Lecture Outline

- Report from Big Data Breakouts
- My lecture
- Individual presentations (~~20~~ 15 min. + 5 min. Q&A)
  - Pilch, Elena: A vision for data science.
  - Schultz, Kate: Big data: How do your data grow
  - Gaebel, Erika: Big data: The future of biocuration
  - Zhang, Liyuan: Big data's big unintended consequences
  - Lee, Chaimi: Measuring the value of Big Data exploitation systems
  - Chander Ravichanderan, Jishnu: A Survey of Big Data Methods, Assessments, and Approaches.
  - Nigam, Arushi: What is big data?

A Taxonomy of Digital File Formats

Agenda

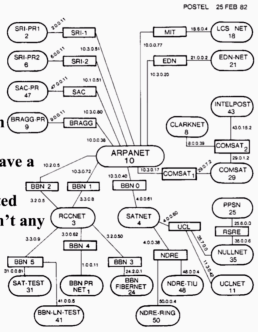
- Motivation
- Request for Comments (RFC) 822, 2045, 2046
- Top Level MIME/media types
- Implementations of the MIME taxonomy

ARPANET NWG History

- Network Working Group
- Laid the foundation for other similar organizations like IETF (Internet Engineering Task Force) and W3C (World Wide Web Consortium)
- In March 1969, RFC 001 was made by Crocker after the meeting in Utah based on the notes taken at the meeting
- RFCs were very successful, ended up becoming the official way to capture the Internet's design decisions, architecture and technical standards

Map of the Internet by Postel

- Map of the Internet circa 1982 posted by Postel



- The Economist wrote in 1997: "God, at least in the West, is often represented as a man with a flowing beard and sandals … if the Net does have a god, he is probably Jon Postel, a man who matches that description to a T." Postel disputed this classification, saying: "Of course there isn't any 'God of the Internet.' The Internet works because a lot of people cooperate to do things together."

2014 and beyond with Postel

- In 2014, almost 15 years after he initiated the controversy, the United States government finally ceded control of key Internet functions to the wider web community. And his name lives on in Postel's Law, a robustness principle formulated in RFC 760 and reworded in RFC 1122: "Be liberal in what you accept, and conservative in what you send."
  - https://howwegettonext.com/meet-jon-postel-the-man-who-invented-dot-com-a917a1faeff8
  - http://www.ntia.doc.gov/press-release/2014/ntia-announces-intent-transition-key-internet-domain-name-functions

Request for Comments

- Request for Comments were invented by Steve Crocker to help provide a record of the dealings of the Networking Group in the design of the ARPANET project
- ARPANET
  - Advanced Research Projects Agency (ARPA) Net (Network)
  - The predecessor to the modern Internet – how the internet was realized – first to implement TCP/IP, developed by Kahn & V. Cerf
- Crocker also developed the NCP tool, Network Control Program – first ARPANet protocol

RFC History - Continued

- In early 1970's, Jon Postel of USC's own Information Sciences Institute (ISI) took over the management of RFCs for the Internet
  - Served for 25 years
  - Authored and led development of more RFCs than anyone else
  - Developed DNS (Domain Name System), File Transfer Protocol (FTP), Telnet, and the Internet Protocol itself

Controversy - 1998

- January 28, 1998 – Postel emailed 12 of the operators of root DNS to have entries removed from USC ISI machine he created as part of ARPANET
  - Gov was trying to regulate the Internet
  - Postel wanted to show them the government couldn't take away the Internet's control from its creators
- A few days later he had the entries put back in
- U.S. presidential science advisor Ira Magaziner told Postel, "You'll never work on the Internet again."

