

*In case you want to follow along...*

---

**Activity 1:** <http://bit.ly/1ctsgpF>

**Activity 2:** <http://bit.ly/1GStpi3>

**Activity 3:** <http://bit.ly/1AJw8If>

---

**Bertram Lyons**  
**Jason Evans Groth**  
*MAC 2015*  
*Lexington, Kentucky*

*Not Everything Digital is a Disk Image*

---

# Using Lightweight Tools to Assess and Profile Digital Collections of Files

**Bertram Lyons**  
**Jason Evans Groth**  
*MAC 2015*  
*Lexington, Kentucky*

---

---

# Part 1. Making files and metadata work for you

---

Using a computer's operating system to quantify file sizes; perform minimal format analysis; and rename, move, and count files.

Using an open-source tool (MDQC) that uses file header metadata to allow you to check the validity of files and compare files against set specifications.

```
yx@ 1 bertramlyons staff 25539229 Sep 28 2012 Bunny W
yx@ 1 bertramlyons staff 16545779 Sep 20 2012 Everybo
yx@ 1 bertramlyons staff 15439927 Sep 28 2012 Fair an
yx@ 1 bertramlyons staff 16405764 Sep 20 2012 Find th
yx@ 1 bertramlyons staff 17079777 Sep 20 2012 Highway
yx@ 1 bertramlyons staff 12145746 Sep 28 2012 I Can't
yx@ 1 bertramlyons staff 5284682 Sep 28 2012 I'm Sob
yx@ 1 bertramlyons staff 10727055 Sep 28 2012 It's Al
yx@ 1 bertramlyons staff 14153132 Sep 20 2012 My Ador
yx@ 1 bertramlyons staff 25873008 Sep 20 2012 The Dre
yx@ 1 bertramlyons staff 22666718 Sep 28 2012 Tinselt
yx@ 1 bertramlyons staff 25334989 Sep 28 2012 Tinselt
yx@ 1 bertramlyons staff 16910466 Sep 20 2012 To the
yx@ 1 bertramlyons staff 6146127 Sep 28 2012 Wishing

Elephant Micah:
?
yx@ 1 bertramlyons staff 32768 Sep 28 2012 Elephant M
yx 1 bertramlyons staff 32768 Sep 28 2012 Elephant M
yx 1 bertramlyons staff 32768 Sep 28 2012 Hindu Wind

Elephant Micah/Elephant Micah - Louder Than Thou:
648
yx 1 bertramlyons staff 14273031 Oct 8 2011 01 Tin
yx 1 bertramlyons staff 17792248 Oct 8 2011 02 Won
yx 1 bertramlyons staff 10742321 Oct 8 2011 03 My C
yx 1 bertramlyons staff 13416215 Oct 8 2011 04 If I
yx 1 bertramlyons staff 16322076 Oct 8 2011 05 Roos
yx 1 bertramlyons staff 15256280 Oct 8 2011 06 Airl

Elephant Micah/Elephant Micah Plays The Songs Of Bible B
6416
yx@ 1 bertramlyons staff 10752148 Aug 30 2012 01 Card
yx@ 1 bertramlyons staff 10126264 Aug 30 2012 02 Free
yx@ 1 bertramlyons staff 4735629 Aug 30 2012 03 Loop
yx@ 1 bertramlyons staff 10537964 Aug 30 2012 04 The
yx@ 1 bertramlyons staff 5631113 Aug 30 2012 05 Imag
yx@ 1 bertramlyons staff 11913040 Aug 30 2012 06 Are
yx@ 1 bertramlyons staff 8466956 Aug 30 2012 07 Para
yx@ 1 bertramlyons staff 8446946 Aug 30 2012 08 Pa F
```

# files

- ❖ At a high level of abstraction, a digital file is a stored segment or block of information that is available to a computer program.



---

# metadata

---

how many times have you seen a slide with this header?

# what i am not talking about

## MARC

```
START  REGULAR  RETURN TO  ANOTHER
OVER   DISPLAY BROWSING  SEARCH  (Search History)
LEADER 00000cam 2200000 a 4500a
001    50906019
003    OCoLC
005    20030715092633.0
008    021023s2003    ilu      b    001 0 eng
010    2002151683
015    GBA3-Y7095
020    0838908470
040    DLC|cDLC|dUKM|dC#P|dXFF|dKSU|dOCoLC
049    KSUU
050 00 2666.5|b.C37 2003
082 00 025.3|221
100 1  Caplan, Priscilla
245 10 Metadata fundamentals for all librarians /|cPriscilla
      Caplan
260    Chicago :|bAmerican Library Association,|c2003
300    ix, 192 p. ;|c28 cm
504    Includes bibliographical references and index
505 00 |tMetadata basics --|tSyntax, creation, and storage --
      |tVocabularies, classification, and identifiers --
      |tApproaches to interoperability --|tMetadata and the Web
      --|tLibrary cataloging --|gThe|tTEI header --|gThe|tDublin
      Core --|tArchival description and the EAD --|tMetadata for
      art and architecture --|tGILS and government information -
      -|tMetadata for education --|tONIX International --
      |tMetadata for geospatial and environmental resources --
      |gThe|tData Documentation Initiative --|tAdministrative
      metadata --|tStructural metadata --|tRights metadata
650 0  Metadata
650 0  Information organization
```

