



The  
University  
Of  
Sheffield.

# A temporal overview of TNA's CDX index

Philip Webster, The University of Sheffield  
Claire Newing, The National Archives

# Contents

- The Project
- The UK Government Web Archive
- CDX processing
- Archive Overview
- Temporal analysis – HTTP codes, media types and protocols

# The Project

- What information can be extracted from CDX files?
- What can analysis of that data tell us about the UK Government Web Archive?
- What are the potential pitfalls with using CDX data in this way?

# UK Government Web Archive

- Administered by The National Archives
- Archived versions of UK central government websites dated from 1996 to present
- Around 4,000 unique websites captured at least once
- Over 4 billion archive entries (DNS, HTTP – images, HTML, page resources and document types)

# UK Government Web Archive

- Data made available as ‘units’
- Semi-arbitrary division of the entire archive over physical drives – 1 drive is 1 ‘unit’
- No guarantee of chronological sequence
- ARC files hold UKGWA archive data
- CDX index available (derived from ARC)

# CDX processing

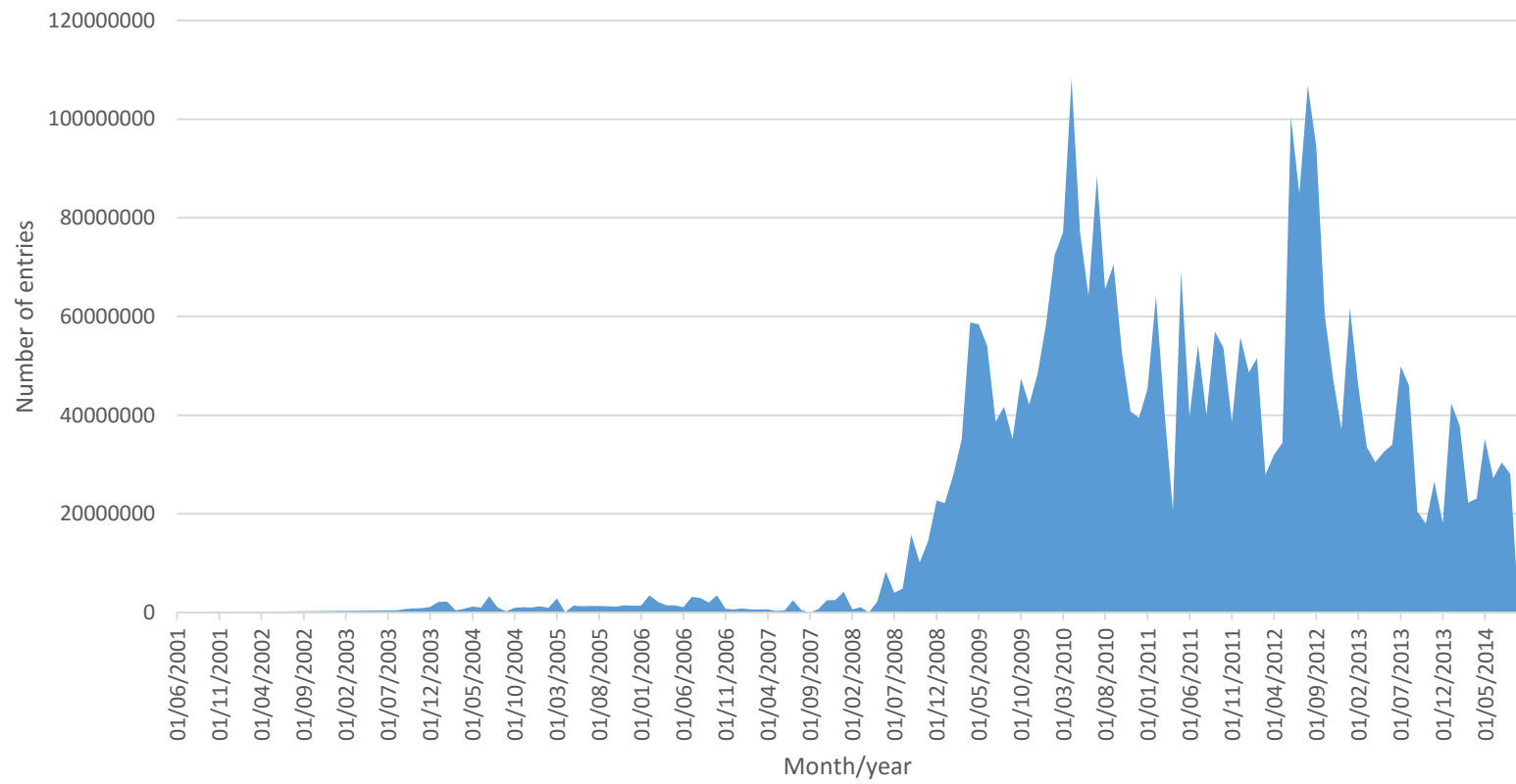
- Text-based index format
- Easily machine-readable
- Inefficient representation of dates and numeric types
- Easy to scan sequentially, but difficult to use for faceted, dynamic querying
- (because it wasn't designed for that)

# Archive overview

- UKGWA composition:
- By media (MIME) type, temporal coverage, file size, HTTP response code
- Temporal range from 1996-present
- Most data from 2008-present

