>
</table>
**Ongoing & Next Steps**

- **Ongoing**
  - Talking to LDC & Sherwood Partners about allowing commercial, as well as non-commercial, access to released collection
  - Talking to Sherwood Partners & Hagley Library (NDP partner) to identify additional email collections

- **Next**
  - Talking to LoC to develop rules and procedures for Brobeck collection
  - Workshop to identify 3rd parties that can “pool” email collections?
Ongoing & Next Steps

- Thank you!
- Questions / comments welcome
- dkirsch@umd.edu
- @darchivist
Backup Slides
Question: What happens to venture backed firms that do not have successful exits?

Answer: Assignment for the Benefit of Creditors or “Workout”—may be thought of as a private bankruptcy.

Board of Directors replaces management with senior executive from specialized consulting firm that oversees liquidation.

- Outsource non-core VC activity
- Specialized team increases value recovered
- Allows investors to maintain control (v. bankruptcy)
One of two Silicon Valley firms that service venture capital market.
- Market share not known at present

Handled several hundred workouts since late-1999
- Mostly IT though some biotech and medical device companies.

Seven year statute of limitations for records retention expiring and new failures arriving…
- Sherwood needs to get rid of some of 7,000+ boxes
Venture Capital's Liquidators

As startups go bust, it's boom times for firms that restructure and sell off tech companies' remains—everything from furniture to patents.

"Really, we're a sophisticated mortuary that monetizes assets," says Fichinson.

By Robert D. Hof

Silicon Valley is famous for its awesome startup machine, but the place is just as brutally efficient at winding down those companies when things turn south. So as the

How Risky Is India?
The Archive Warehouse
September
2008