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

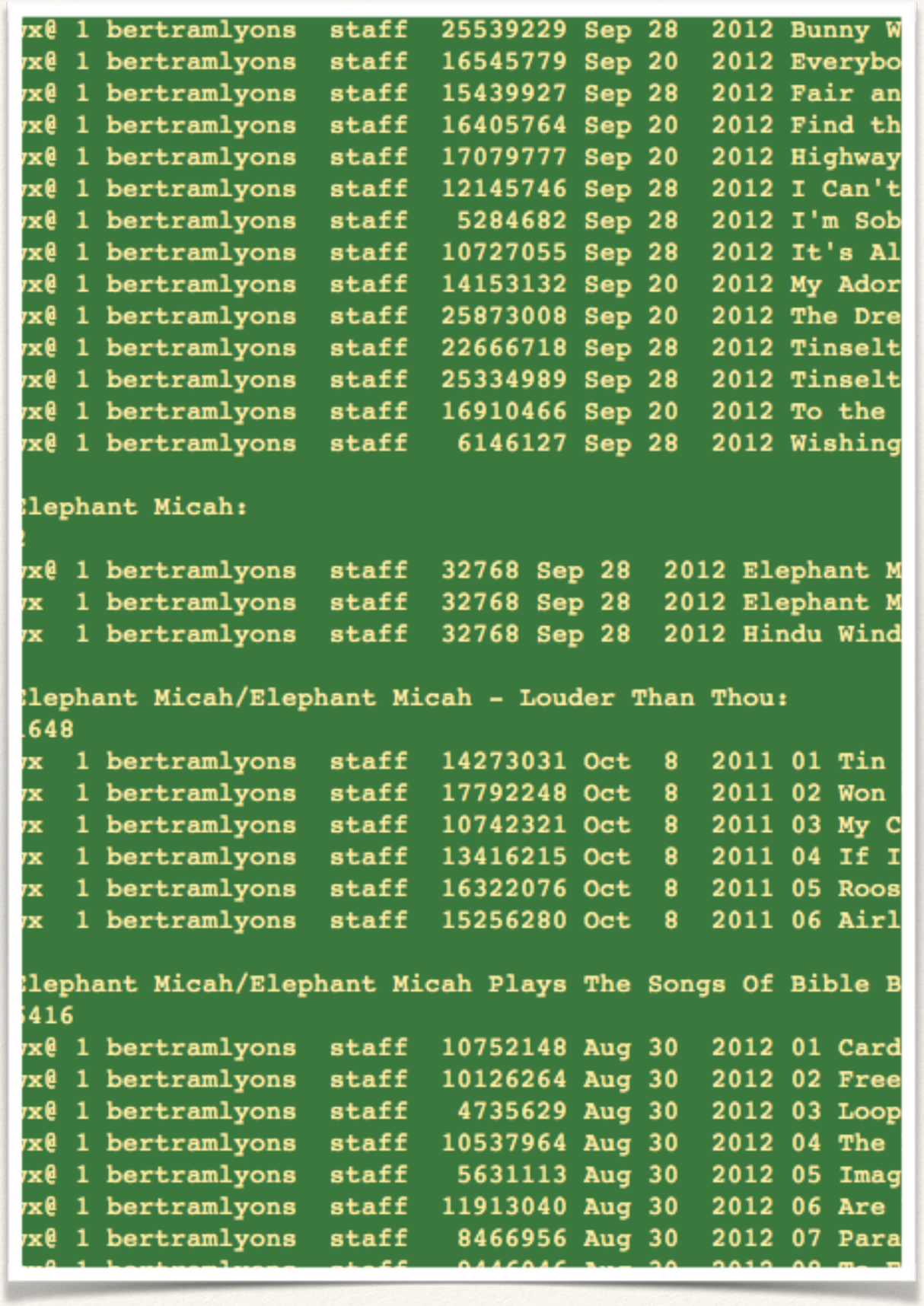
Using Lightweight Tools to Assess and Profile Digital Collections of Files

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.



