Pitfall! Working with Legacy Born Digital Materials in Special Collections

Don Mennerich and Mark A. Matienzo

Code4Lib 2013, Chicago, Illinois

Level 1. Disk Imaging

Image Credit: S. Soosay

Understanding Disk Images

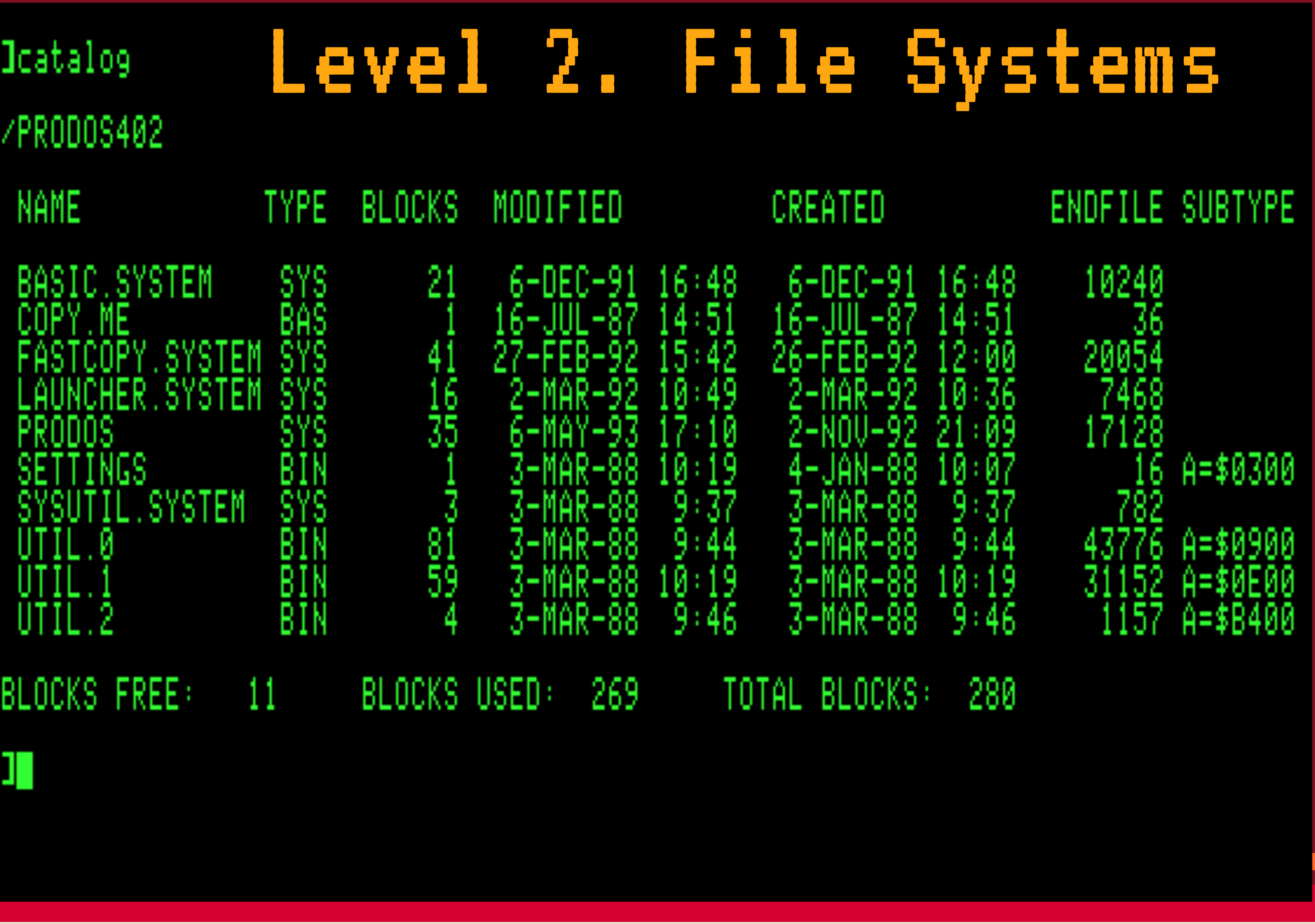
- What process are you using to image?
  - "Stream": digitized analog magnetic signal
  - "Sector": stream decoded using algorithm(s)

- What is the object you're trying to acquire?
  - "Physical": entirety of device
  - "Logical": some subset; volume/set of files

Pitfalls

- Disk image formats mean different things
- Communities of practice (forensics vs. retrocomputing) use different kinds of container formats
- There is no single solution: depends on what your workflow is/what tools you use

Level 2. File Systems



```
]catalog

/PRODOS402

 NAME           TYPE  BLOCKS  MODIFIED         CREATED          ENDFILE SUBTYPE

 BASIC.SYSTEM    SYS      21   6-DEC-91 16:48   6-DEC-91 16:48    10240
 COPY.ME         BAS       1  16-JUL-87 14:51  16-JUL-87 14:51       36
 FASTCOPY.SYSTEM SYS      41  27-FEB-92 15:42  26-FEB-92 12:00    20054
 LAUNCHER.SYSTEM SYS      16   2-MAR-92 10:49   2-MAR-92 10:36     7468
 PRODOS          SYS      35   6-MAY-93 17:10   2-NOV-92 21:09    17128
 SETTINGS        BIN       1   3-MAR-88 10:19   4-JAN-88 10:07       16 A=$0300
 SYSUTIL.SYSTEM  SYS       3   3-MAR-88  9:37   3-MAR-88  9:37      782
 UTIL.0          BIN      81   3-MAR-88  9:44   3-MAR-88  9:44    43776 A=$0900
 UTIL.1          BIN      59   3-MAR-88 10:19   3-MAR-88 10:19    31152 A=$0E00
 UTIL.2          BIN       4   3-MAR-88  9:46   3-MAR-88  9:46     1157 A=$B400

BLOCKS FREE:   11     BLOCKS USED:  269     TOTAL BLOCKS:  280

]
```

