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

BUSINESS PLAN ARCHIVE

Brobeck
Closed Archive

ENRON

- How will we learn about what really goes on inside modern firms?

- Surviving firms unlikely to grant researchers access to large, un-vetted 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)
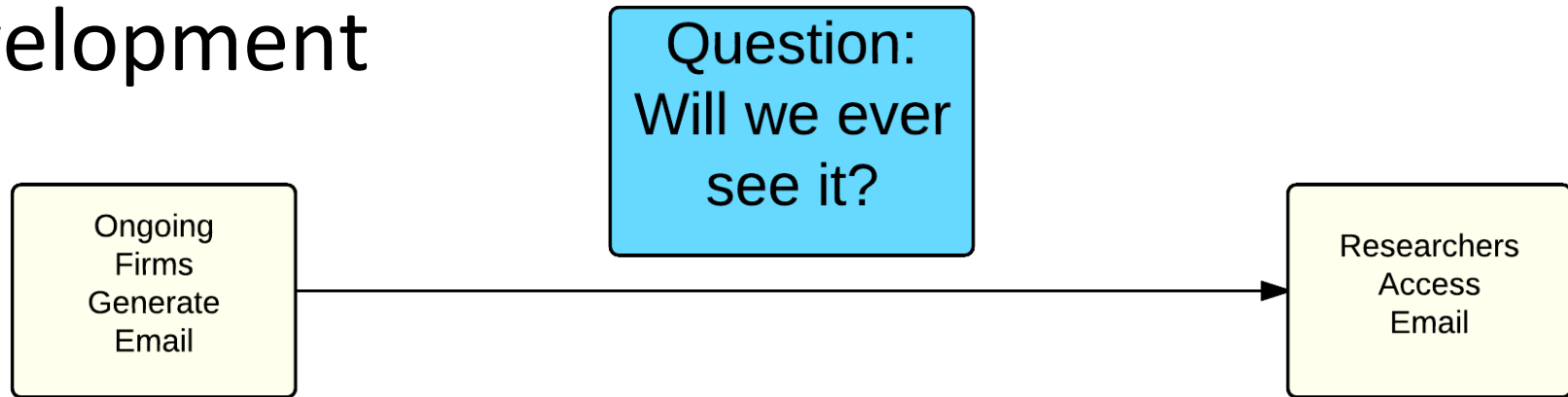 - LoC / NDIIPP (DABODCE)
  - NSF (eDiscovery)
- **Scholars**
  - RRCHNM at GMU
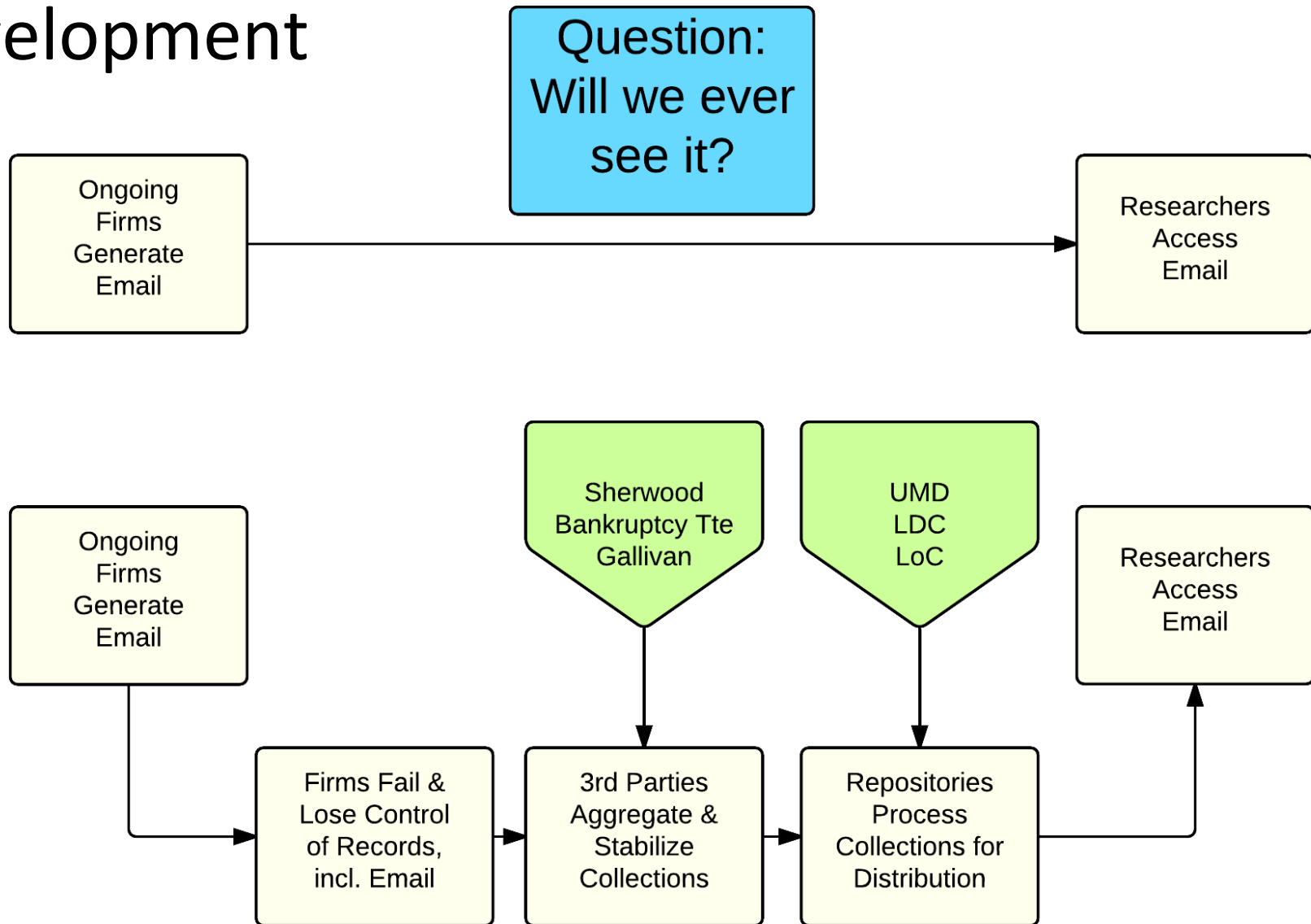 - UMD i-School / UMIACS
  - Advisory Council
- **Repositories**
  - Hagley Museum & Library (NDP)
  - Linguistic Data Consortium at UPenn
  - LC
  - Others?