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 OVER | REGULAR DISPLAY | RETURN TO BROWSING | ANOTHER 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 Z666.5|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

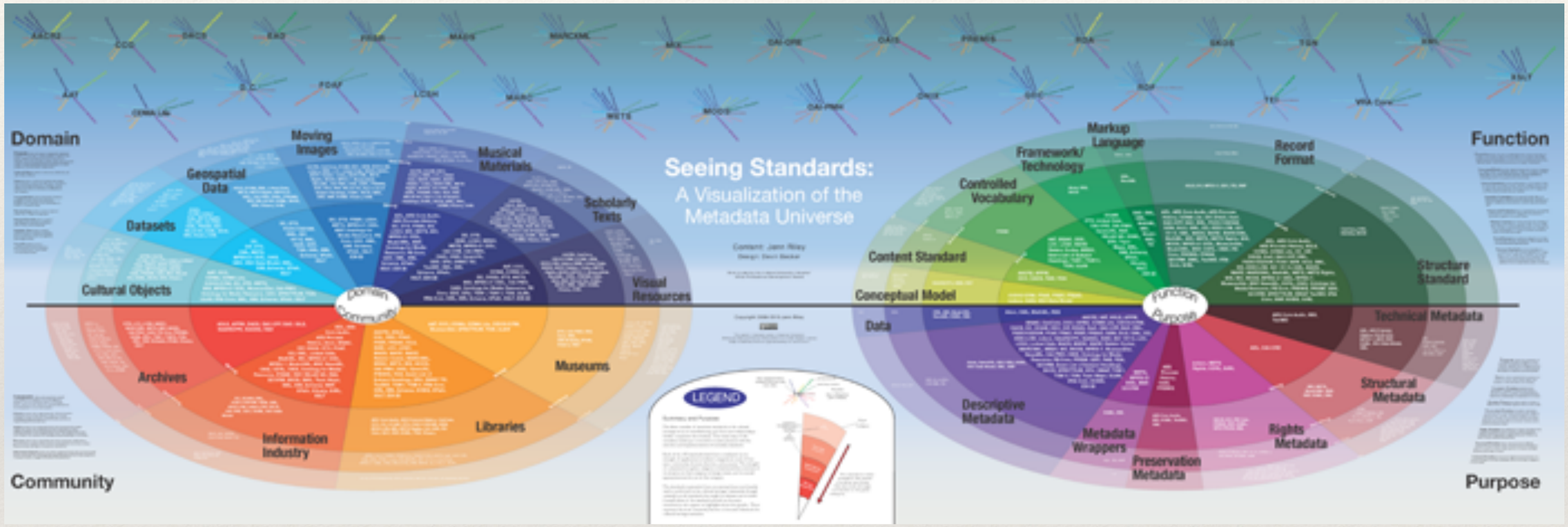
```
<?xml version="1.0" encoding="UTF-8"?>
<ead xmlns="urn:isbn:1-931666-22-9"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="urn:isbn:1-931666-22-9 http://www.loc.gov/ead/ead.xsd">
    <eadheader audience="internal" countryencoding="iso3166-1" dateencoding="iso8601">
        <eadid countrycode="US">Basham Kelly Papers, MS-F24</eadid>
        <filedesc>
            <titlestmt>
                <titleproper>Guide to the Basham Kelly Papers, 1936-1988</titleproper>
                <subtitle></subtitle>
                <author>Processed by Judith Morgan; finding aid prepared by Diana Elizabeth</author>
