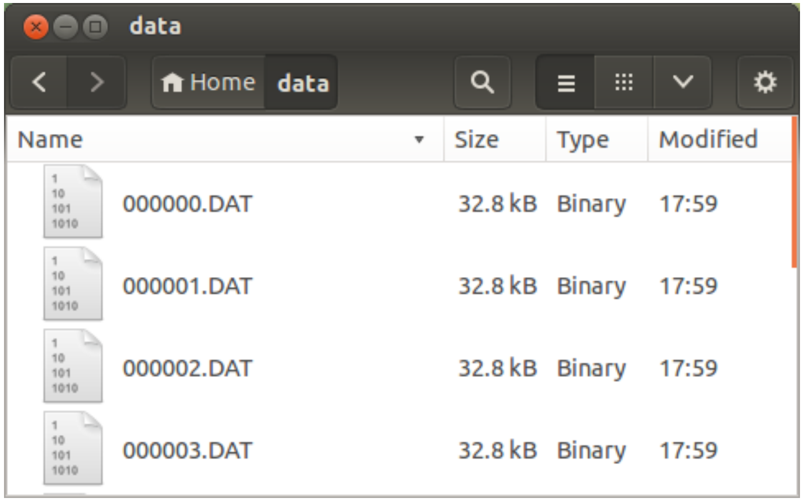
Digging into File Formats:

Poking around at data using `file`, DROID, JHOVE, and more

Presented by Stephen Eisenhauer
UNT Libraries TechTalks
February 12, 2014

Why?

- We handle a lot of digital information
- It's not always readily identifiable
  - Names/extensions can be meaningless
  - Recovered data may have no names or metadata at all



Name	Size	Type	Modified
000000.DAT	32.8 kB	Binary	17:59
000001.DAT	32.8 kB	Binary	17:59
000002.DAT	32.8 kB	Binary	17:59
000003.DAT	32.8 kB	Binary	17:59

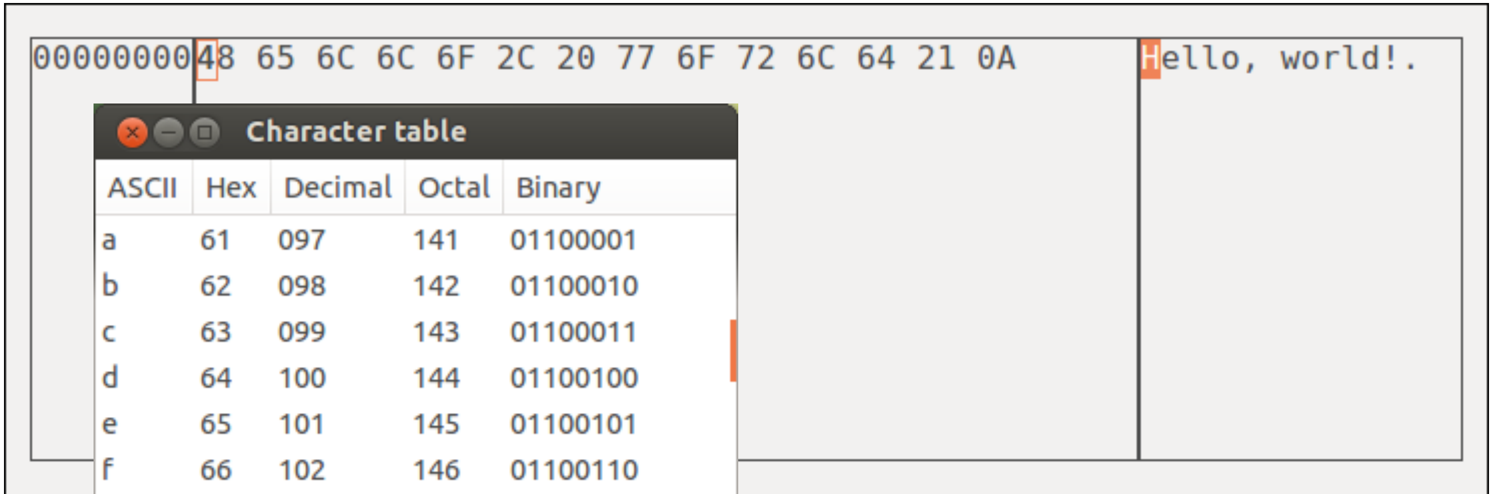
"I totally know what's in this folder."

Why?

- Sometimes we just need to verify a file is what it is supposed to be
  - "Why won't this video open?"
    "What? It's actually a HTML 404 document??"
- Maybe you want to automate
  - Statistical analysis, reporting, workflow, etc.

What's in a file, anyway?