# what i am not talking about

## EAD

```
EADrecord.xml x
1 <?xml version="1.0" encoding="UTF-8"?>
2 <ead xmlns="urn:isbn:1-931666-22-9"
3   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4   xsi:schemaLocation="urn:isbn:1-931666-22-9 http://www.loc.gov/ead/ead.xsd">
5   <eadheader audience="internal" countryencoding="iso3166-1" dateencoding="iso8601">
6     <eadid countrycode="US">Basham Kelly Papers, MS-F24</eadid>
7     <filedesc>
8       <titlestmt>
9         <titleproper>Guide to the Basham Kelly Papers, 1936-1988</titleproper>
10        <subtitle></subtitle>
11        <author>Processed by Judith Morgan; finding aid prepared by Diana Elizabeth</author>
12        <sponsor></sponsor>
13      </titlestmt>
14      <publicationstmt>
15        <date>1992</date>
16        <publisher>University Archives, Rodgers Library, Bluegrass State University</publisher>
17      </publicationstmt>
18      <notestmt>
19        <note>
20          <p>
21            <subject source="lcsh">English literature</subject>
22            <subject source="lcsh">College teachers</subject>
23            <subject source="lcsh">Archival resources</subject>
24          </p>
25        </note>
26      </notestmt>
27    </filedesc>
28    <profiledesc>
29      <creation>The collection was processed at the University Archives in 1992 by Judith Morgan. The finding aid was prepared by
30      <language>English</language>
31      <desrules>Find aid prepared using Describing Archives, a Content Standard</desrules>
32    </profiledesc>
33  </eadheader>
34  <frontmatter>
35    <titlepage>
36      <titleproper>Guide to the Basham Kelly Papers</titleproper>
37      <date>1936-1988</date>
38      <num type="Collection number">MS-F24</num>
```