                <sponsor></sponsor>
            </titlestmt>
            <publicationstmt>
                <date>1992</date>
                <publisher>University Archives, Rodgers Library, Bluegrass State University</publisher>
            </publicationstmt>
            <notestmt>
                <note>
                    <p>
                        <subject source="lcsh">English literature</subject>
                        <subject source="lcsh">College teachers</subject>
                        <subject source="lcsh">Archival resources</subject>
                    </p>
                </note>
            </notestmt>
        </filedesc>
        <profiledesc>
            <creation>The collection was processed at the University Archives in 1992 by Judith Morgan. The finding aid was prepared by
            <langusage>English</langusage>
            <descrules>Find aid prepared using Describing Archives, a Content Standard</descrules>
        </profiledesc>
    </eadheader>
    <frontmatter>
        <titlepage>
            <titleproper>Guide to the Basham Kelly Papers</titleproper>
            <date>1936-1988</date>
            <num type="Collection number:">MS-F24</num>
```

what i am not talking about

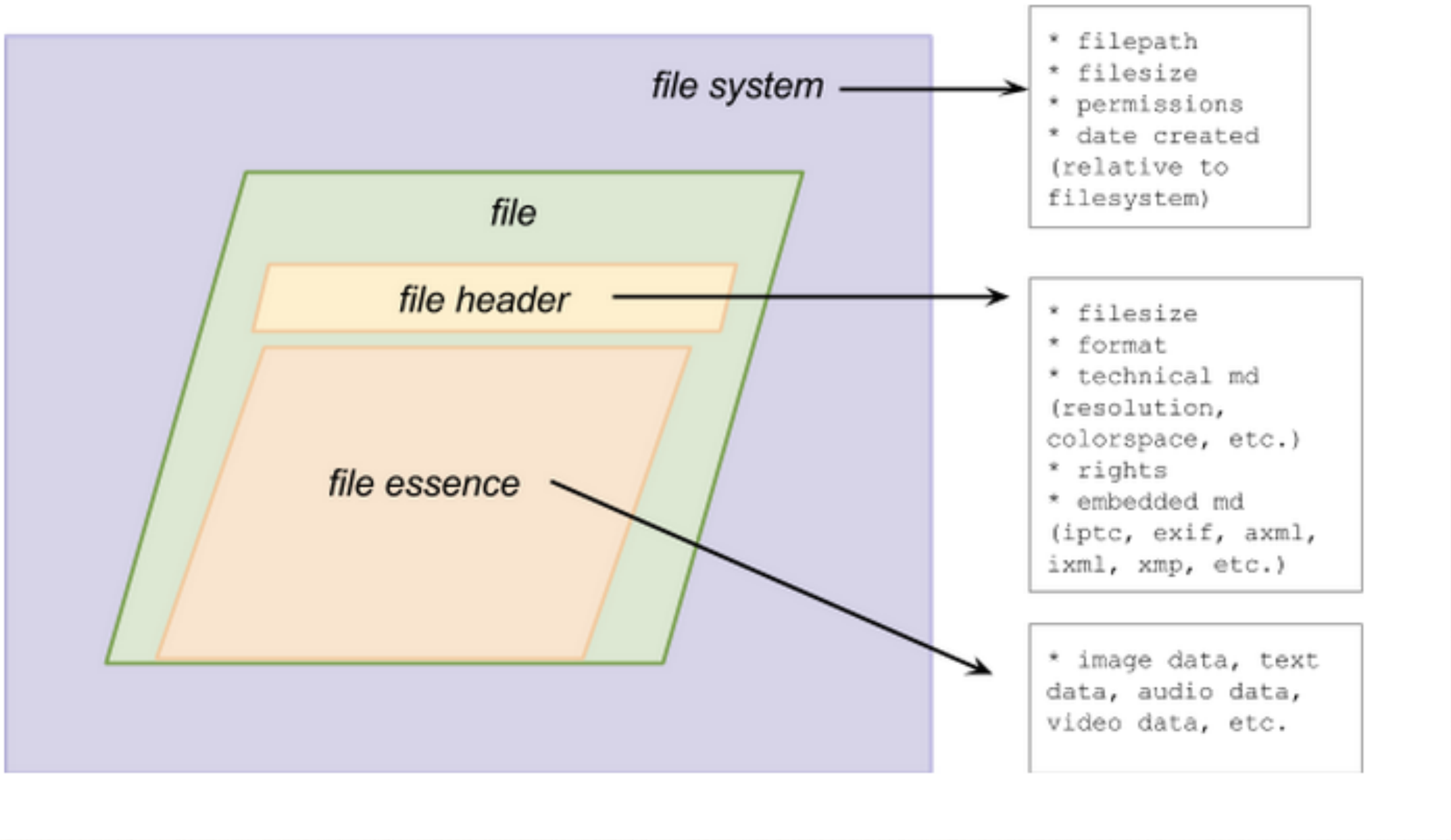
Dublin Core

```
- <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

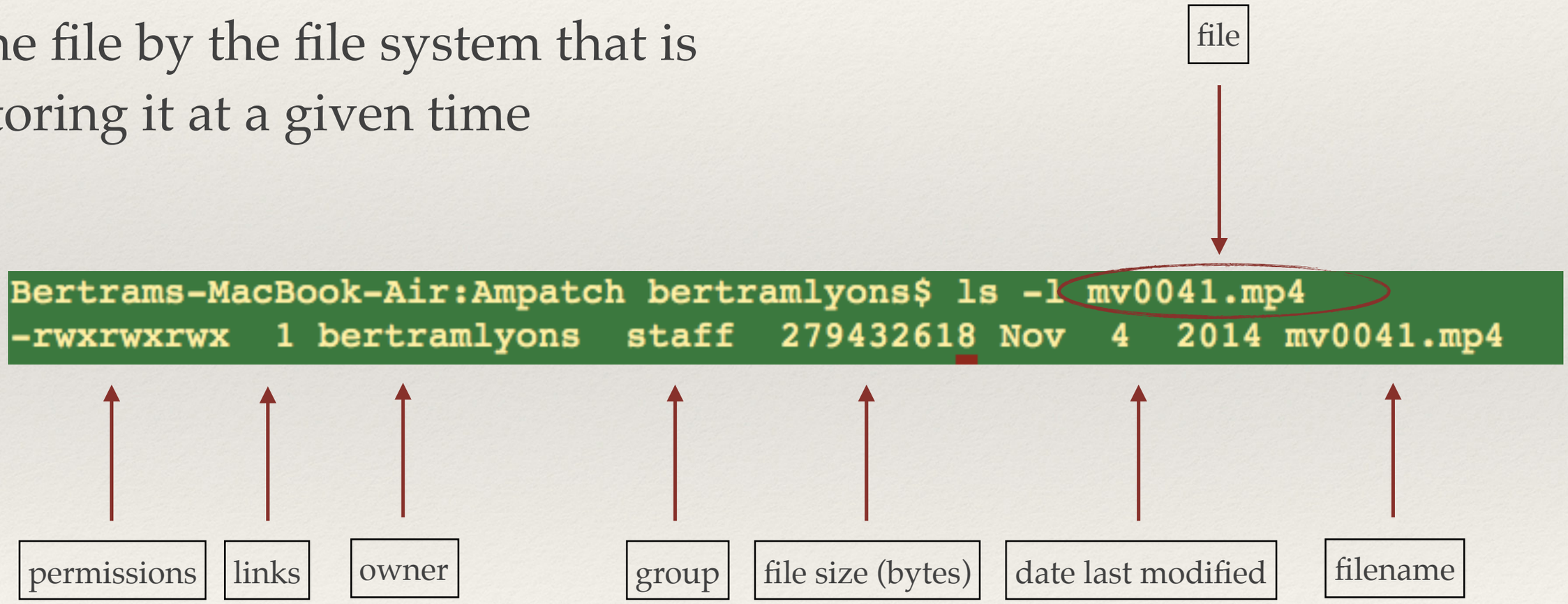


what i AM talking about



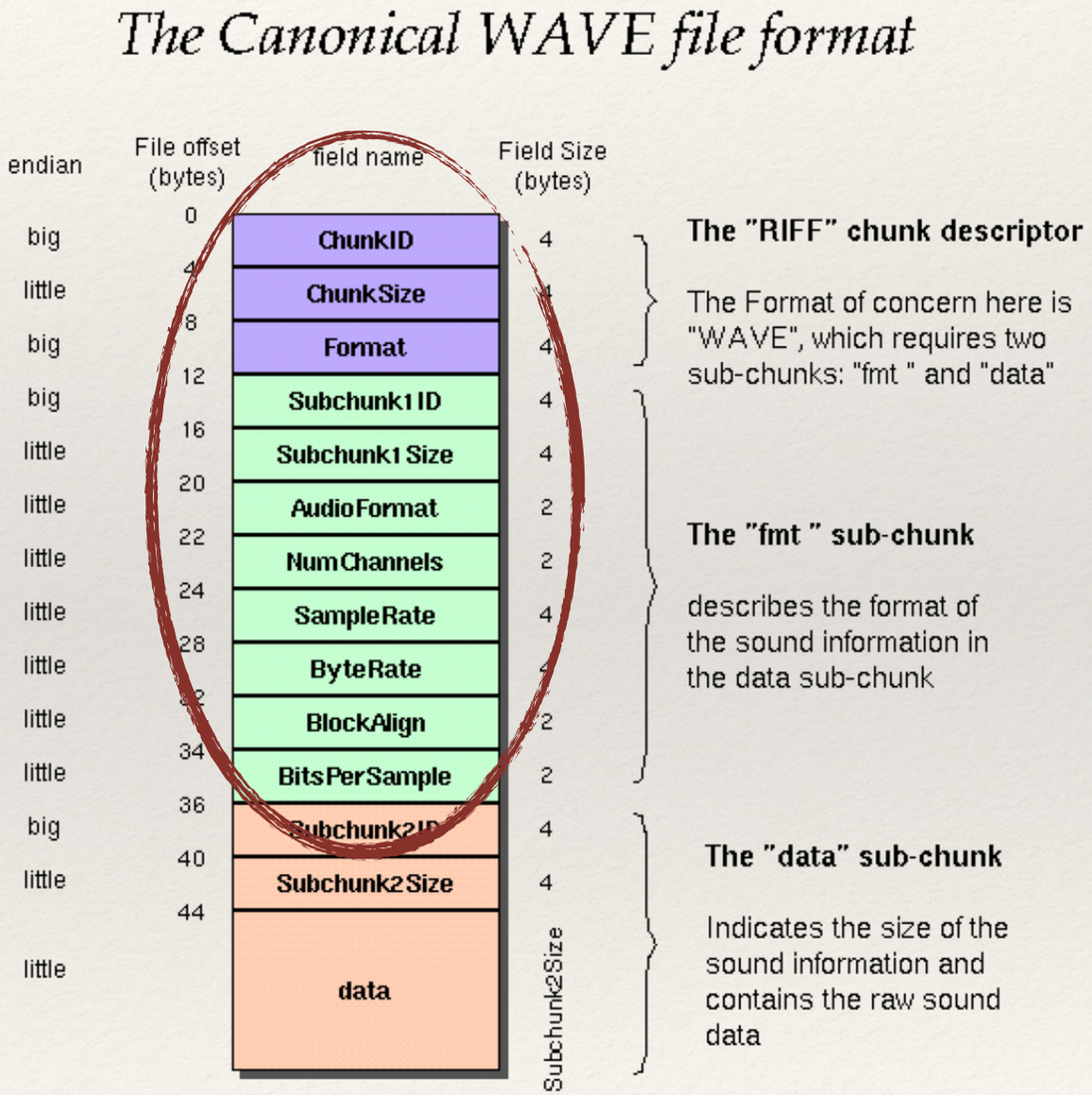
file system metadata

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



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.



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.)

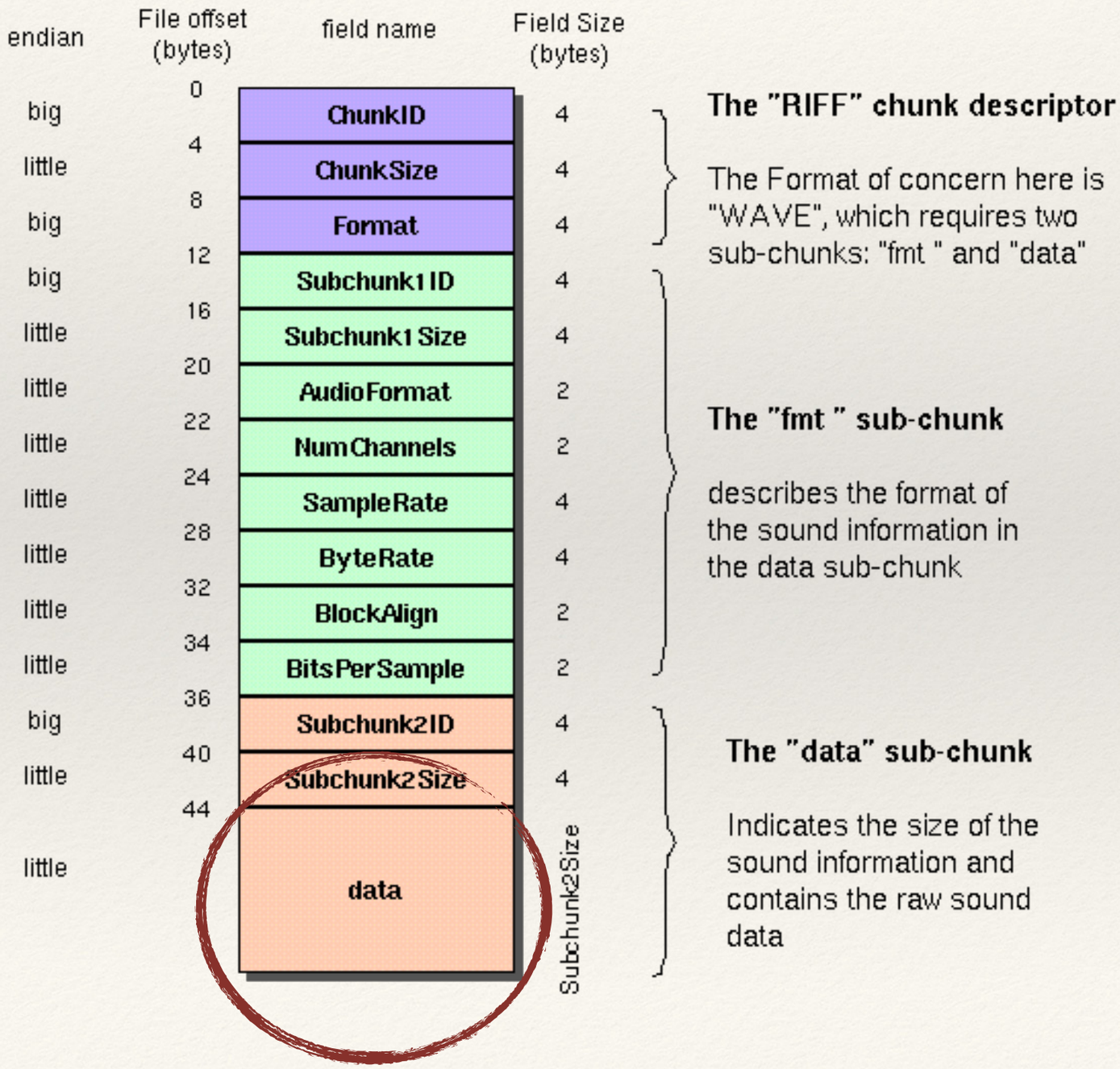
* video data (for mov, mp4, mxf, avi, etc.).