- Files are sequences of numeric values
- Those values are meaningless if you don't know what they represent
  - ASCII characters? Colors? Something more complex?



```
00000000 48 65 6C 6C 6F 2C 20 77 6F 72 6C 64 21 0A    Hello, world!.
```

ASCII	Hex	Decimal	Octal	Binary
a	61	097	141	01100001
b	62	098	142	01100010
c	63	099	143	01100011
d	64	100	144	01100100
e	65	101	145	01100101
f	66	102	146	01100110

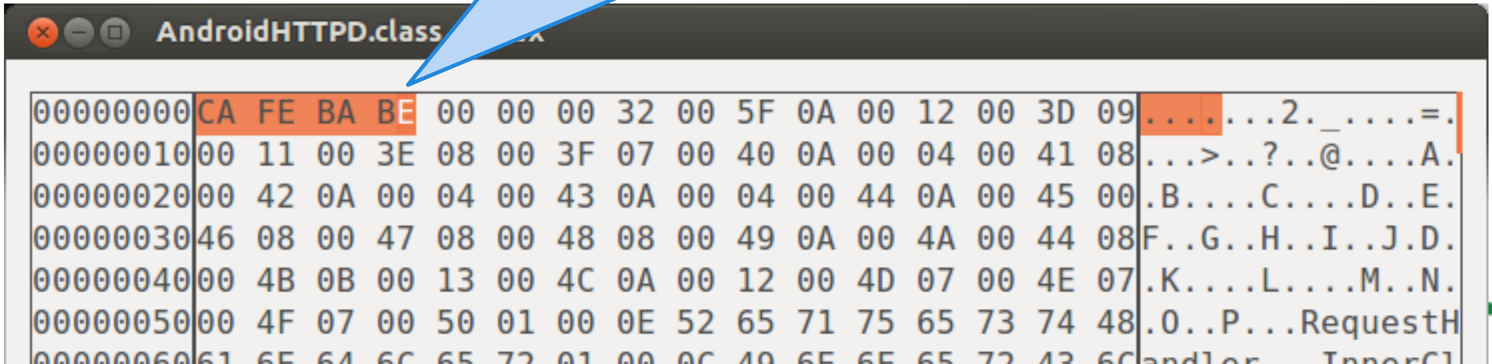
What’s in a file, anyway?

- Filenames aren’t stored inside the file
- File extensions are really just hints
- Metadata only exists within a file if the format specifies it (MP3, PDF, DOC…)

So, how can we tell what’s in mystery data?

- File Identification Tools: Software trained to look for certain special patterns in data to determine its file format
- Usually known as “magic numbers”

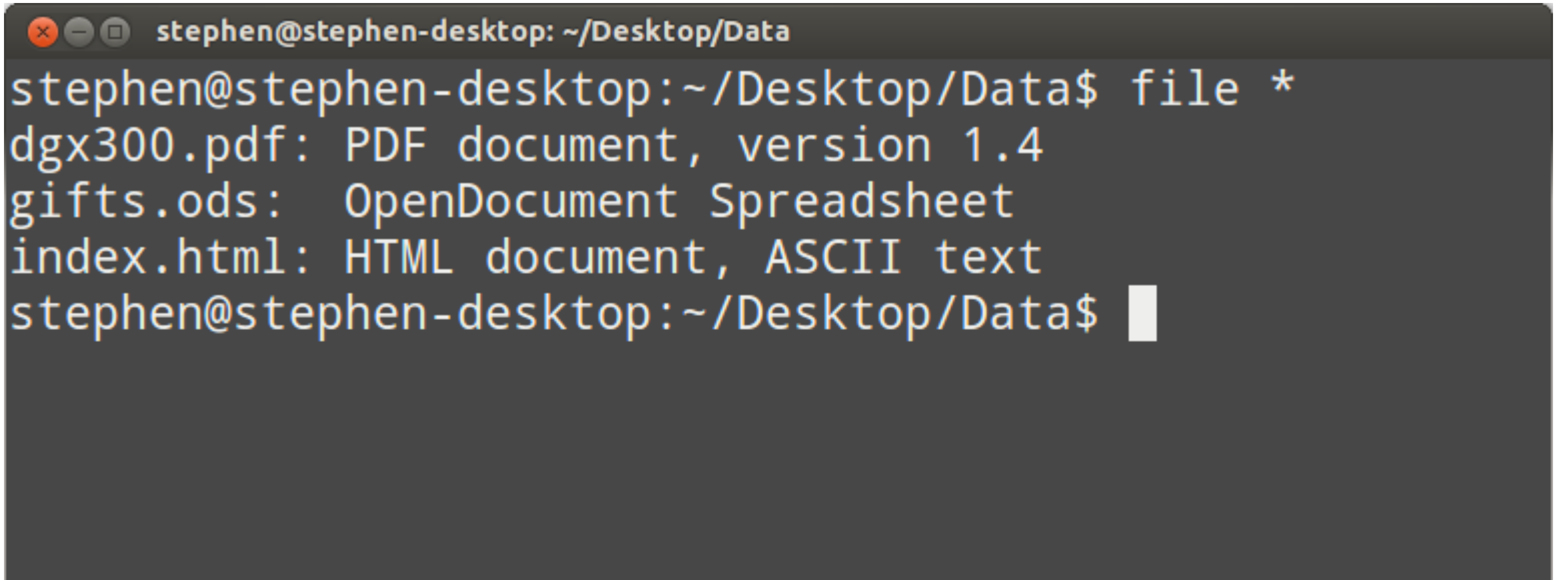
Fun fact: Java class files all start with the hexadecimal number **CAFEBABE** or **CAFED00D**.



```
00000000CA FE BA BE 00 00 00 32 00 5F 0A 00 12 00 3D 09.......2._....=.
0000001000 11 00 3E 08 00 3F 07 00 40 0A 00 04 00 41 08...>..?..@....A.
0000002000 42 0A 00 04 00 43 0A 00 04 00 44 0A 00 45 00.B....C....D..E.
0000003046 08 00 47 08 00 48 08 00 49 0A 00 4A 00 44 08F..G..H..I..J.D.
0000004000 4B 0B 00 13 00 4C 0A 00 12 00 4D 07 00 4E 07.K....L....M..N.
0000005000 4F 07 00 50 01 00 0E 52 65 71 75 65 73 74 48.O..P...RequestH
0000006061 6E 64 6C 65 72 01 00 0C 49 6E 6E 65 72 43 6Candler....InnerCl
```