File System/Volume Formats

Windows

FAT (12,16,32)

NTFS

Linux

EXT3

Mac

HFS extended

Unix

UFS

File Systems supported in TSK

ntfs

iso9660

hfs+

fat12

fat16

fat32

ext2

ext3

ufs1

ufs2

The Sleuthkit can generate DFXML +1

```
  <fileobject>
      <filename>ACCESS</filename>
      <partition>1</partition>
      <id>3</id>
      <name_type>r</name_type>
      <filesize>1829</filesize>
      <alloc>1</alloc>
      <inode>5</inode>
      <mtime>1990-12-07T20:17:50Z</mtime>
      <byte_runs>
       <byte_run file_offset='0' fs_offset='17408'
img_offset='17408' len='1829'/>
      </byte_runs>
      <hashdigest
type='md5'>f79b7ab9b0b41794b34afd3a83479688</hashdigest>
```

```
<!-- plugin_process -->
      <pronomMatch>true</pronomMatch>
      <pronomPuid>x-fmt/393</pronomPuid>
      <pronomMimeType />
      <pronomFormat>WordPerfect for MS-DOS
Document</pronomFormat>
      <pronomFormatVersion>5.0</pronomFormatVersion>
      <pronomIdentificationMethod>binary
signature</pronomIdentificationMethod>
      <virusFound>false</virusFound>
```

But it does not support Apple HFS Volumes -1

File systems supported in Forensic Toolkit

FAT 12

FAT 16

FAT 32

NTFS

Ext2

HFS

HFS+

Ext3

CDFS

Ext4FS

exFAT

ReiserFS

VxFS

UFS1

UFS2

And many optical formats

Forensic Toolkit cannot generate DFXML -1

But it supports Apple HFS Volumes +1

But What About ...?

Apple/Macintosh

ProDOS

MFS - Macintosh file system

Amiga

OFS - old file system

FFS - fast file system

PFS - professional files system

Commodore, CP/M, Solaris ZFS, BeOS BFS...

Pitfalls

- Significant changes to Apple´s file system format has generally made preservation more difficult than windows file systems for legacy collections

- The more ´exotic´ the file system, the more difficult to integrate into a workflow

Level 3. Files

```
.AAC     .ACE     .ALZ     .APK     .AT3     .ARC     .ARJ     .BIG     .BIK
    .CAD     .cgr     .DRW     .DWG     .DFT     .DGN     .DGK     .DMT     .DXF
.BKF     .BMP     .BLD     .CAB     .DAA     .DEB     .DMG     .DDZ     .DPE
    .DWB     .DWF     .EMB     .ESW     .EXP     .FMZ     .GLM     .GRB     .GTC
.EEA     .EGT     .ESS     .GHO     .IPG     .JAR     .LBR     .LQR     .LHA
    .IAM     .ICD     .IDW     .IFC     .IPN     .IPT     .MCD     .OCD     .PAR
.LZO     .LZX     .MPQ     .NTH     .PAK     .RAR     .RAG     .RPM     .SEN
    .PLN     .PRT     .PSM     .PWI     .PYT     .SKP     .RLF     .RVT     .RFA
.SKB     .TAR     .TIB     .UHA     .VIV     .VOL     .VSA     .WAX     .ZOO
    .STL     .TCT     .TCW     .UNV     .VC6     .VLM     .WRL     .BRD     .CDL
.ZIP     .ISO     .NRG     .IMG     .ADF     .ADZ     .DMS     .DSK     .D64
    .CPF     .DEF     .HEX     .LEF     .LIB     .SDC     .SDF     .UPF     .VCD
.SDI     .MDS     .MDX     .DMG     .CDI     .CUE     .CIF     .C2D     .DAA
    .WGL     .ADT     .APR     .BOX     .DAF     .DAT     .DAT     .DBF     .EGT
.B6T     .ACP     .AMF     .ART     .ASC     .ASM     .CCC     .CCM     .CCS
```