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

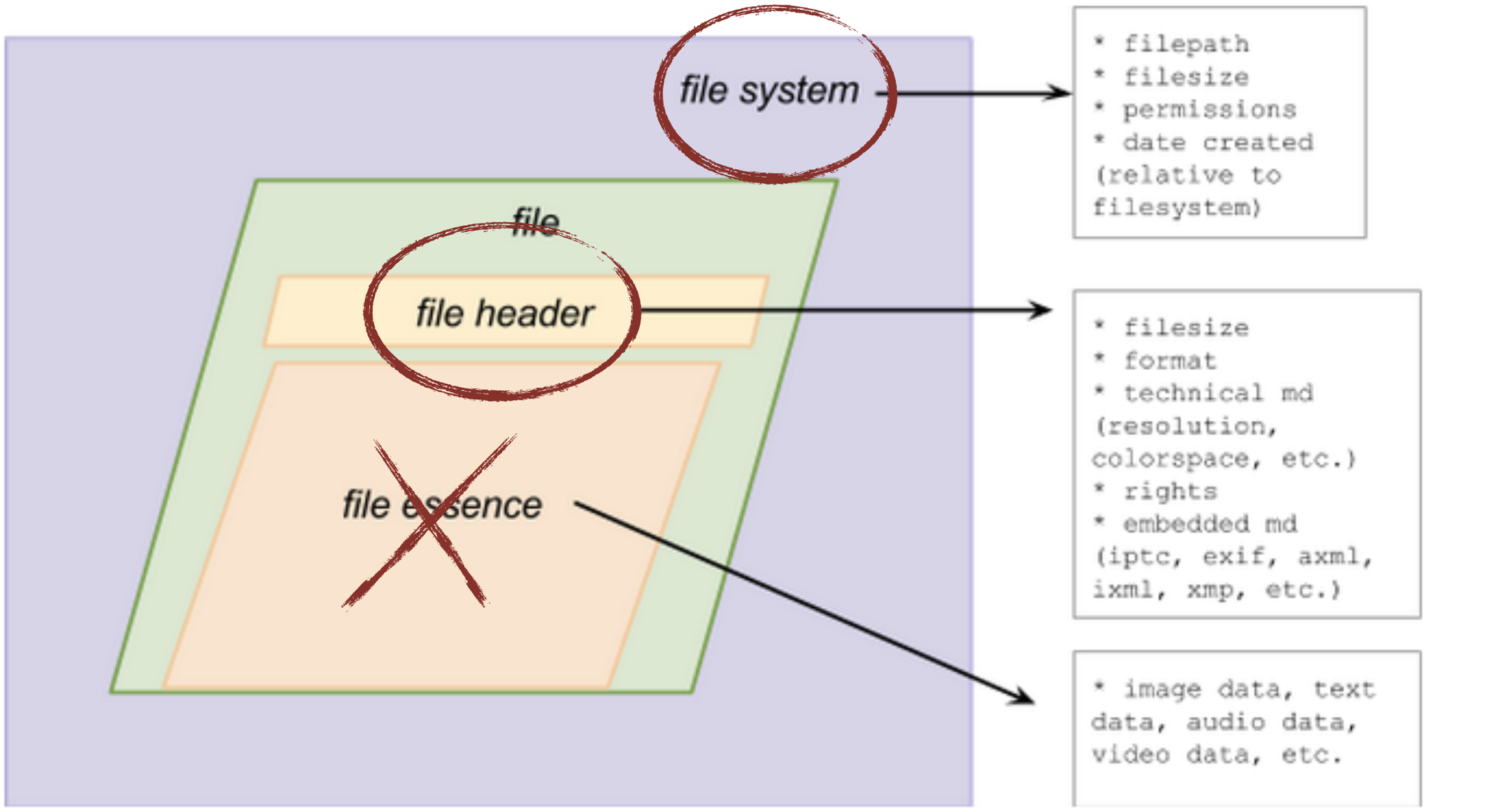
* etc.