Collection Development

Ongoing
Firms
Generate
Email

Question:
Will we ever
see it?

Researchers
Access
Email

Collection Development

Question:
Will we ever
see it?

Ongoing
Firms
Generate
Email

Researchers
Access
Email

Ongoing
Firms
Generate
Email

Sherwood
Bankruptcy Tte
Gallivan

UMD
LDC
LoC

Researchers
Access
Email

Firms Fail &
Lose Control
of Records,
incl. Email

3rd Parties
Aggregate &
Stabilize
Collections

Repositories
Process
Collections for
Distribution

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

Sample Collections: AvocadoIT

Avocado Research Email Collection

Item Name:	Avocado Research Email Collection
Author(s):	Douglas Oard, William Webber, David Kirsch, Sergey Golitsynskiy
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

Sample Collections: AvocadoIT


```
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.
```

SAMPLE COLLECTIONS: AVOCADOIT
[COLLECTION STATISTICS]

There are a total of 2,033,740 items in the collection, broken up into the following categories:

Type	Total	Redact	Dup	Final	No Text	Text
Email	938035	0	323574	614461	0	614461
Attachment	325506	110023	126411	89072	23731	65341
Extracted file	556286	298022	172870	85394	14775	70619
Contact	59455	0	0	59455	59455	0
Appointment	76902	0	0	76902	0	76902
Task	15474	0	0	15474	0	15474
Schedule	26980	0	0	26980	0	26980
Report	7206	0	0	7206	7206	0
Journal	4204	4204	0	0	0	0
Stickynote	2232	2232	0	0	0	0
Other	21460	21460	0	0	0	0
Total	2033740	435941	622855	974944	105167	869777

SAMPLE COLLECTIONS: BROBECK
[DOCUMENT CATOGORIES]

DOCUMENT TYPE	COUNT	PERCENT OFF TOTAL
Word Processor [.DOC, .WPF]	3195672	57%
Image Format [.TIF, .GIF]	831289	15%
System Configuration Files [.DLL, .INI, ...]	215274	5%
Email Message [.EML]	178503	3%
Spreadsheet [.XLS]	168206	3%
Text Format [.TXT, .RTF]	160438	3%
Database Files [.SQL,.MDB, ...]	65709	1%
Error Log Files [.BAD]	66380	1%
Web Page [.HTM]	22343	< 1%
Portable Document [.PDF]	22287	< 1%
Data file [.DAT]	21614	< 1%
Miscellaneous [.LNK,.PST,.ZIP...]	-	Approx. 10 -11%

SAMPLE COLLECTIONS: BROBECK
DESCRIPTIVE STATISTICS OF TOP 100 EXTRACTED .PST FILES
[DOCUMENT CATOGORIES]

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

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

Venture Workouts

- **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)

Sample – Sherwood Partners

- 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

Business Week:

BusinessWeek

HOME   INVESTING   COMPANIES   TECHNOLOGY   INNOVATION   MANAGING   SMALL BIZ   B-SCHOOLS

Current Issue   Past Issues   Cover Story Podcasts   Figures of the Week   Small Biz Magazine   BW TV   Subscribe   FA

SILICON VALLEY December 4, 2008, 5:00PM EST   text size: T | T

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 Pichinson Gabriela Hasbun

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

BW MAGAZINE

How Risky Is India?

The ~~Archive~~ Warehouse

IP DEMO
DO NOT DISTURB

CURLY

CB2

POWER 2

CB1

PWR OK
FAULT
TEMP
I LIM

PWR OK
FAULT
TEMP
I LIM

MOE

IF
Equipment

LARRY

September 2008