RFCs – 1980s

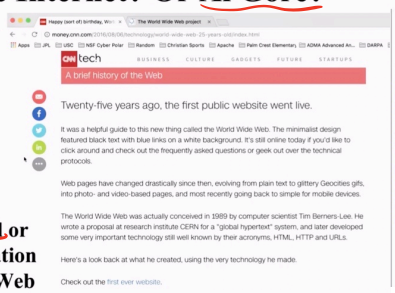
- Joyce K. Reynolds took over from Postel in 1983 and continues to manage RFCs until her death from cancer in December 2015
  - Member of IETF
  - Member of IANA
  - Helped develop Multimedia Mail system, Post Office Protocol, Telnet, and updated FTP

- Robert Braden
  - RFC co-editor for IETF
  - Developed the requirements for Internet Hosts
  - Contributed to FTP, RSVP and the Internet protocol itself

- Elizabeth Feinler
  - Lead, Stanford Research Institute Network Information Center
  - Contributed to and developed many RFCs
- RFC 2555 – a 30 year history of RFCs http://www.rfc-editor.org/rfc/rfc2555.txt

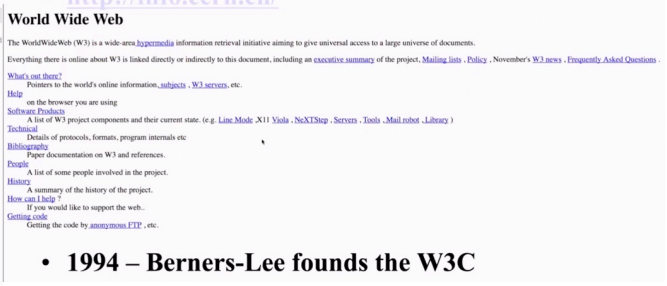
But wait! Didn't Tim Berners Lee invent the Internet? Or Al Gore?

- Berners-Lee while at CERN conceived of a global hyper-text media system
- HyperText and Hyper Text Transfer Protocol or (HTTP) is foundation for *browsing* the Web



Check out the first "web site"

- http://info.cern.ch/



- 1994 – Berners-Lees founds the W3C

Inventors of the Web

- Of course there is the ARPANET the modern *networking* required and associated protocols for the Internet including TCP/IP
- What goes on TOP of that network layer is where you typically hear the WWW/HTML/Berners-Lee notion of who invented the web
- There are many inventors of it!

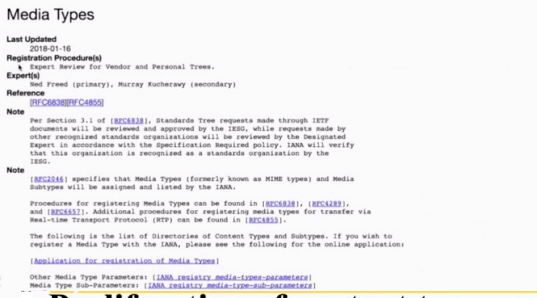
MIME, Content Type, Media Type?

- First, some terminology
- MIME = Multipurpose Internet Mail* Extensions
  - Classification of file types motivated by *email* standard
- Content Type~=MIME
  - Grew out of Internet age and identifying what *content* was available (movies, images, HTML, etc.) – Google calls them *file types*
    - https://support.google.com/webmasters/answer/35287?hl=en
- Media Type
  - First defined in RFC 2045, elaborated in 2046
  - https://www.ietf.org/rfc/rfc2045.txt

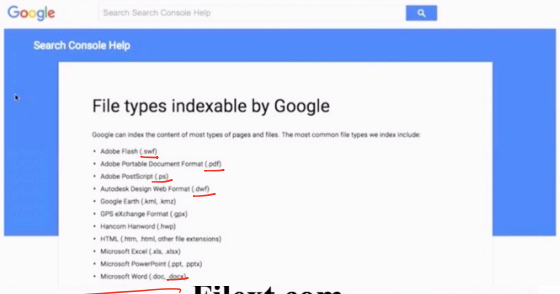
IANA List of Media Types

- Authoritative Source of Types
  - http://www.iana.org/assignments/media-types/media-types.xhtml
- Also available as XML, HTML, Plain Text
- Each individual top level type of registrations available as CSV, e.g., for the application/* type
  - http://www.iana.org/assignments/media-types/application.csv