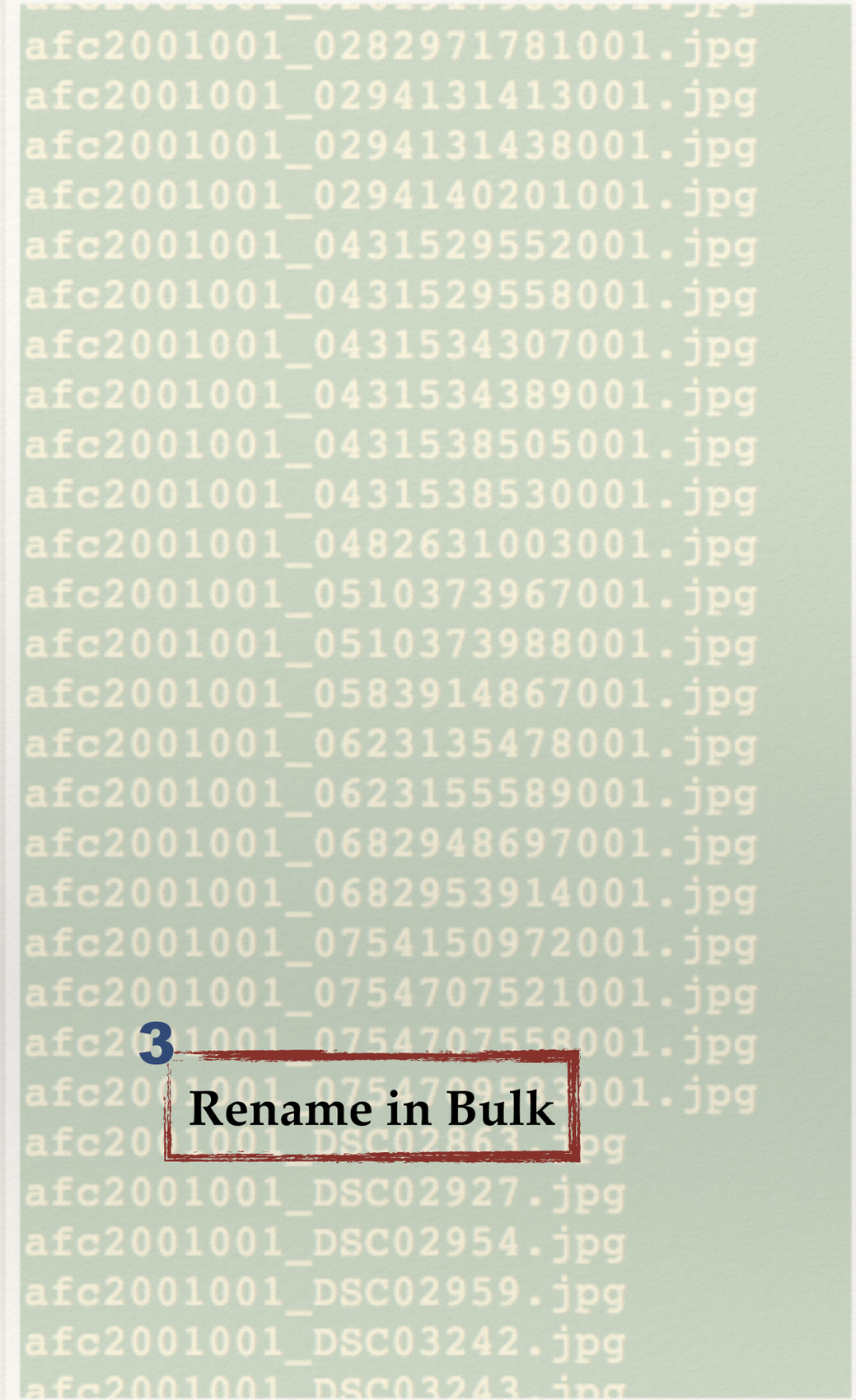
file essence (or content)

The Canonical WAVE file format

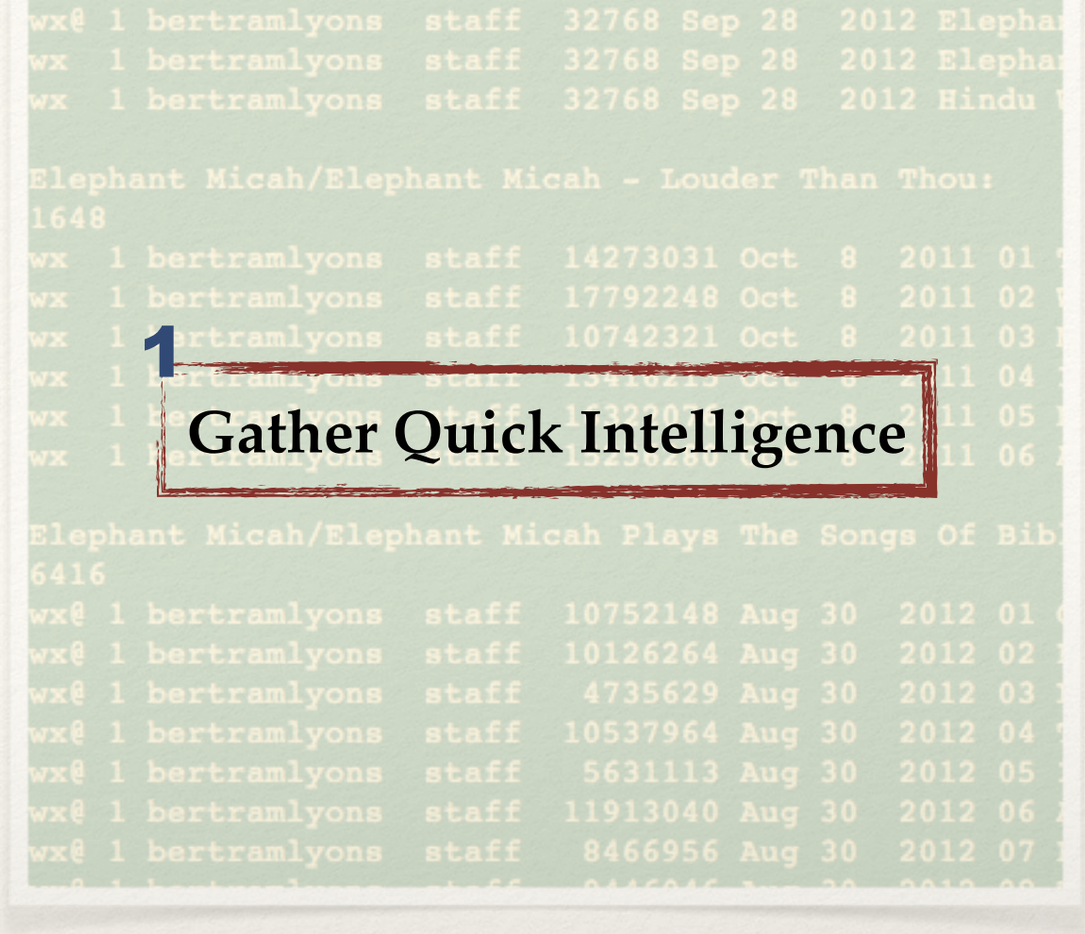


what i AM talking about



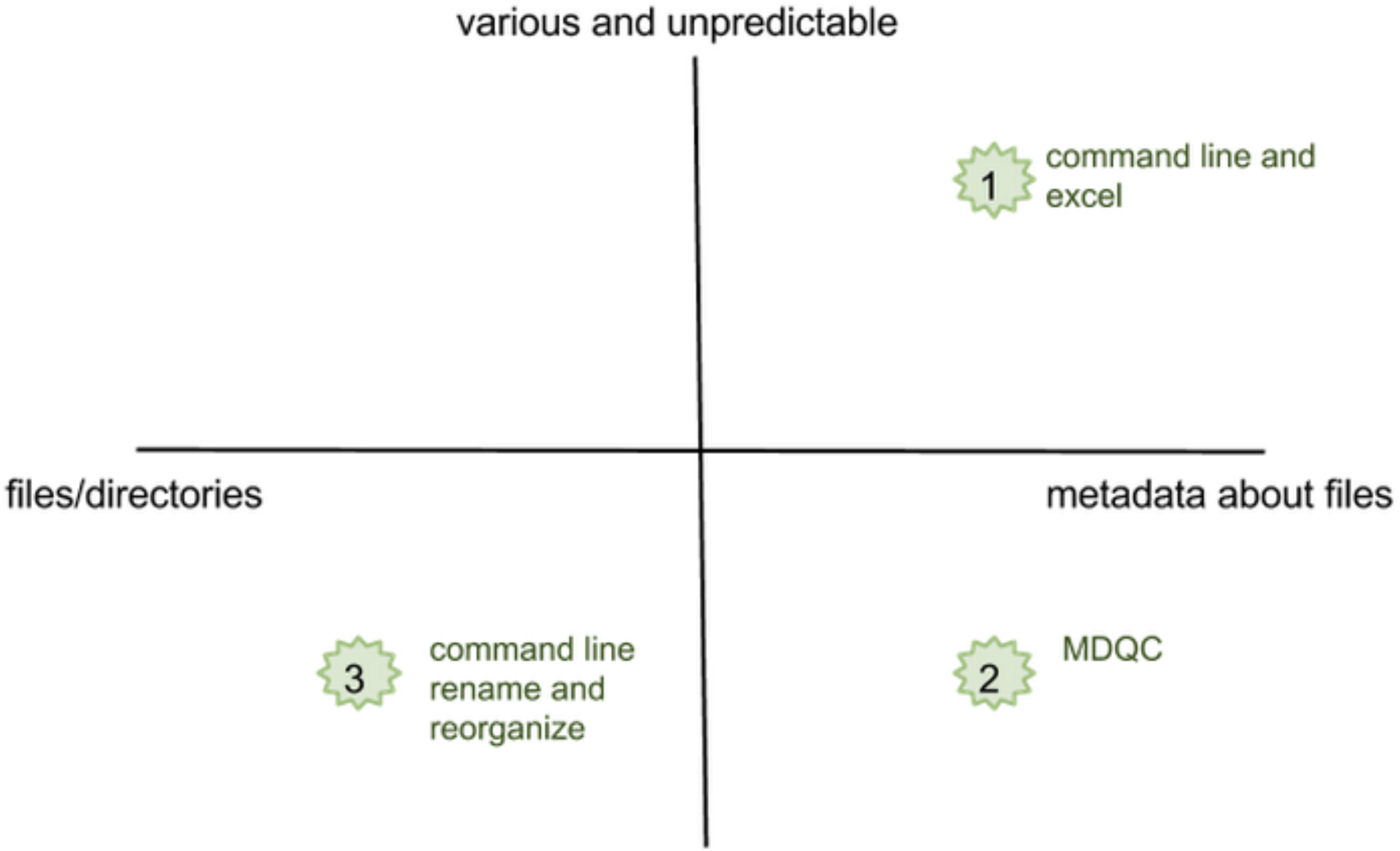


3 **Rename in Bulk**