File Identification

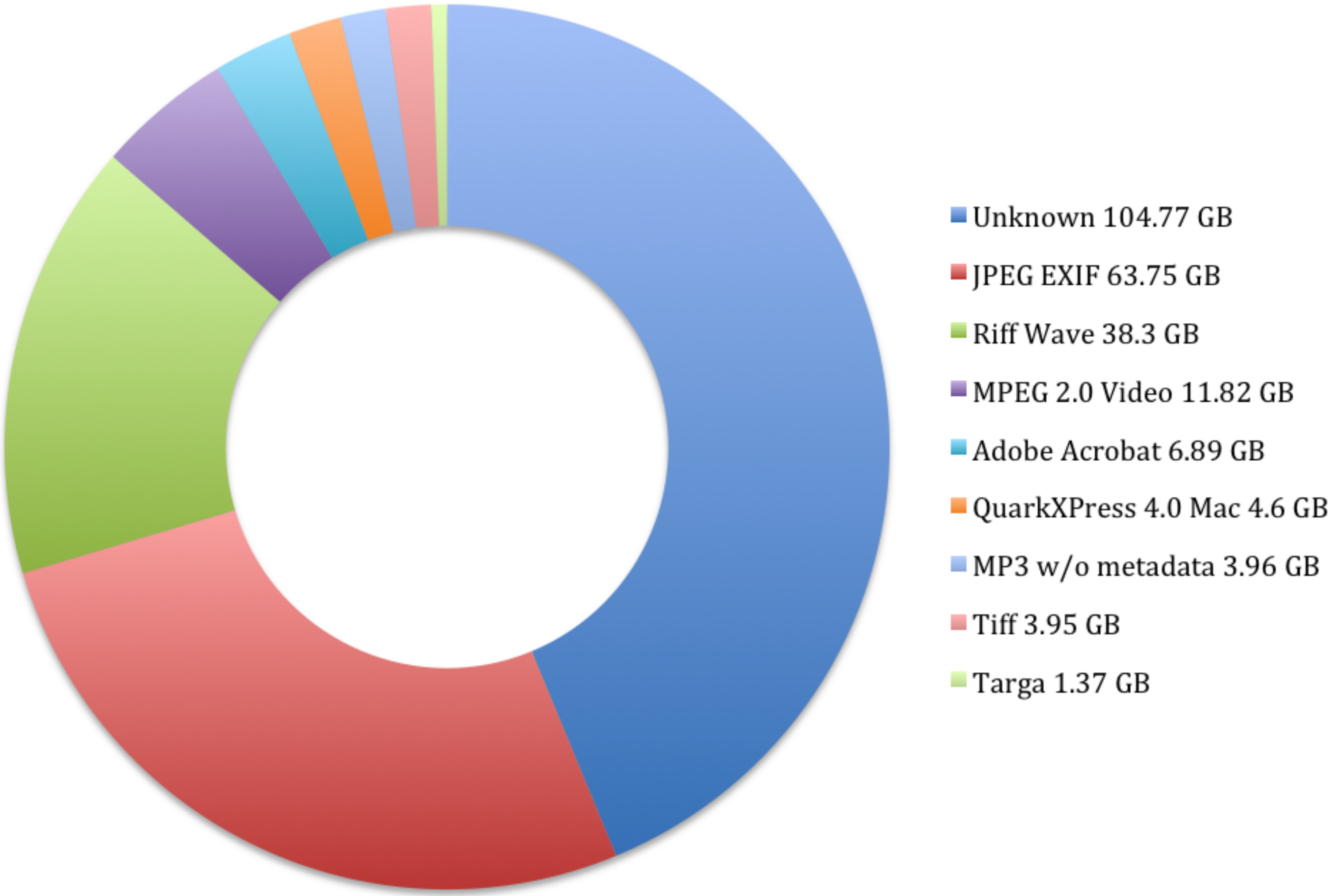
- Characterize collection to better understand a collection for management and planning
- Automate tasks based on format type

PRONOM / FIDO

```
FIDO v1.1.1
OK
524
x-fmt/263,
"ZIP Format",
"ZIP format",
18143146,
"/Users/dm/DMDM.pptx",
"application/zip",
"signature"
```

OTHER FILE IDENTIFICATION TOOLS and Resources

- Unix ‘file’ command
- Apache Tika (tika.apache.org)
- Forensics Toolkit / FTK Imager (accessdata.com)
- Unified Digital Format Registry (udfr.org)
- Archives Team ‘Just Solve the Problem’ wiki http://fileformats.archiveteam.org/wiki


Unknown 104.77 GB
JPEG EXIF 63.75 GB
Riff Wave 38.3 GB
MPEG 2.0 Video 11.82 GB
Adobe Acrobat 6.89 GB
QuarkXPress 4.0 Mac 4.6 GB
MP3 w/o metadata 3.96 GB
Tiff 3.95 GB
Targa 1.37 GB

Pitfalls

- No format registry can be all encompassing
- Matches by extension can often be misleading

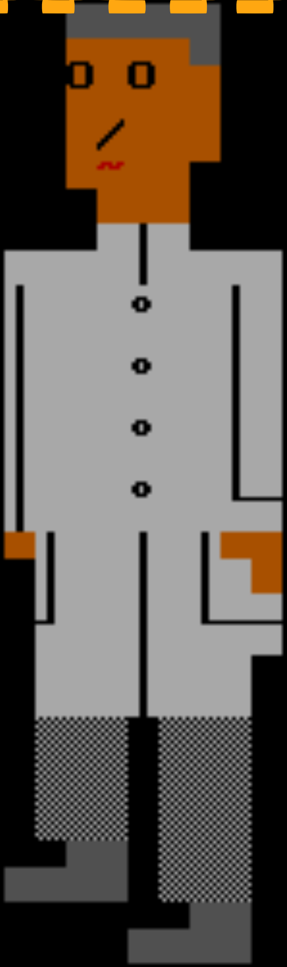
  OS9 Mac .doc vs. Windows 2003-2007 .doc vs. WordPerfect .doc