IANA List of Media Types

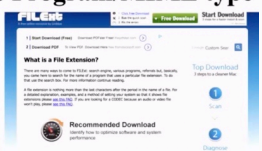


File Types Indexed by Google



Filext.com

- We currently have 26,024 records in the main database; 51,537 registered filetype records; and, 16,344 records in the Program/MIME type database.
- Course-grained estimate
- Looks at web log data
  - End of the URL suffix



Proliferation of content types available

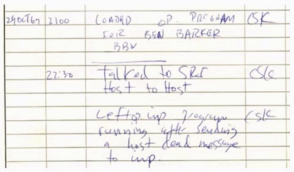
- By some accounts, 16K to 51K content types*
- What to do with content types?
  - Parse them
    - How?
    - Extract their text and structure
  - Index their metadata
    - In an indexing technology like Lucene, Solr, or Compass, or in Google Appliance
  - Identify what language they belong to
    - Ngrams

Why all this focus on email with MIME*?

- Today: over 2.6 billion active users and over 4.6 billion email accounts in operation
  - Most widely used communications medium on the Internet
  - https://phrasee.co/a-brief-history-of-email/
- Then (ARPANET)
  - Core purpose: communication within an organization
  - How to enable near instantaneous communication between machines within an organization
- Mail: became a key part of this!

The first email

- On October 29th, 1969, the first message was sent from a computer on ARPANET
  - 29 OCT 69 2100 LOADED OP PROGRAM FOR BEN BARKER
  - 22:30 TALKED TO SRS HOST TO HOST
  - LEFT IP PROGRAM RUNNING AFTER SENDING A HOST DEAD MESSAGE TO IMP



Roy Tomlinson

- American computer programmer
  - Came up with the idea of HOW to address mail on the internet
  - The concept of "@"
  - username@name of computer
  - Died in March 2016

- By 1976 75% of all ARPANET traffic was electronic mail
  - So successful idea came up to send mail *outside* of the network, *became idea for Internet itself*

Email in the 1980s

- As more and more inter-organizational emails were being sent, there became a great need for software to handle this
- Internet Service Providers began offering "hosting sites" for email
- For many people during this time, E-mail was the most exciting use for the "Internet"
- In 1993, "electronic mail" becomes "e-mail" in the public lexicon

Email in the 1990s

- America Online (AOL), Echomail, Hotmail, Yahoo
  - Shaped the Internet landscape
  - Poured money to increase the accessibility, utility, and ease of use of email
- By the late 90's Internet use exploded going from 55 million users worldwide in 1997 to 400 million by 1999
- During this time, e-mail SPAM also became a huge problem
- By the end of the 90's, early 2000's, to have an "email" address was akin to having a phone number and as important

One of the biggest uses of email is sending files

- Originally the simple mail transport protocol (SMTP) was 7-bit ASCII text only
  - Text files were emailed by including them in the message body
- 1980s → attachment become possible
  - UNIX tools like bundle and shar (shell archive) allow grouping of multiple files in email by attaching a file in the message body and unpacking on the other end with a single shell command
  - https://en.wikipedia.org/wiki/Email_attachment

1980s attaching non-text files to emails

- Manually encode 8 bit files using uuencode developed by Mary Ann Horton
  - Or BinHex or xxencode later tools developed
  - Paste the resulting encoded text into the body of the message

- 1985
  - "cc:Mail" program includes an "attachment" user interface
    - Uses uuencode under the hood
    - Microsoft mail later copies this

1990s attaching files to emails

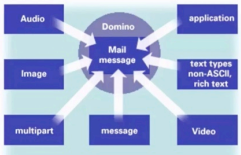
- MIME internet standard
- Developed by Nathaniel Borenstein and Ned Freed
- March 11, 1992
  - First MIME email attachment
  - RFC 2045 released in 1996
- MIME
  - Message and all of its attachments encapsulated within a multipart message with base64 encoding and converted into binary (7-bit) or (8-bit with 8BITMIME extension)


Nathaniel Borenstein


Ned Freed

One Application: Detecting MIME Types

- Identify and classify file types
  - MIME detection
    - Glob pattern
      - *.txt
      - *.pdf
    - URL
      - http://…pdf
      - ftp://myfile.txt
    - Magic bytes
    - Combination of the above means