1 **Gather Quick Intelligence**

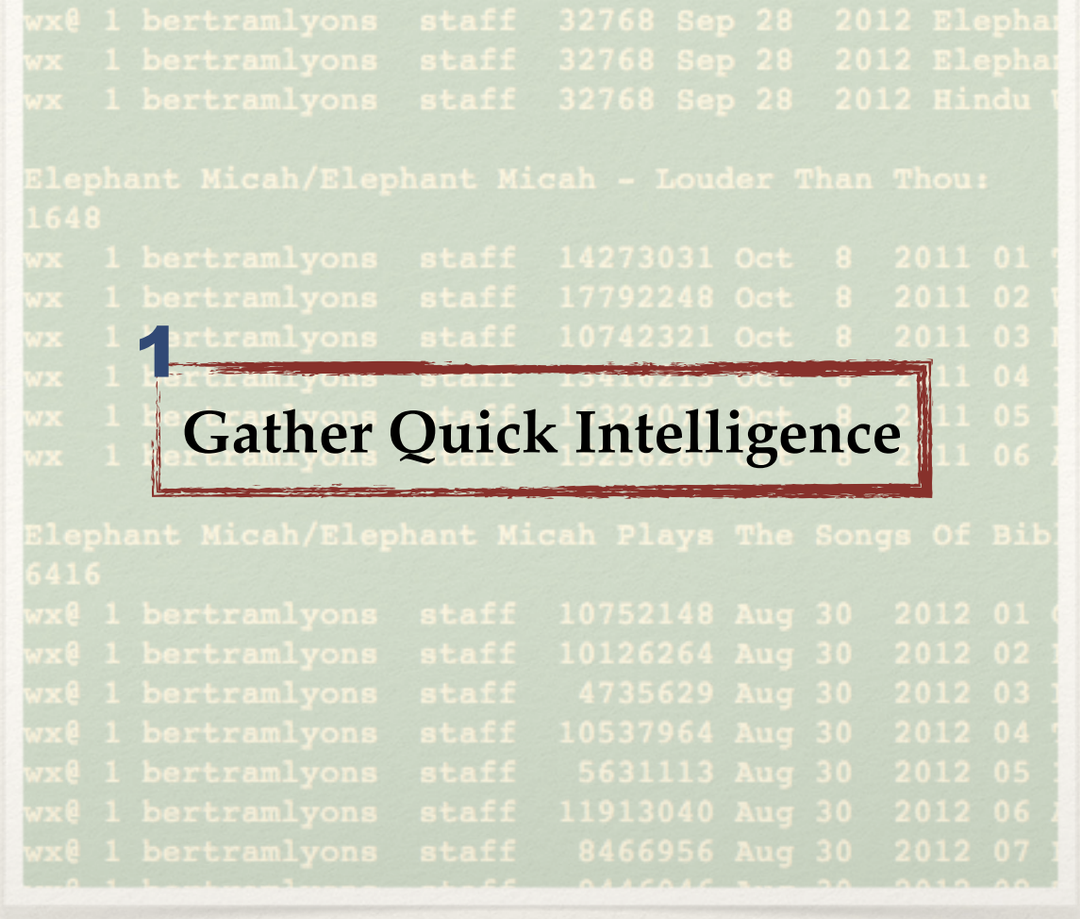
2 **Perform Quality Control**


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

Activity 2

Homogeneous and Predictable Aggregation of Files