---

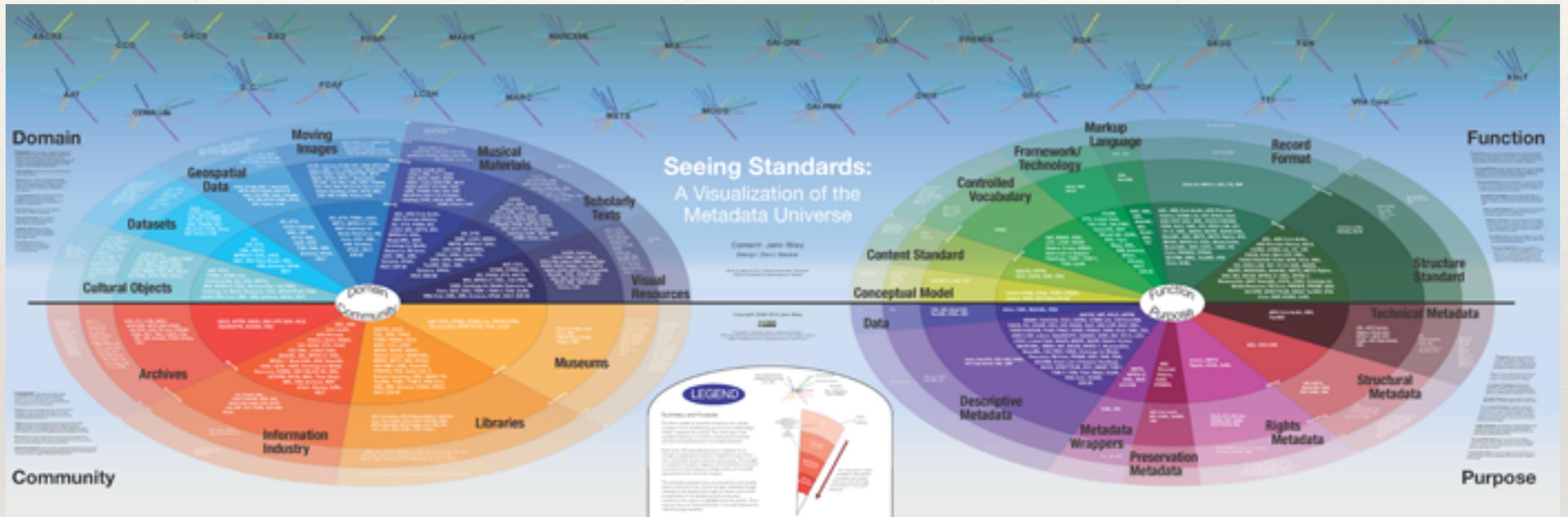
# what i am not talking about

---

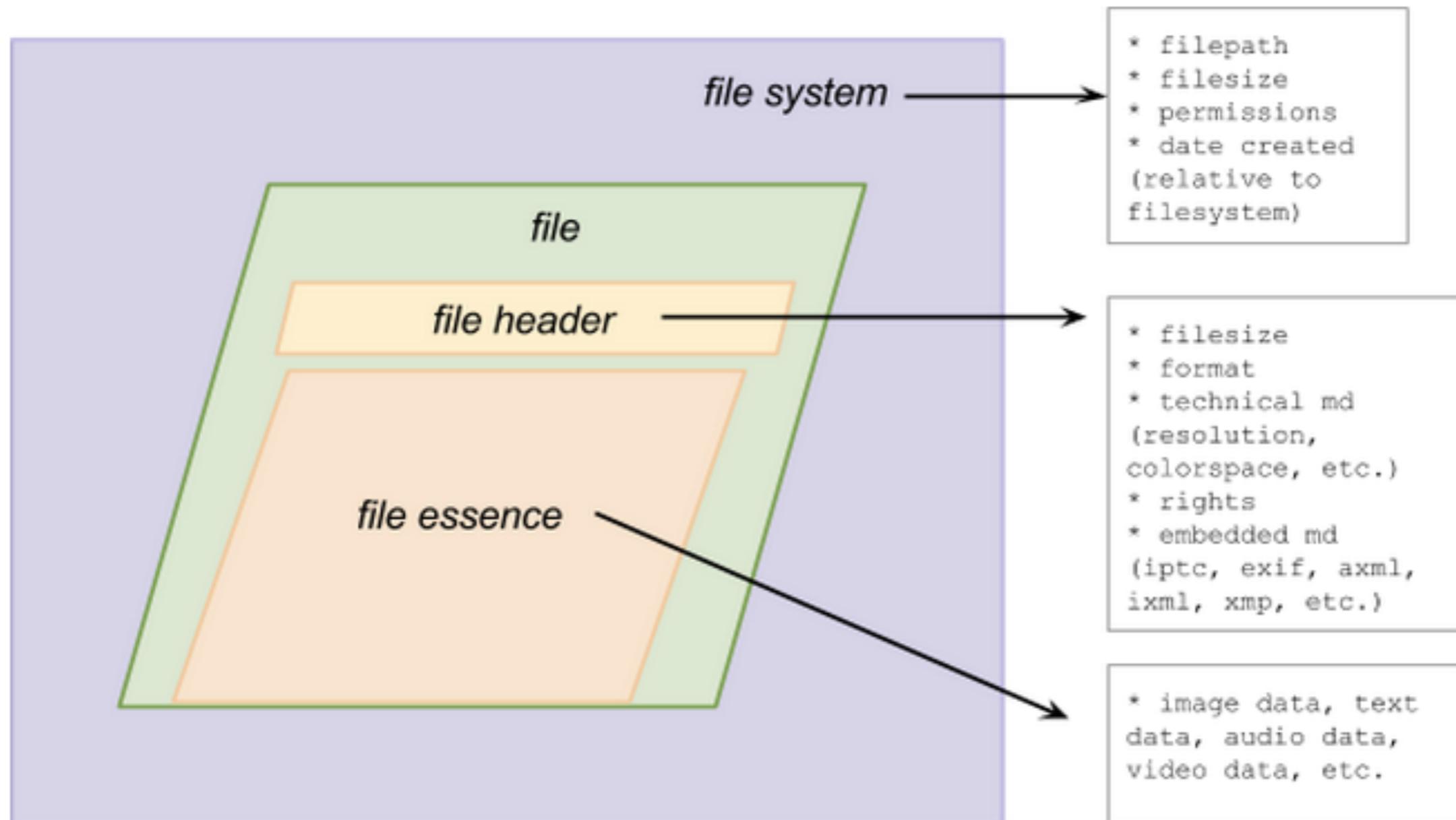
## Dublin Core

```
- <METS:dmdSec>
  - <METS:mdWrap>
    - <METS:xmlData>
      - <DC:dc>
        <DC:title>Bm:display_title</DC:title>
        <DC:description>Bm:abstract</DC:description>
        <DC:contributor>Cs:name</DC:contributor>
        <DC:coverage>B:url</DC:coverage>
        <DC:creator>D:display_name</DC:creator>
        <DC:date>B:date_added_to_library</DC:date>
        <DC:source>Cs:name</DC:source>
        <DC:rights>B:rights_holders</DC:rights>
        <DC:type>Bm:type</DC:type>
        <DC:subject>Bm:genre</DC:subject>
        <DC:publisher>Bm:publisher_name</DC:publisher>
        <DC:language>Bm:book_language</DC:language>
        <DC:identifier>B:id</DC:identifier>
        <DC:format>JPEG/TIFF/DJVU/PDF</DC:format>
        <DC:relation>N/A</DC:relation>
      </DC:dc>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
```

# what i am not talking about



# what i AM talking about



# file system metadata

*file system metadata* - all the information that is known about the file by the file system that is storing it at a given time

```
Bertrams-MacBook-Air:Ampatch bertramlyons$ ls -l mv0041.mp4  
-rwxrwxrwx  1 bertramlyons  staff  279432618 Nov  4  2014 mv0041.mp4
```

permissions

links

owner

group

file size (bytes)