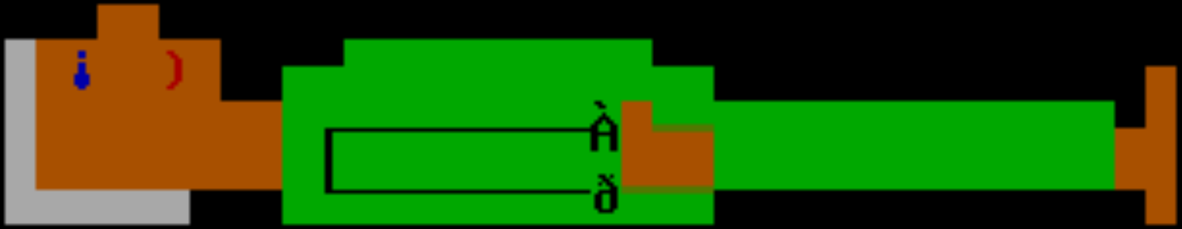
- Chronological 'arrangement' can be difficult due to inconsistent 'date' metadata

Level 4.

The Quest for Access

One moment...I'm now reviewing your decisions.





Access Decisions

Format

Disk image
Original files (or media)
Migrated version
Emulated version/system

User Location

In Person / Reading room
Remote access

Permissions for Interaction

Discover via metadata
Discover via content
View
Download
Use on researcher machine
Manipulate / Edit
Text / data mining, etc.

Pitfalls

- There is no ideal single model, even when backed by policy
- Decisions throughout the life cycle have an impact on access
- Capacities of your institution
- Levels of researcher support: what do you expect them to know, do, etc.?

Bonus Round 1. Mark

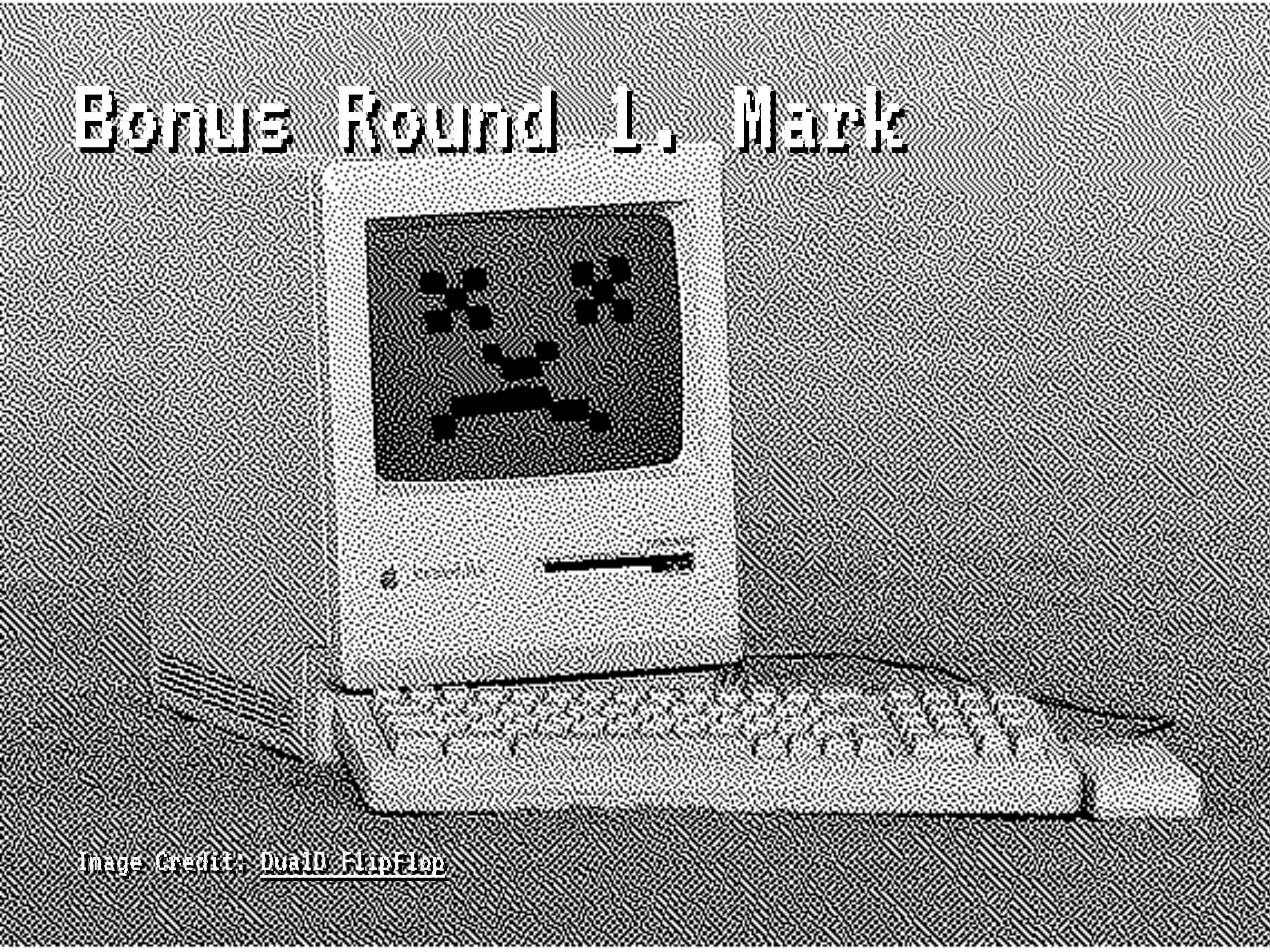


Image Credit: DualD FlipFlop

The Collection: Faculty Papers

- 162 floppies / 35 linear feet
- 5.25" and 3.5" PC disks; 3.5" Mac/HFS disks
- 14 disks were in box labeled "Backup 12/30/94"
- Strong assumption that backup was important as creator died shortly before this date
- Little info about creator's tech environment