Using file header metadata to perform quality control

Goals of this method

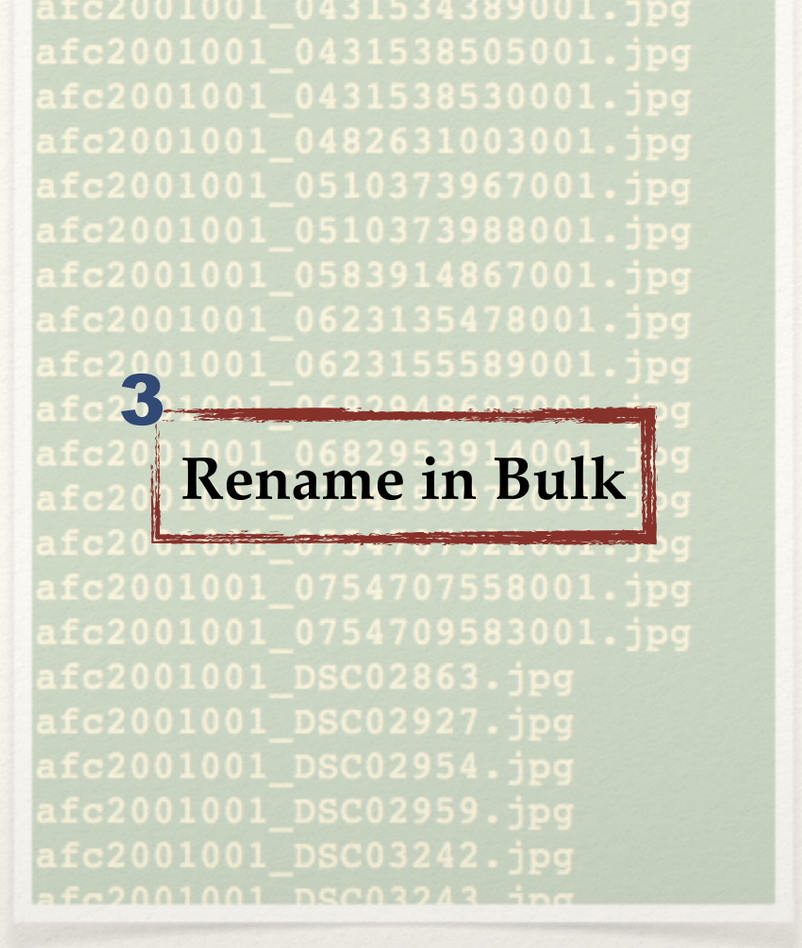
1. Do all files in the batch meet expectations?

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