# Archive overview

Temporal distribution of CDX index entries in the UKGWA



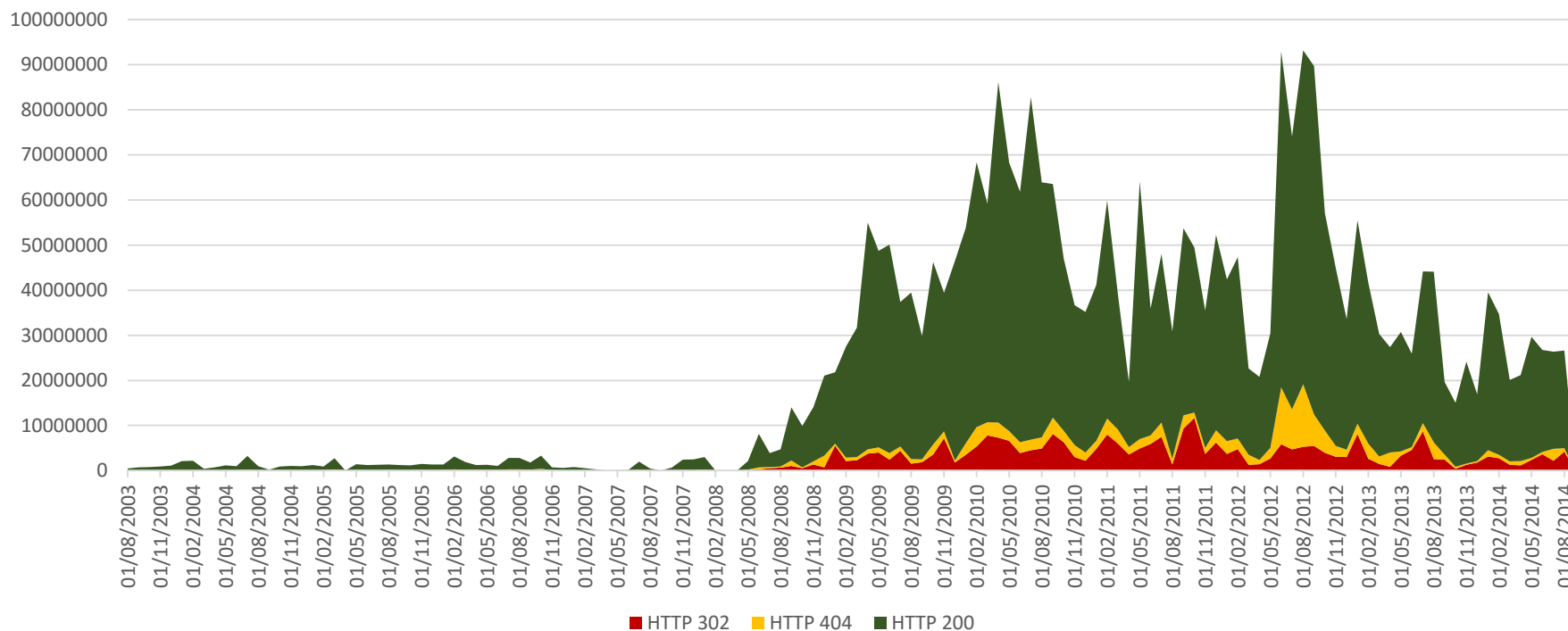


# HTTP status codes

- 3.3 billion HTTP status codes (3,279,650,659)
- Data range: August 2003 to August 2014

# HTTP status codes (absolute)

Raw frequency data for 200, 302 and 404 HTTP responses during crawls, 2003-2014



# HTTP status codes (absolute)

- Absolute frequency counts highlight peaks and troughs in crawl frequency.
- Data is sparse prior to 2008
- Proportion must be used to identify shifts in frequency

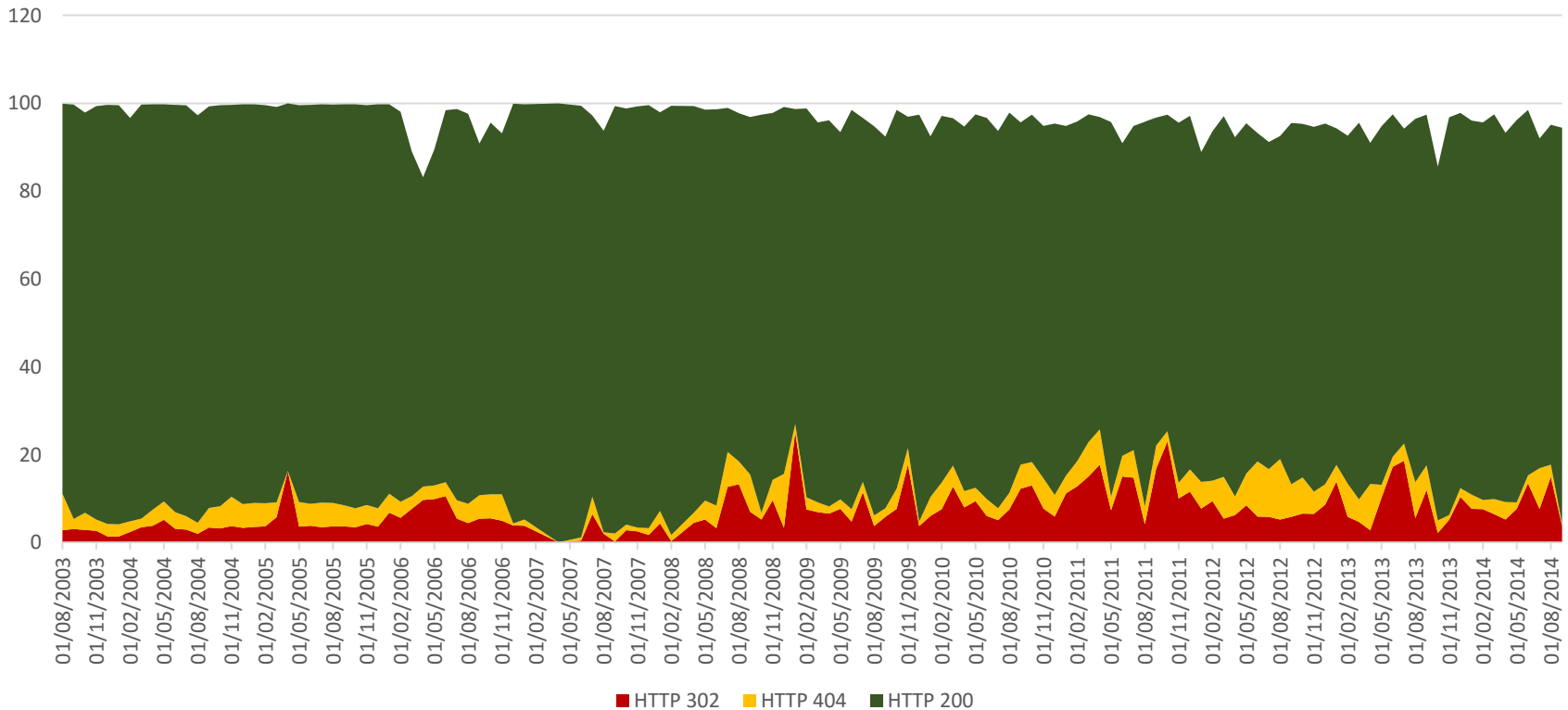
# HTTP status codes (relative)

- Graph restricted to 3 key HTTP status codes:
- 200 (success)
- 302 (redirect)
- 404 (not found)
- Other codes (500 etc.) excluded.