The unix `file` command

- Comes installed on Mac OSX and most Linux operating systems
- (For Ubuntu, just install the “file” package)
- A very quick way to spot-check files



```
stephen@stephen-desktop:~/Desktop/Data$ file *
dgx300.pdf: PDF document, version 1.4
gifts.ods:  OpenDocument Spreadsheet
index.html: HTML document, ASCII text
stephen@stephen-desktop:~/Desktop/Data$ 
```

DROID: Digital Record and Object Identification

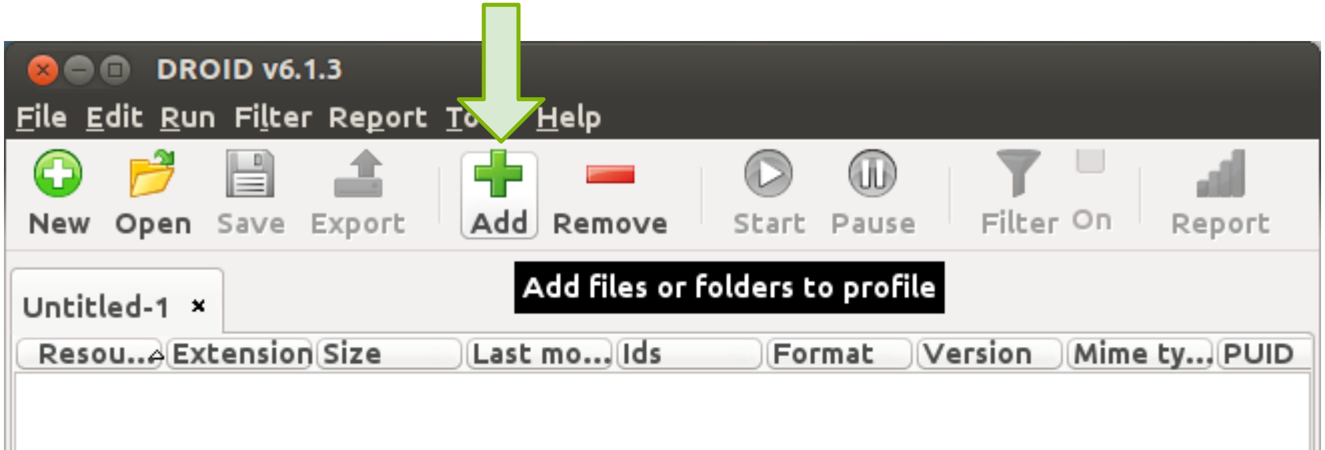
- Developed by the U.K. National Archives
- Fully free and Open Source
- Uses the industry-standard PRONOM registry of file format information
- Oriented toward large batches of files
- Comes with a graphical user interface in addition to a command-line tool

Getting DROID

- Works on Windows, Mac, Linux
- Requires Java
- Download from nationalarchives.gov.uk
  a. Click "Download the current version of DROID"
  b. Extract the ZIP file to your Desktop (anywhere, really)
  c. Run **droid.bat** (or **droid.sh** on Linux/OSX) to launch