The Goal:

- "Recover" backup into something usable by other tools (FTK, emulators, etc.)
- Make minimal changes to files within backup, or their metadata (especially timestamps)
- Document process so it could hypothetically be repeatable

Phase I. Preparation

- Imaged floppies using CatWeasel (PCI floppy controller; now no longer produced/supported)
- Because of no HFS support in Sleuth Kit, used FTK Imager to analyze images
- Found single SEA file on each backup disk; extracted files using FTK Imager

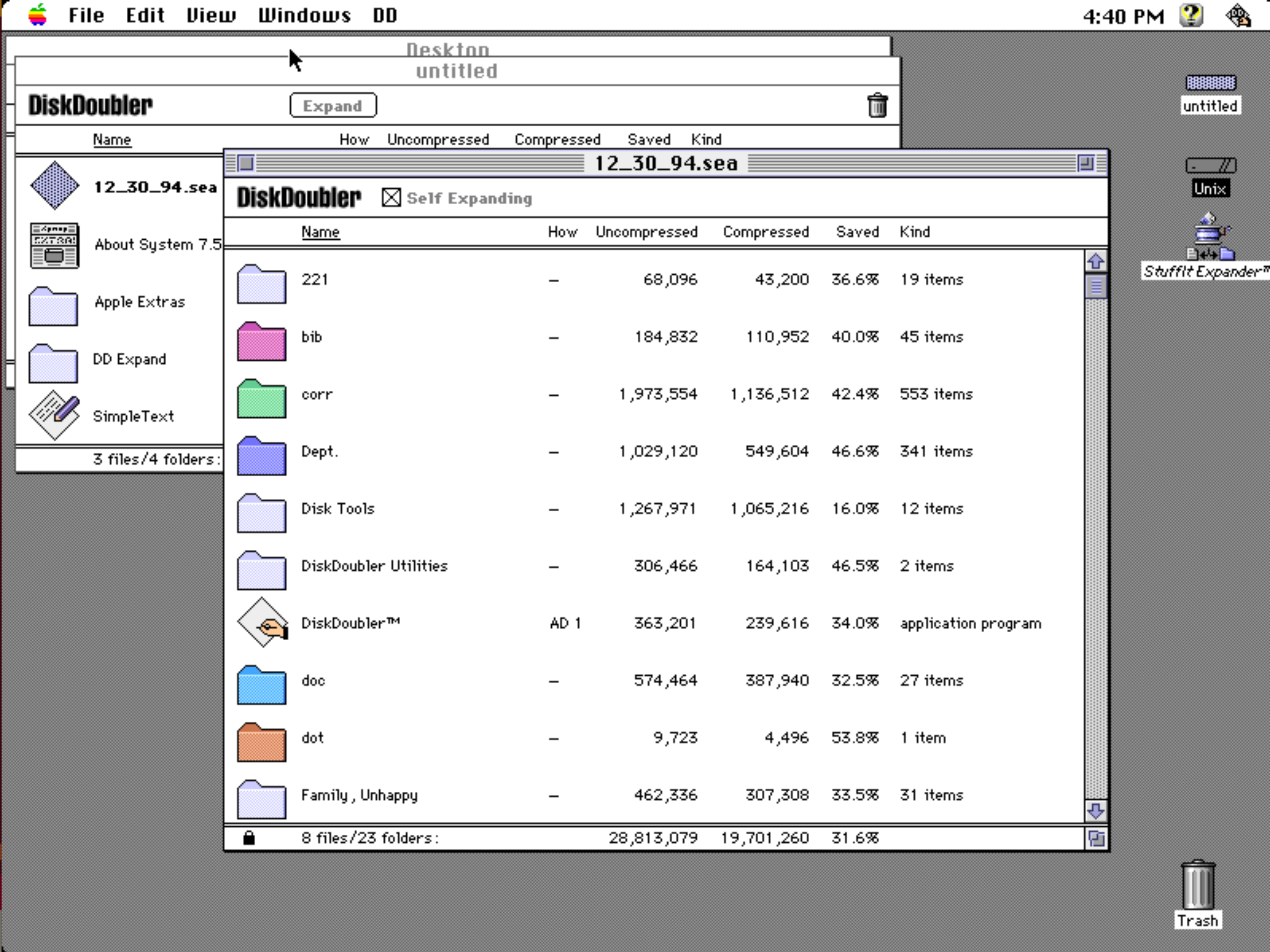
Phase II. Reconstruction

- FTK Imager extracted data forks only

- SEA files: multi-part self-extracting archive

- Used emulated environment to assign type and creator codes and create single SEA with DiskDoubler (lots of trial and error)

- Expanded SEA to an empty 30 MB HFS image to be loaded into Forensic Toolkit for processing



untitled

DiskDoubler

Name	How	Uncompressed	Compressed	Saved	Kind
12_30_94.sea					
About System 7.5					
Apple Extras					
DD Expand					
SimpleText					
3 files/4 folders:					