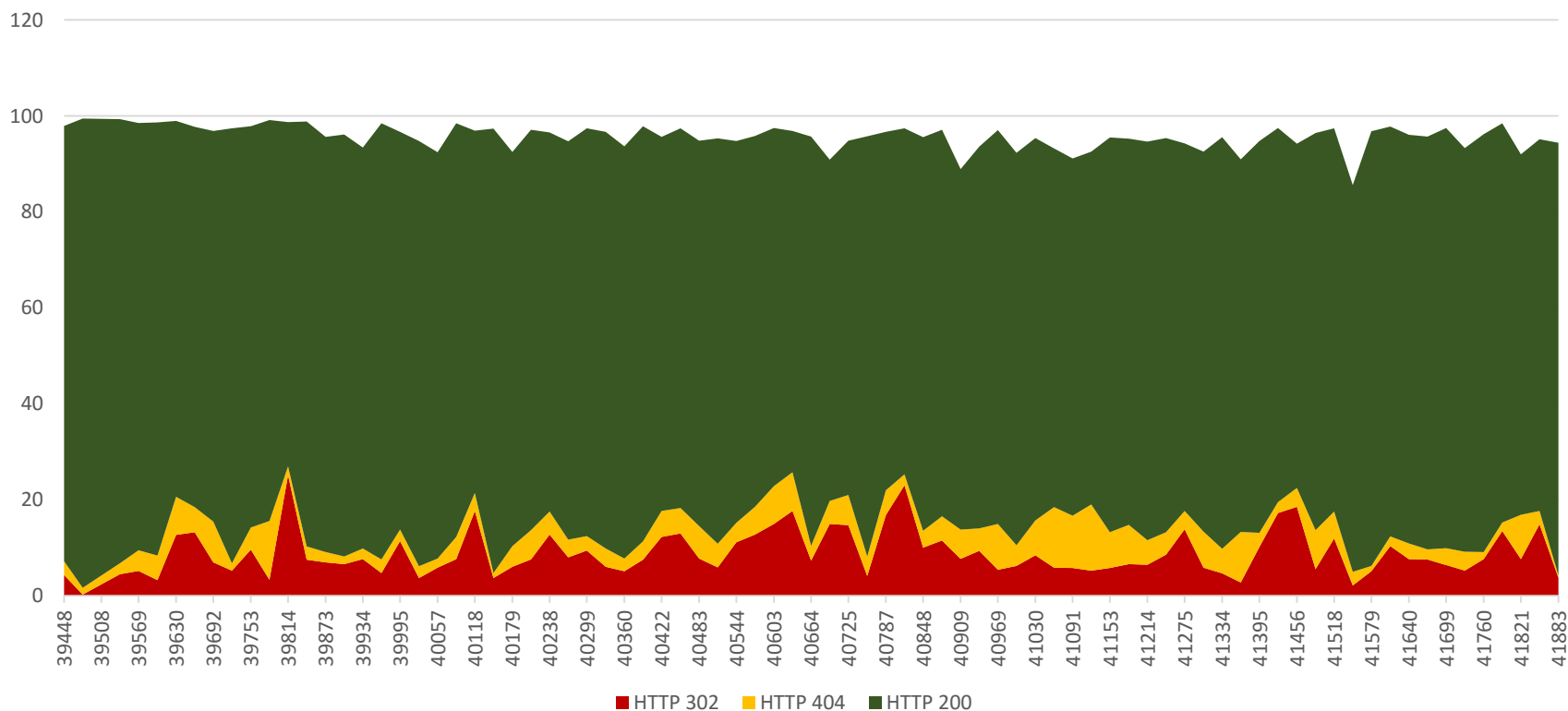
# HTTP status codes (relative)

Percentage data for 200, 302 and 404 HTTP responses during crawls, 2003-2014



# Post-2008 HTTP status codes

Percentage data for 200, 302 and 404 HTTP responses during crawls, 2008-2014



# HTTP status codes - trends

- Gradual increase in non-success response codes
- Possibly indicative of increased use of dynamic sites (HTTP 500), access control, or indicators of site closure.
- Gradual increase in the number of redirects (302) and not found (404) codes.

# HTTP status codes - issues

- Data is known to be influenced by short term changes in crawler focus.
- Shifting focus to specific domains of interest can skew results .
- Researchers using archive CDX data should consider this.