date last modified

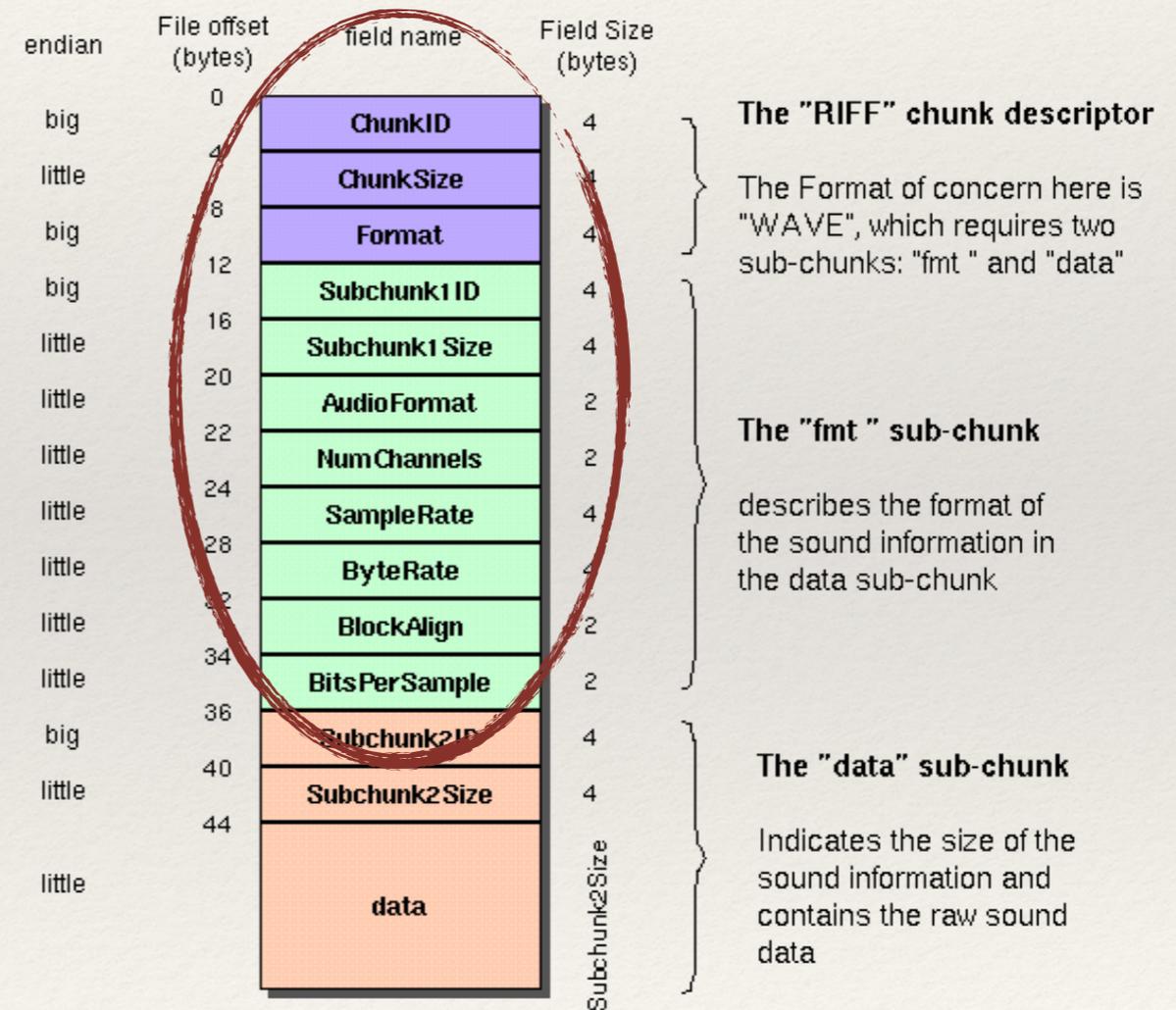
filename

file

# file header metadata

*file header metadata* - information contained within a file to help software interpret and decode the bits so that a human can understand them how they were intended to be understood.