12_30_94.sea

DiskDoubler ☒ Self Expanding

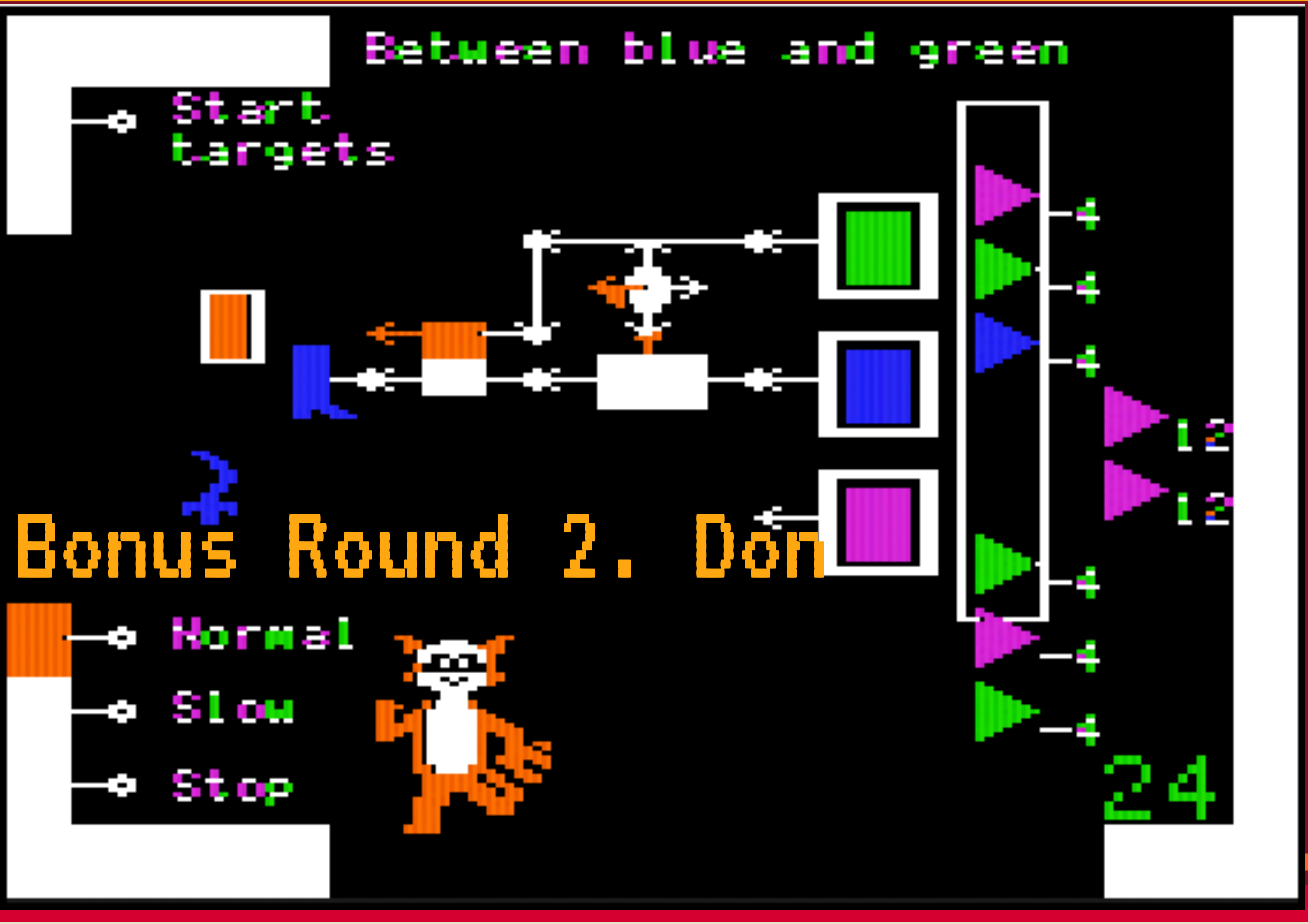
Name	How	Uncompressed	Compressed	Saved	Kind
221	–	68,096	43,200	36.6%	19 items
bib	–	184,832	110,952	40.0%	45 items
corr	–	1,973,554	1,136,512	42.4%	553 items
Dept.	–	1,029,120	549,604	46.6%	341 items
Disk Tools	–	1,267,971	1,065,216	16.0%	12 items
DiskDoubler Utilities	–	306,466	164,103	46.5%	2 items
DiskDoubler™	AD 1	363,201	239,616	34.0%	application program
doc	–	574,464	387,940	32.5%	27 items
dot	–	9,723	4,496	53.8%	1 item
Family, Unhappy	–	462,336	307,308	33.5%	31 items
8 files/23 folders:		28,813,079	19,701,260	31.6%	

Phase III. Documentation

- No good format for documentation!

- Ultimately just added a note in a bag-info.txt file in BagIt packages for disk images:

> Extracted from 2011-M-034.0078 to 2011-M-034.0091. Original files placed in 01-original_sea_files directory. Files comprise multipart DiskDoubler SEA file. Transferred to Basilisk II emulator and joined using DiskDoubler 4.2 after setting proper type/creator code (DDSP/DDAP). The resulting file was of file type/creator APPL/DSE2 and is located in the 02_intermediary_file directory. DiskDoubler was used to expand the files into a blank HFS disk image. The resulting disk image is located in 03_derived_disk_image.