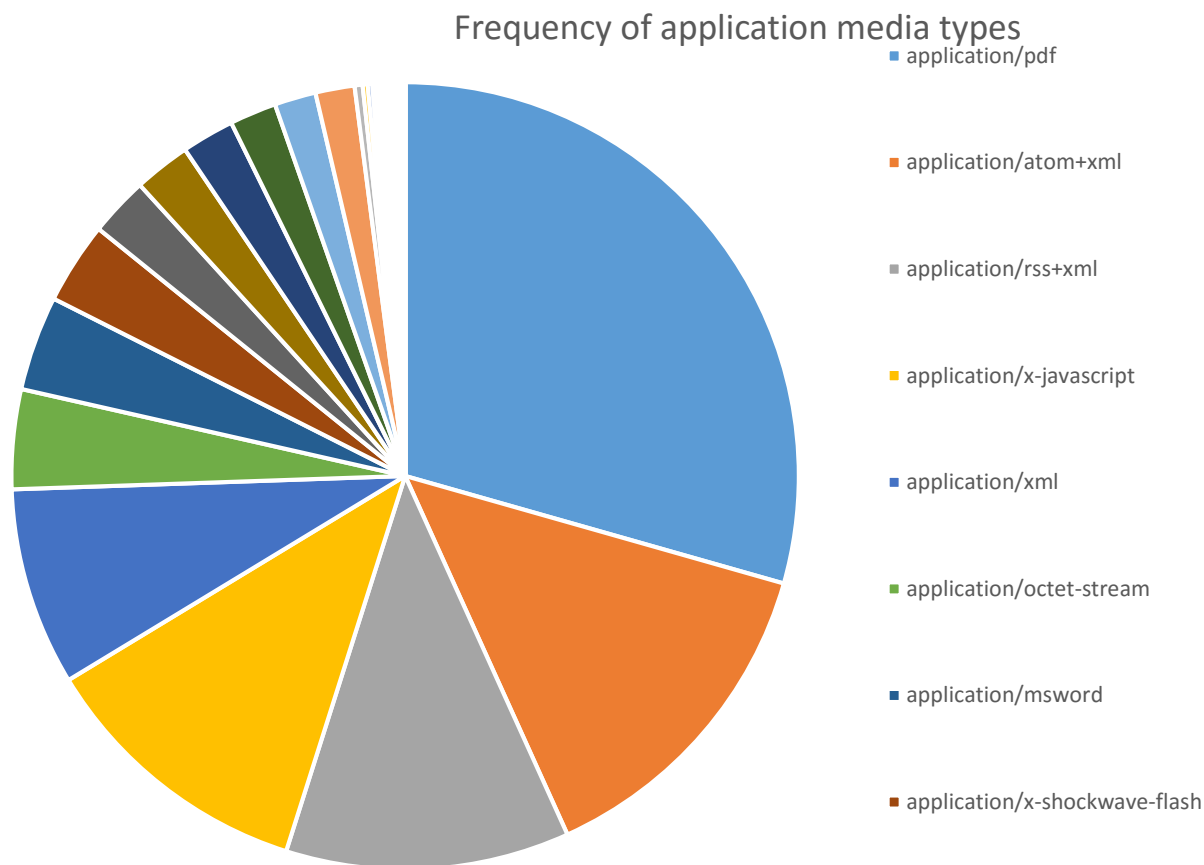
# Media types (MIME types)

- Restricted to 4 media type groups:
- application/\*
- image/\*
- text/\*
- video/\*
- Significant media types within these groups selected for investigation

# Media types - application

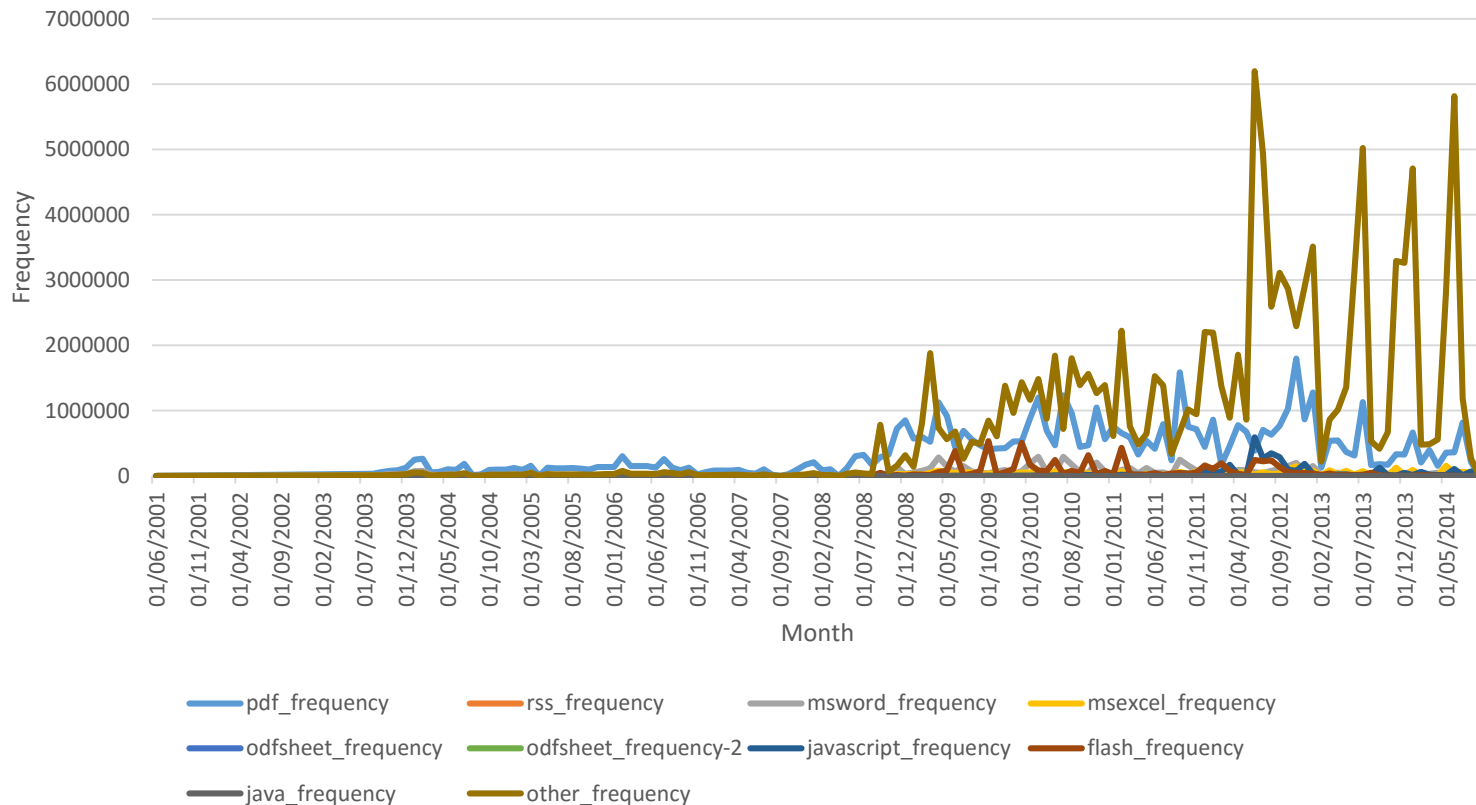
- application/x-shockwave-flash
- application/x-java
- application/java-byte-code
- application/javascript
- application/msword
- application/pdf
- application/rtf
- application/vnd.ms-excel
- application/xml
- application/zip