## The Canonical WAVE file format

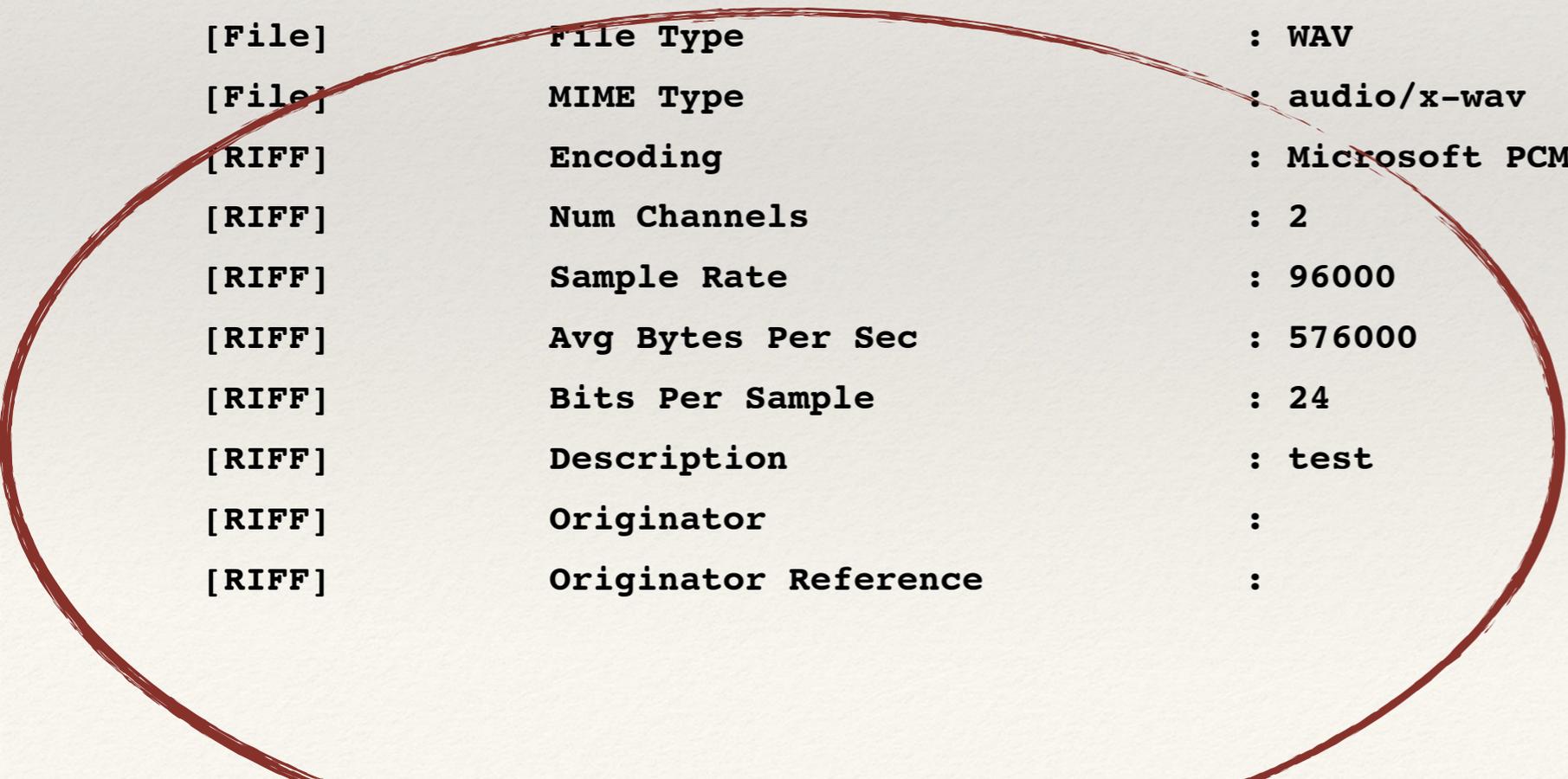


---

# file header metadata

---

```
bertramlyons$ exiftool -u -G test.wav
[ExifTool]      ExifTool Version Number      : 9.47
[File]          File Name                    : test.wav
[File]          Directory                  : .
[File]          File Size                 : 1719 kB
[File]          File Modification Date/Time : 2014:10:09 01:54:00-04:00
[File]          File Access Date/Time      : 2015:05:05 20:42:30-04:00
[File]          File Inode Change Date/Time : 2015:05:05 20:42:30-04:00
[File]          File Permissions          : rw-r-----
[File]          File Type                 : WAV
[File]          MIME Type                 : audio/x-wav
[RIFF]          Encoding                  : Microsoft PCM
[RIFF]          Num Channels                : 2
[RIFF]          Sample Rate                : 96000
[RIFF]          Avg Bytes Per Sec          : 576000
[RIFF]          Bits Per Sample            : 24
[RIFF]          Description                 : test
[RIFF]          Originator                  :
[RIFF]          Originator Reference       :
```



---

# file header metadata

---

```
bertramlyons$ tiffutil -info IMG_0565.tif
```

```
Directory at 0x12c106
```

```
Image Width: 640
```

```
Image Length: 480
```

```
Resolution: 72, 72
```

```
Resolution Unit: pixels/inch
```

```
Bits/Sample: 8
```

```
Sample Format: unsigned integer
```

```
Compression Scheme: none
```

```
Photometric Interpretation: RGB color
```

```
Alpha: Present
```

```
Date & Time: "2010:10:03 12:37:19"
```

```
Host Computer: "Mac OS X 10.6.2"
```

```
Software: "QuickTime 7.6.3"
```

```
Make: "Canon"
```

```
Model: "Canon PowerShot S90"
```

```
Orientation: row 0 top, col 0 lhs
```

```
Samples/Pixel: 4
```

---

# file essence (or content)

---

*file essence* - the information contained within the file that represents the encoded content that the file was intended to transmit.

\* textual data (for xml, txt, doc, excel, json, pdf, etc.)

\* image data (for jpegs, tiffs, dng, nef, dpx, etc.)

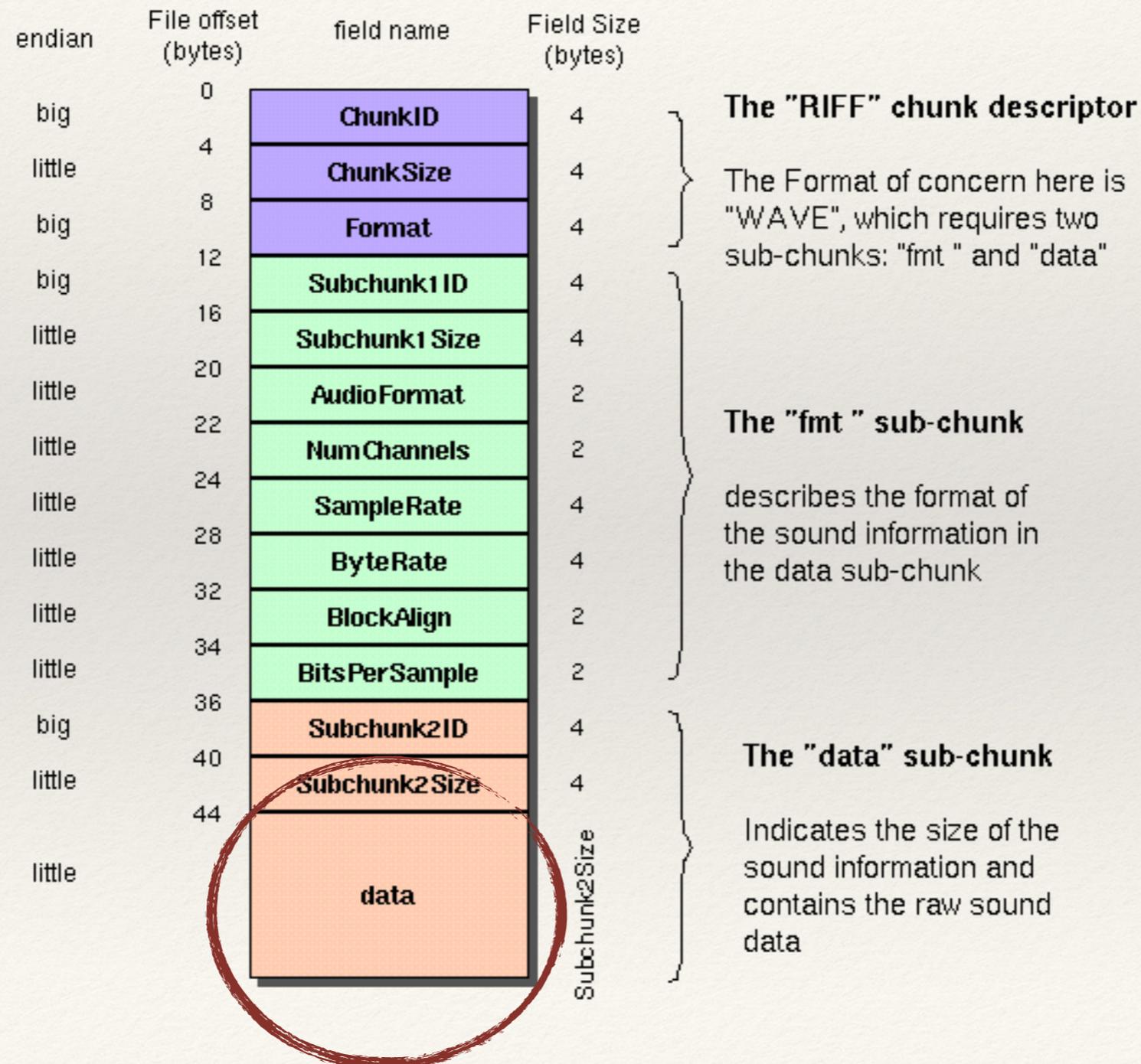
\* video data (for mov, mp4, mxf, avi, etc.).