Between blue and green

Bonus Round 2. Don

The Collection:
The Vito Russo Papers

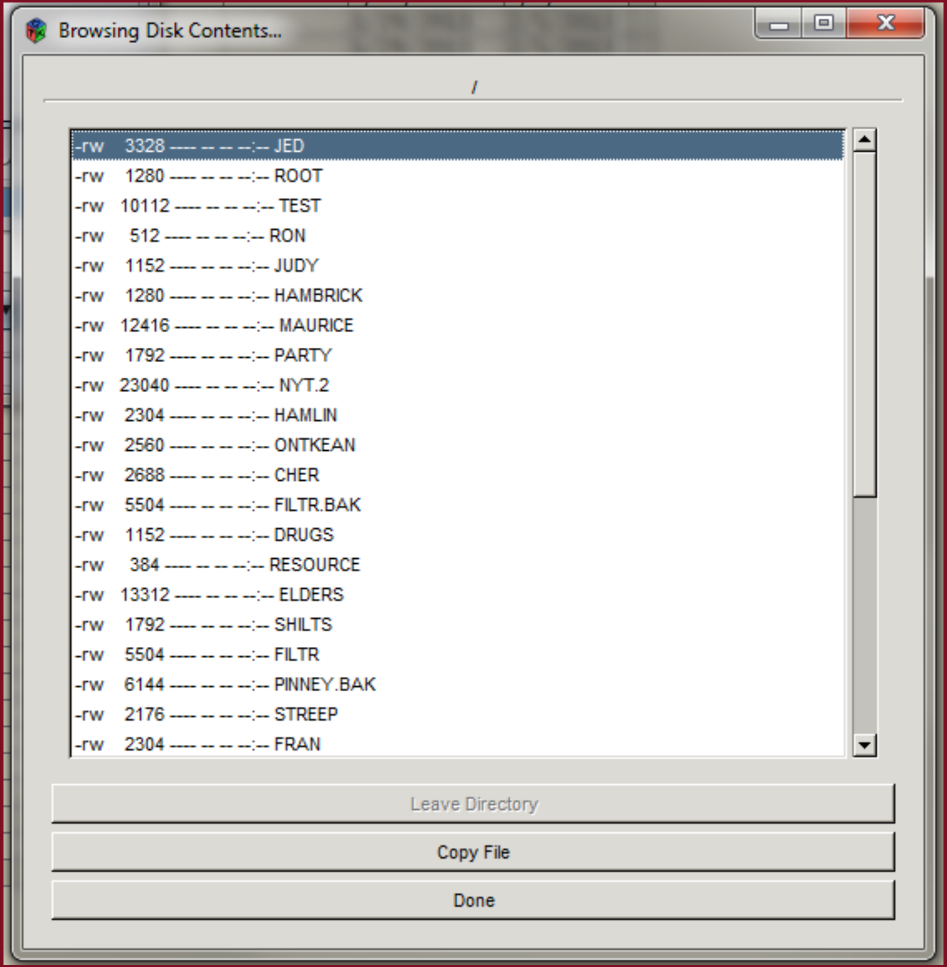
- 18 5.25" Kaypro IV disks
- Wordstar word processing documents dated from the mid 1980s to the early 1990s
- Hard copies printed on a borrowed Kaypro IV, sometime in the 1990s
- 2 of the 18 disks were marked as being ‘unreadable’

KAYPRO
POWER

The Goal:

- Preserve originals
- Describe and arrange files
- Provide access to files or content in the files
- If possible, migrate to format supported by Tika to create a fulltext index

Phase I. Imaging



```
-rw   3328 ---- -- -- --:-- JED
-rw   1280 ---- -- -- --:-- ROOT
-rw  10112 ---- -- -- --:-- TEST
-rw    512 ---- -- -- --:-- RON
-rw   1152 ---- -- -- --:-- JUDY
-rw   1280 ---- -- -- --:-- HAMBRICK
-rw  12416 ---- -- -- --:-- MAURICE
-rw   1792 ---- -- -- --:-- PARTY
-rw  23040 ---- -- -- --:-- NYT.2
-rw   2304 ---- -- -- --:-- HAMLIN
-rw   2560 ---- -- -- --:-- ONTKEAN
-rw   2688 ---- -- -- --:-- CHER
-rw   5504 ---- -- -- --:-- FILTR.BAK
-rw   1152 ---- -- -- --:-- DRUGS
-rw    384 ---- -- -- --:-- RESOURCE
-rw  13312 ---- -- -- --:-- ELDERS
-rw   1792 ---- -- -- --:-- SHILTS
-rw   5504 ---- -- -- --:-- FILTR
-rw   6144 ---- -- -- --:-- PINNEY.BAK
-rw   2176 ---- -- -- --:-- STREEP
-rw   2304 ---- -- -- --:-- FRAN
```

Phase II. File System

- No support in TSK or FTK
- Command line binaries: CPMTools
- No disk defs for KayPro disks for CPMTools
- Need for automation

Phase III. Files

- FTK supports WordStar format, but has no facility to convert
- WordPerfect X6 has the ability to convert but not batch convert
- Corel would write a WordPerfect batch processor but would not support converting to .docx or .odt