# Media types - application



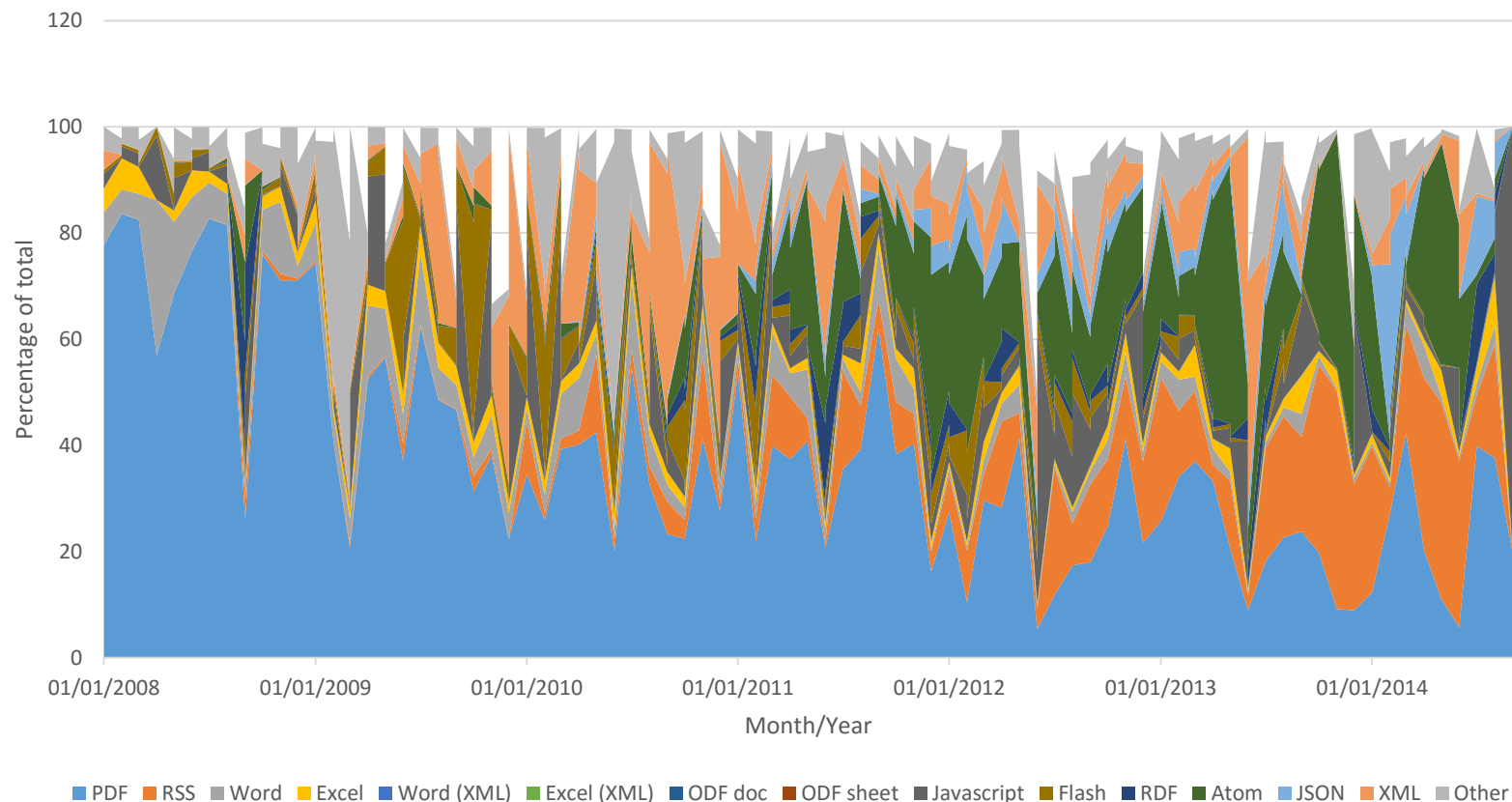
# Media types - application

Application media type frequency over time



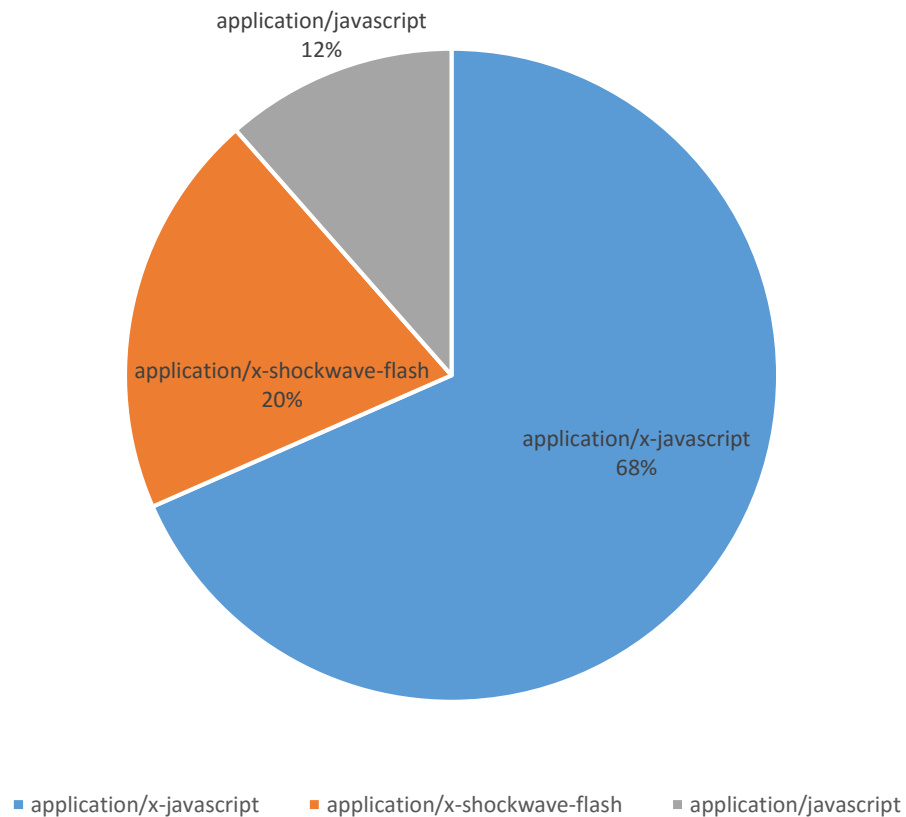
# Media types - application

Application media types as percentage of total, 2008-2014



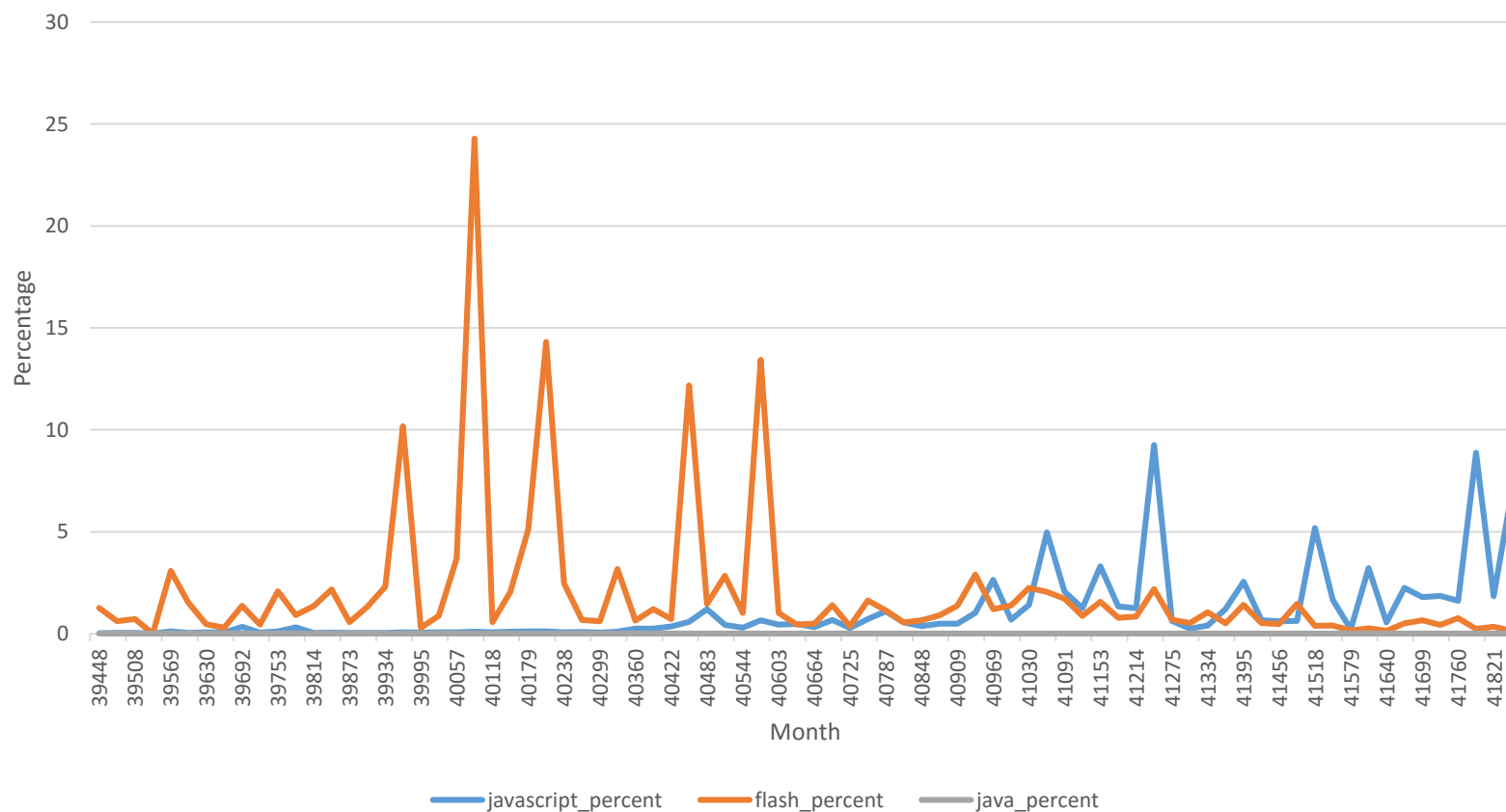