\* audio data (for wav, mp3, aiff, etc.)

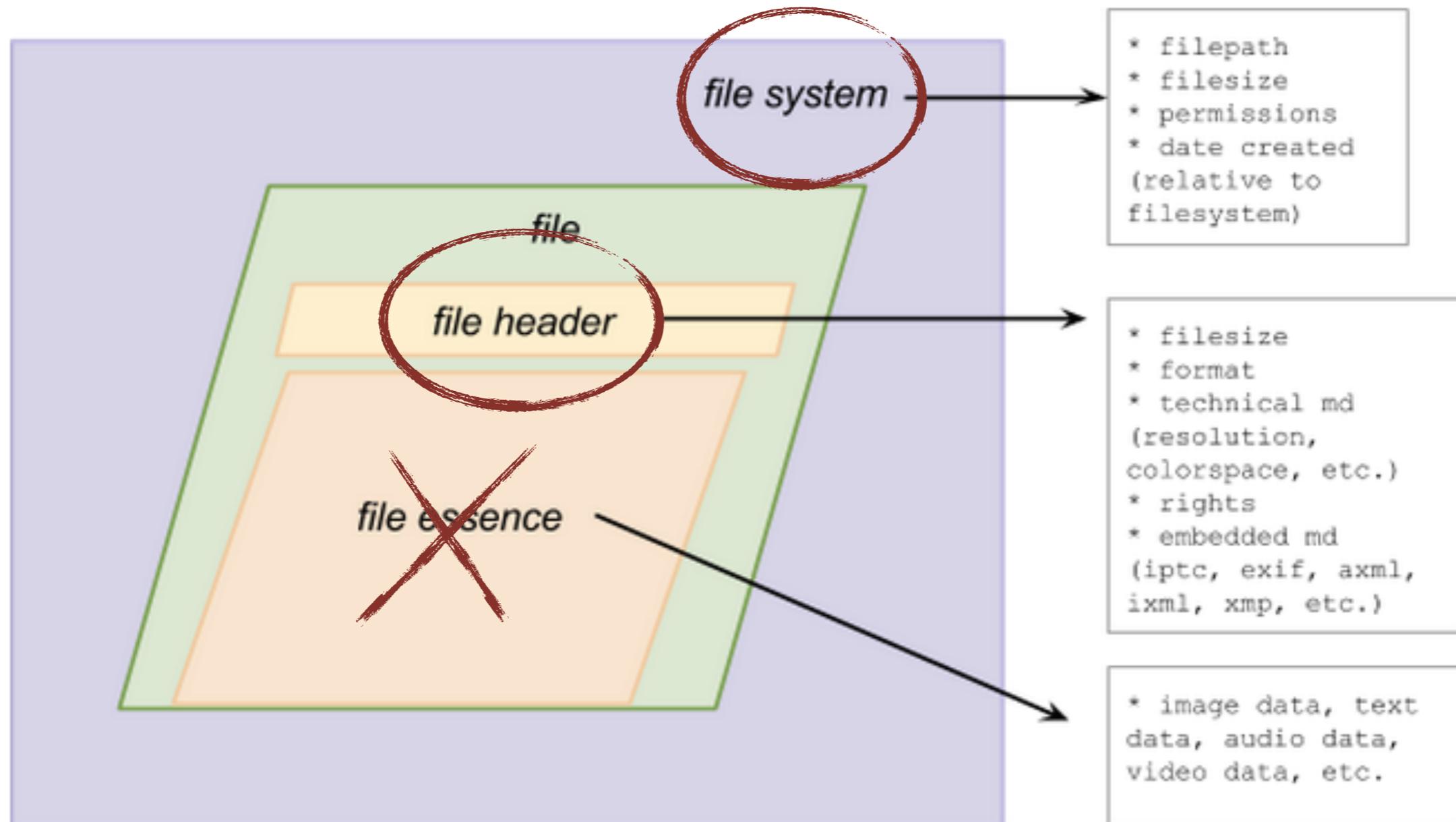
\* etc.

# file essence (or content)

## *The Canonical WAVE file format*



# what i AM talking about



afc2001001\_0282971781001.jpg  
afc2001001\_0294131413001.jpg  
afc2001001\_0294131438001.jpg  
afc2001001\_0294140201001.jpg  
afc2001001\_0431529552001.jpg  
afc2001001\_0431529558001.jpg  
afc2001001\_0431534307001.jpg  
afc2001001\_0431534389001.jpg  
afc2001001\_0431538505001.jpg  
afc2001001\_0431538530001.jpg  
afc2001001\_0482631003001.jpg  
afc2001001\_0510373967001.jpg  
afc2001001\_0510373988001.jpg  
afc2001001\_0583914867001.jpg  
afc2001001\_0623135478001.jpg  
afc2001001\_0623155589001.jpg  
afc2001001\_0682948697001.jpg  
afc2001001\_0682953914001.jpg  
afc2001001\_0754150972001.jpg  
afc2001001\_0754707521001.jpg  
afc2001001\_0754707558001.jpg  
afc2001001\_0754709503001.jpg  
afc2001001\_DSC02863.jpg  
afc2001001\_DSC02927.jpg  
afc2001001\_DSC02954.jpg  
afc2001001\_DSC02959.jpg  
afc2001001\_DSC03242.jpg  
afc2001001\_DSC03243.jpg