Sample MIME message content types

sort of a little bit like string at the very beginning of a file which says sort of what one important details of the file, or the type of file it is

- Classification means reaction can be targeted

MIME Types Defined

- RFC 822 https://www.ietf.org/rfc/rfc0822.txt
  - Standard for the Format of ARPA Internet Text Messages - 1982
  - Envelope (Headers incl Email/Date/Time) + Contents
- RFC 2045 https://www.ietf.org/rfc/rfc2045.txt
  - 822 inadequate for including other file types in Email
  - Adds MIME-Version, Content-Type, Content-Transfer-Encoding and e.g., Content-ID, Content-Description
  - Included Charset definition (non US-ASCII)
  - Initial definition of MIME taxonomy

Magic Bytes (Examples)

List of file signatures

From Wikipedia, the free encyclopedia

This is a dynamic list and may never be able to satisfy particular standards for completeness. You can help by adding missing items with reliable sources.

This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Find sources: "List of file signatures" – news – newspapers – books – scholar – JSTOR (May 2021) (*Learn how and when to remove this template message*)

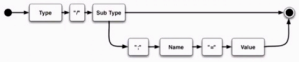
This is a list of file signatures, data used to identify or verify the content of a file. Such signatures are also known as magic numbers or Magic Bytes.

Many file formats are not intended to be read as text. If such a file is accidentally viewed as a text file, its contents will be unintelligible. However, sometimes the file signature in the common ISO 8859-1 encoding, with unprintable characters represented as the control code abbreviation or symbol, or codepage 1252 character where available, may be recognizable.

Hex signature	ISO 8859-1	Offset	Extension	
23 21	#!	0		Script or data to be passed to the program following the shebang (#!)
64 C3 B2 A1 (little-endian)	dÃ²¡	0	pcap	Libpcap File Format[7]
A1 B2 C3 D4 (big-endian)	¡²Ã			
4D 3C B2 A1 (little-endian)	M<²¡	0	pcap	Libpcap File Format (nanosecond-resolution)[7]
A1 B2 3C 4D (big-endian)	¡²<M			
0A 0D 0D 0A	----	0	pcapng	PCAP Next Generation Dump File Format[8]
ED AB EE DB	íë¾Û	0	rpm	RedHat Package Manager (RPM) package[9]
53 51 4C 69 74 65 20 66 6F 72 6D 61 74 20 33 00	SQLite format 3~	0	sqlite3, sqlite, db	SQLite Database[10]
53 50 38 31	SP81	0	bin	Amazon Kindle Update Package[11]
			PIC	IBM Storyboard bitmap file
00	~	0	PIF	Windows Program Information File
			SEA	Mac Stuffit Self-Extracting Archive

MIME Types Defined (part 2)

- RFC 2045
  - Content-Type:
    - Type / subtype
    - Types: message, multi-part, text, image, audio, video, application, extension-token
    - Parameters
      - Attribute=value
      - E.g., encoding=UTF-8
      - E.g., charset=us-ascii



- Parameters are case-insensitive
- 7 Initial Types defined, with mechanism to add new types defined by IETF

MIME Types Defined (part 3)

- RFC 2046 http://www.hjp.at/doc/rfc/rfc2046.html
  - Augments 2045 with definitions of MIME taxonomy types and initial sub-types
  - Plain, unrecognized sub-types defined
  - Octet-Stream sub-type of application
  - Discrete Types (atomic)
    - Application, image, video, audio, text
  - Composite Types
    - Message, multi-part
  - Extension Types
    - Formal way of adding new x-types, without requiring IANA

Character Set Defined

- A "character set" is used in MIME to refer to a method of converting a sequence of octets into a sequence of characters
  - RFC 2045
- It's a table
  - This set of octets/hex/binary data maps to these *symbol*
- Many of them
  - US-ASCII, UTF-8, UTF-16, ISO-8859-1, etc.