# Media types - executable

Executable media types



# Media types - executable

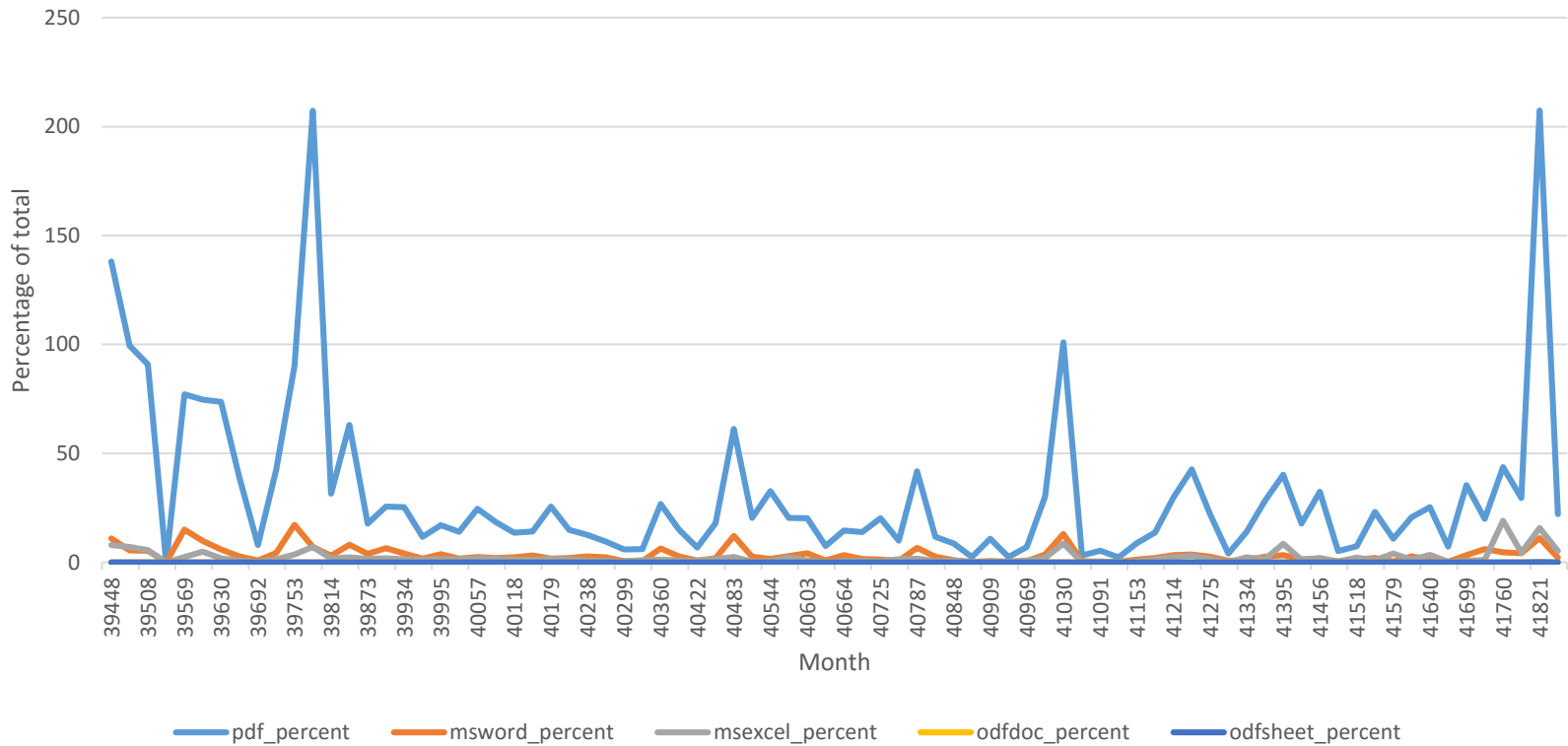
Executable content percentage over time





# Media types - document

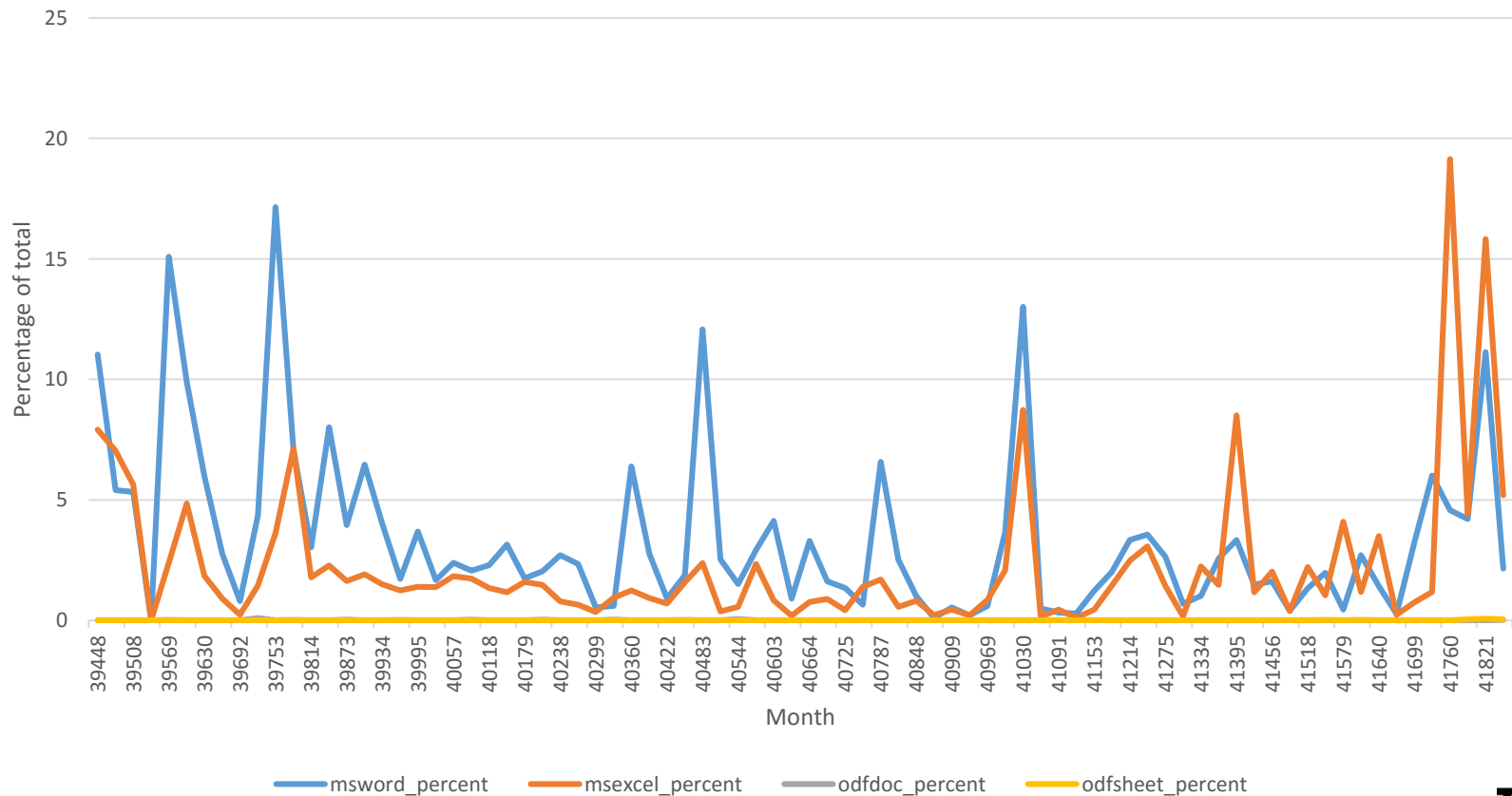
Document media types over time





# Media types - document

Document media types over time, excluding PDF