3

**Rename in Bulk**

```
wx@ 1 bertramlyons staff 32768 Sep 28 2012 Elepha  
wx 1 bertramlyons staff 32768 Sep 28 2012 Elepha  
wx 1 bertramlyons staff 32768 Sep 28 2012 Hindu V  
  
Elephant Micah/Elephant Micah - Louder Than Thou:  
1648  
wx 1 bertramlyons staff 14273031 Oct 8 2011 01  
wx 1 bertramlyons staff 17792248 Oct 8 2011 02  
wx 1 bertramlyons staff 10742321 Oct 8 2011 03  
wx 1 bertramlyons staff 13416215 Oct 8 2011 04  
wx 1 bertramlyons staff 11322075 Oct 8 2011 05  
wx 1 bertramlyons staff 15256280 Oct 8 2011 06  
  
Elephant Micah/Elephant Micah Plays The Songs Of Bib  
6416  
wx@ 1 bertramlyons staff 10752148 Aug 30 2012 01  
wx@ 1 bertramlyons staff 10126264 Aug 30 2012 02  
wx@ 1 bertramlyons staff 4735629 Aug 30 2012 03  
wx@ 1 bertramlyons staff 10537964 Aug 30 2012 04  
wx@ 1 bertramlyons staff 5631113 Aug 30 2012 05  
wx@ 1 bertramlyons staff 11913040 Aug 30 2012 06  
wx@ 1 bertramlyons staff 8466956 Aug 30 2012 07  
wx@ 1 bertramlyons staff 9446046 Aug 30 2012 08
```

1

**Gather Quick Intelligence**

Rule Generation

Set Rules

File Permissions	[Ignore Tag]	:	rxr-xr-x	+
File Size	Is At Least	:	50000	+
File Size	Is At Most	:	100000	+
File Type	Is	:	PNG	+
Filter	[Ignore Tag]	:	Adaptive	+
Gamma	[Ignore Tag]	:	2	+
Green X	[Ignore Tag]	:	0.3	+
Green Y	[Ignore Tag]	:	0.6	+
Image Height	Is Greater Than	:	100	+
Image Size	[Ignore Tag]	:	100x100	+

2

**Perform Quality Control**

various and unpredictable

1 command line and excel

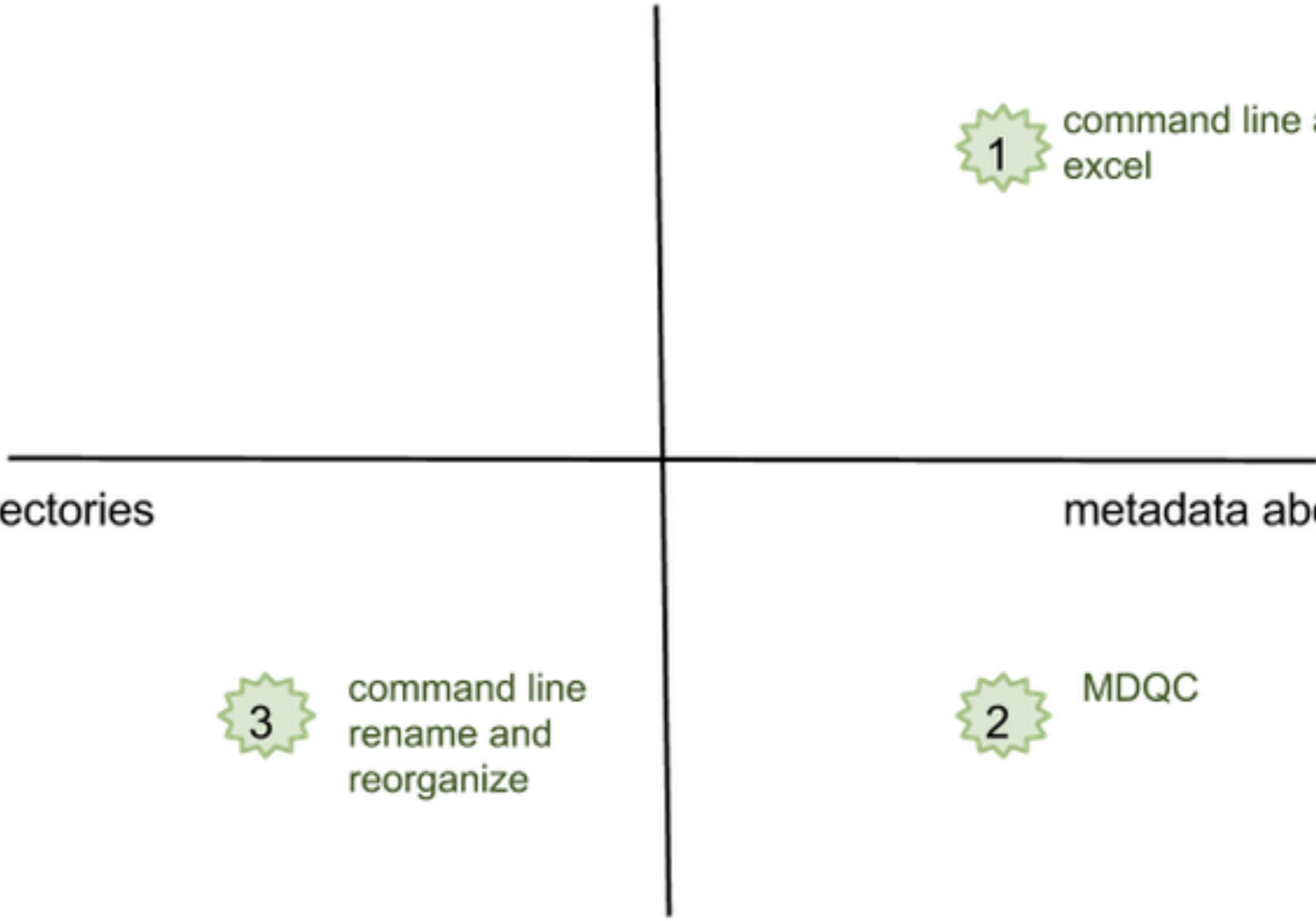
files/directories

metadata about files

3 command line rename and reorganize

2 MDQC

homogeneous and predictable



# Activity 1

*Various and Unpredictable  
Aggregation of Files*

Using file system metadata  
to establish a preliminary profile

## Goals of this method

1. Quick browsable manifest of the contents of the drive
2. Quantity of files
3. Total size of delivery in bytes
4. List of formats (generally)
  1. Quantity per format
  2. Total bytes per format type
5. List of file names