1. Create new logical images of each disk in FTK for analysis and description

2. Migrated access copies created through a two-step conversion using Corel batch converter and MS Word macro

3. Full text indexing of files with Apache Tika and entity extraction with Apache NLP

4. Access in reading room to originals using Quickview Pro

Query: Russo

Results

Number of records located: 10

collection: The Vito Russo Papers

id: M2654.0009.0001

filename: abstract

filetype: Wordstar 4.0

filesize: 22337

modification date:

language: en

names:

Janet, Vito Russo, Vincent Price, Sue Lyons, Anne Bancroft, Vesta Tilley, Fred McMurray, Edward G. Robinson "Dancing, Young Woman, Eddie Cantor, Alan Mowbray

organizations:

MGM, United Artists, CAESAR, Marines, Professional SissyFranklin Pangborn, MOMAMAEDCHEN, STRANGELOVEBONNIE, Marches Post Stonewall WORD

Locations:

United States, WASHINGTON

Conclusions (1)

- Legacy materials can be extremely time consuming to manage or 'process' archivally
- Technological problems for legacy materials can require significant resources to solve and may never occur again within the collections of a repository

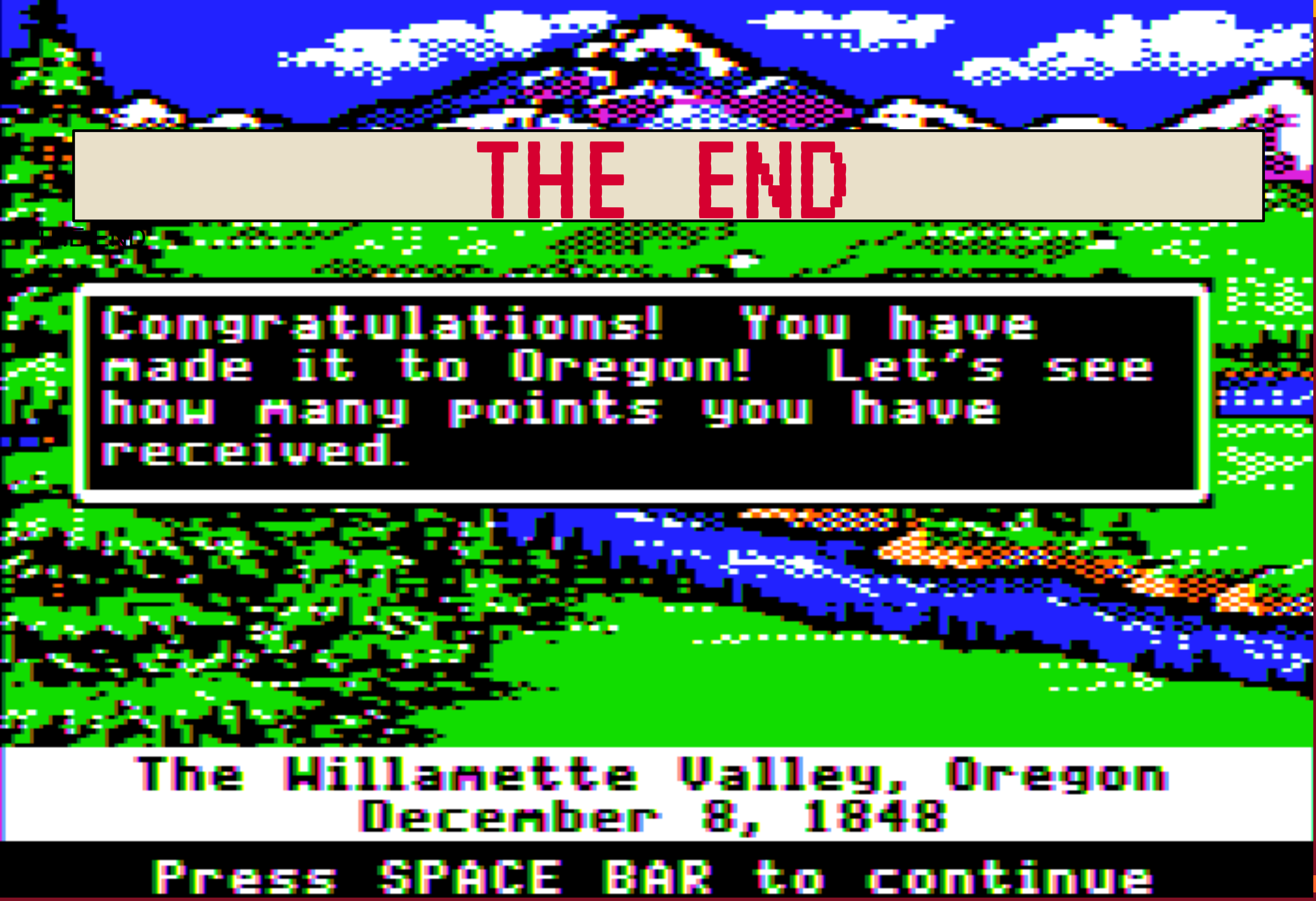
Conclusions (2)

- Acknowledge to researchers about limits on what resources we can provide for access and what their responsibilities are

- Our community of practice would be better served by common practice for documentation and better tools for knowledge sharing

THE END

THE END



Congratulations! You have made it to Oregon! Let's see how many points you have received.

The Willamette Valley, Oregon
December 8, 1848

Press SPACE BAR to continue

@anarchivist @mennerich