Top-level Media Types: Text

- Text/plain
  - General catch-all usually for servers that cannot accurately discern type – unrecognized subtypes should use this
  - Default charset: US-ASCII
- Code
  - text/x-java-source, text/x-c++src
- Charsets matter here
  - One way of detecting text – high probability of text if US-ASCII – if unrecognized charset and text, default to application/octet-stream

Top-level Media Types: Image

- Names are not case-sensitive for sub-types
  - JPEG, GiF, Png
  - Mostly all lowercase
- Sub-type of application/octet-stream, especially if sub-type not discernable
- JPEG, jp2 (JPEG-2000), GIF, Portable Network Graphics (PNG)

Top-level Media Types: Audio

- Initial sub-type of basic
  - single channel audio encoded using 8bit ISDN mu-law [PCM] at a sample rate of 8000 Hz.
- Sub-type of application/octet-stream, especially if sub-type not discernable
- Midi, MPEG layer 3, WAV (windows audio-video interleaved), MPEG layer 4, etc.

Top-level Media Types: Video

- Time-varying picture image possibly with color and coordinated sound – RFC 2046
- May include animated drawings
- Sub-type of application/octet-stream, especially if sub-type not discernable
- May use video/video for subtypes that are not discernable

Top-level Media Types: Application

- Typically content that must be processed by an application before being "useful"
- Expected uses
  - file transfer, spreadsheets, data for mail-based scheduling systems, and languages for "active" (computational) material
- Sub-types divided into
  - Octet-stream and postscript
  - Octet-stream includes PADDING and TYPE parameters
- Sub-type of application/octet-stream, especially if sub-type not discernable

Top Level Composite Type: Multipart

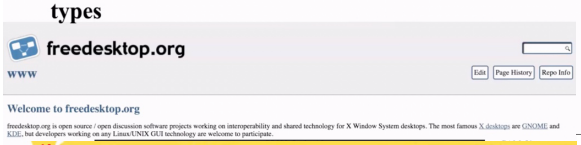
- One or more MIME types inside
  - Delimited by a –boundary– header
  - Two hyphen characters is important
- Body part is NOT an RFC 822 message body
- Digest sub-type

Top Level Composite Type: Message

- Encapsulates sending another message within a message
- Subtypes indicate the encoding
- Partial sub-type for split messages across several messages
- Must include ID field header

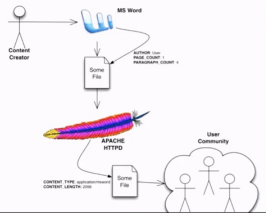
Example implementation: Freedesktop.org

- http://www.freedesktop.org/wiki/Specifications/shared-mime-info-spec/
- Specifies an XML schema and DTD for MIME
- Needed because desktop applications and a desktop environment needs to know about MIME types



Example implementation: Apache HTTPD

- Web server needs to understand how content is touched and moved around
- Needs to know MIME database



Example Implementation: Tika

extract the important metadata even when it's not obvious

- Apache Tika MIME DB includes
  - Glob file extension (*.ppt)
  - Digital File Signatures (MIME Magic)
    - Offsets
    - Priority
    - Way to combine them
  - Comments
  - Parent/Subtype (from IANA hierarchy)
  - XML hierarchy including XMLNS
    - Parent/sub-type with schema

Summary

- Motivation for MIME taxonomy
- History of MIME taxonomy
- Top Level Types
- Sub Types

In-Class Activity

5min

- Identify how you would characterize the MIME types of 3 of your real world datasets from last group discussion
  - Give general broad strokes about
    - How many of each type?
    - How does the type correspond with the 5 Vs? Are certain types typically at more velocity? More volume? Variety?
  - What applications go along with generating and modifying data belonging to those types?
    - Think about 2-3 software applications or architectures that modify and edit/create/update/delete those types
- Make slides we will discuss the outcomes of this next class

Refresher

- Volume (size)
- Value (utility)
- Variety (diversity of data)
- Velocity (speed of data coming in or out)
- Veracity (accuracy of data)
- (sometimes) Variability (how stable is the data, e.g., does the context of words change over time)