<http://bit.ly/1ctsgpF>

```
wx@ 1 bertramlyons staff 32768 Sep 28 2012 Elepha
wx 1 bertramlyons staff 32768 Sep 28 2012 Elepha
wx 1 bertramlyons staff 32768 Sep 28 2012 Hindu

Elephant Micah/Elephant Micah - Louder Than Thou:
1648
wx 1 bertramlyons staff 14273031 Oct 8 2011 01
wx 1 bertramlyons staff 17792248 Oct 8 2011 02
wx 1 bertramlyons staff 10742321 Oct 8 2011 03
wx 1 bertramlyons staff 13416213 Oct 8 2011 04
wx 1 bertramlyons staff 15322075 Oct 8 2011 05
wx 1 bertramlyons staff 15256280 Oct 8 2011 06

Elephant Micah/Elephant Micah Plays The Songs Of Bib
6416
wx@ 1 bertramlyons staff 10752148 Aug 30 2012 01
wx@ 1 bertramlyons staff 10126264 Aug 30 2012 02
wx@ 1 bertramlyons staff 4735629 Aug 30 2012 03
wx@ 1 bertramlyons staff 10537964 Aug 30 2012 04
wx@ 1 bertramlyons staff 5631113 Aug 30 2012 05
wx@ 1 bertramlyons staff 11913040 Aug 30 2012 06
wx@ 1 bertramlyons staff 8466956 Aug 30 2012 07
wx@ 1 bertramlyons staff 8446046 Aug 30 2012 08
```

**1 Gather Quick Intelligence**

## Activity 2

*Homogeneous and Predictable  
Aggregation of Files*

Using file header metadata  
to perform quality control

Goals of this method

1. Do all files in the batch meet expectations?

<http://bit.ly/1GStpi3>



# Activity 3

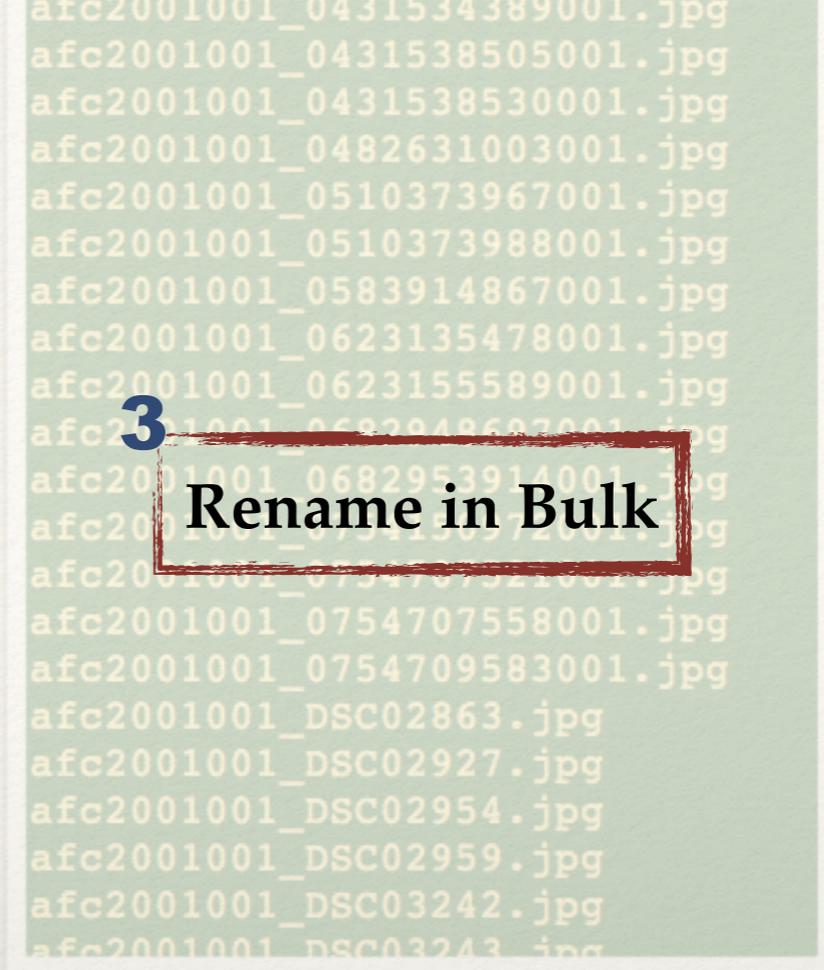
*Homogeneous and Predictable  
Aggregation of Files*

Using file system metadata  
to rename files

Goals of this method

1. Use logic and/or substitution to rename files in bulk

<http://bit.ly/1AJw8If>



*Thank you.*

---

**Activity 1: <http://bit.ly/1ctsgpF>**

**Activity 2: <http://bit.ly/1GStpi3>**

**Activity 3: <http://bit.ly/1AJw8If>**

---

**Bertram Lyons, CA**

**AVPreserve**

**[bertram@avpreserve.com](mailto:bertram@avpreserve.com)**