Let’s make our first DROID profile

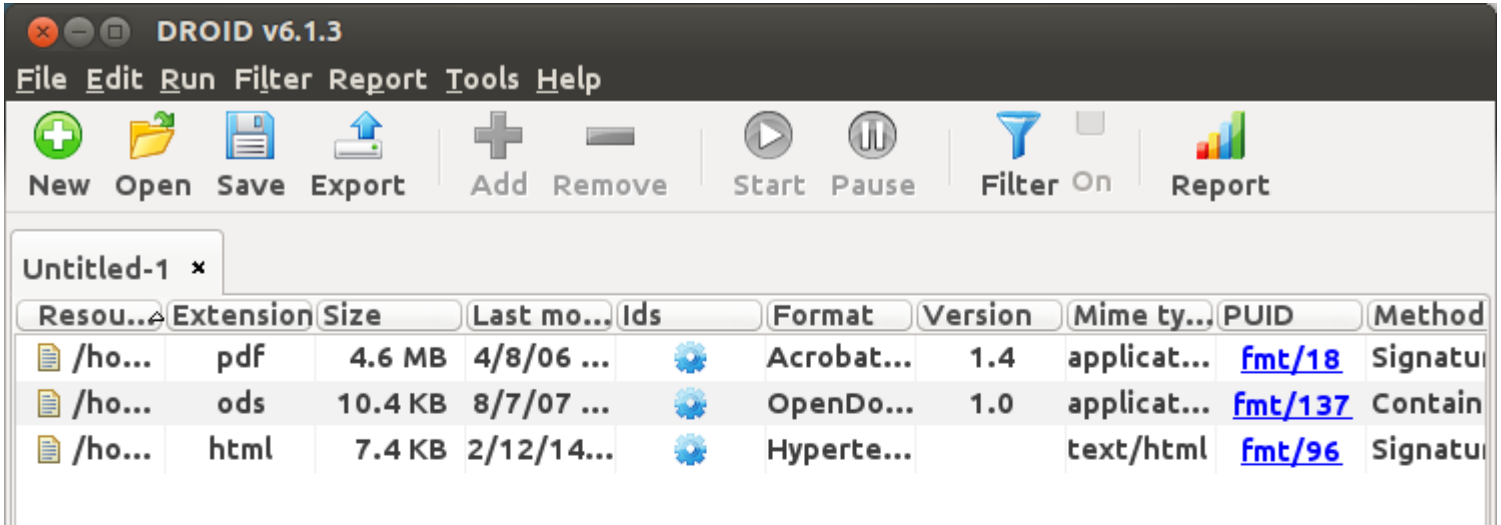
- After checking for updates, you’ll see an empty workspace labeled “Untitled-1”
- This is a “profile” in DROID terms; it represents a set of data you’re working on
- Add some files/folders to this profile using the **Add (+)** button



UNIVERSITY LIBRARIES UNT

Let DROID do its thing

- Once you've added the files you want to analyze, click the **Start** button
- When DROID is finished, you will see the columns in the profile populate with information



Resou...	Extension	Size	Last mo...	Ids	Format	Version	Mime ty...	PUID	Method
/ho...	pdf	4.6 MB	4/8/06 ...	⚙	Acrobat...	1.4	applicat...	fmt/18	Signatu
/ho...	ods	10.4 KB	8/7/07 ...	⚙	OpenDo...	1.0	applicat...	fmt/137	Contain
/ho...	html	7.4 KB	2/12/14...	⚙	Hyperte...		text/html	fmt/96	Signatu

We did it!

- You can now save this profile and open it later using DROID without needing to analyze the data again
- You can also use DROID's handy features:
  - **Export** lets you save the analysis as a CSV spreadsheet
  - **Filter** lets you drill down if your dataset is large
  - **Report** offers a range of statistical reports that can be generated with the analysis results

What’s the impact?

- `file` and DROID are commonly used within the digital preservation scene
- Institutional repositories often integrate with these tools
- Data curators use these tools when ensuring quality and integrity
- Software package including Archivematica, FITS, and the FCLA Description Service integrate with these tools out-of-the-box

Other tools to be aware of

- JHOVE (and JHOVE2)
  - Determines whether data of a known format is *valid*
- FITS (File Information Tool Set)
  - Analyzes data using a wide range of tools (including DROID and JHOVE) to look at it from every angle
- FCLA Description Service
  - Web-based application that analyzes a single file using DROID and JHOVE and produces a PREMIS XML document containing the results

Links to project web sites

DROID: http://nationalarchives.gov.uk/information-management/projects-and-work/droid.htm

JHOVE: http://jhove.sourceforge.net/

JHOVE2: https://bitbucket.org/jhove2/main/wiki

FITS: http://fitstool.org

Description Service: http://description.fcla.edu/

And quick primer on all of these tools:

http://metaarchive.org/imls/index.php/Format_Recognition_Tools_Documentation_for_ETDs

Thanks for attending!