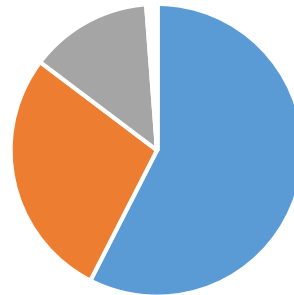


# Media types - image

- Consists of inline images appearing in documents, plus icons:
- image/jpeg
- image/png
- image/gif
- image/x-icon
- Occasional use of non-standard media type labels was ignored for this analysis

# Media types - image

Image media types

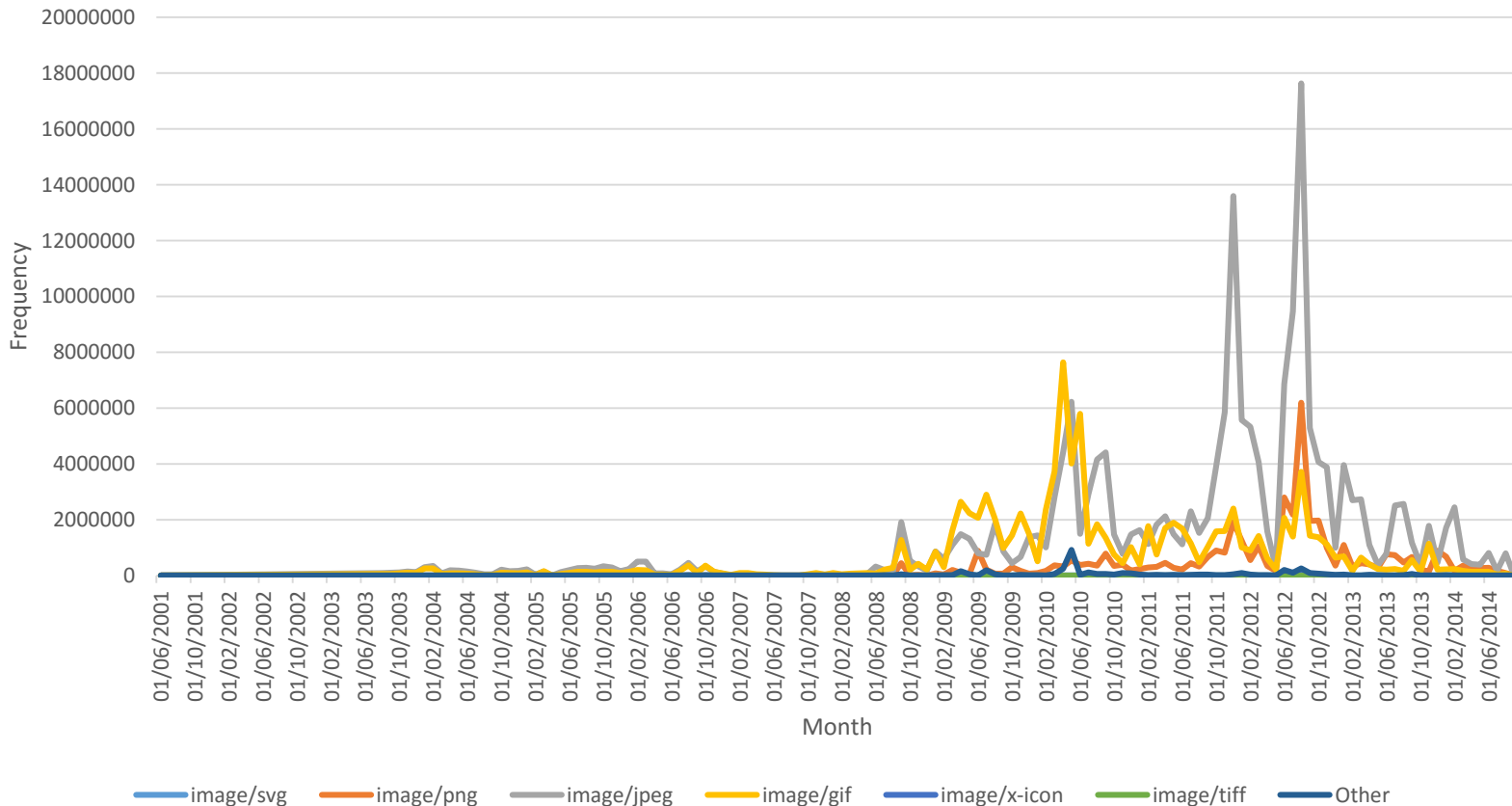


- image/jpeg
- image/gif
- image/png
- image/pjpeg
- image/jpg
- image/x-icon
- image/bmp
- image/svg+xml
- image/x-png
- image/vnd.microsoft.icon
- image/JPEG
- image/vnd.wap.wbmp
- image/\$inputFileExtension
- image/JPG
- image/tiff
- image/\*,%20image/gif
- image/.jpg



# Media types - image

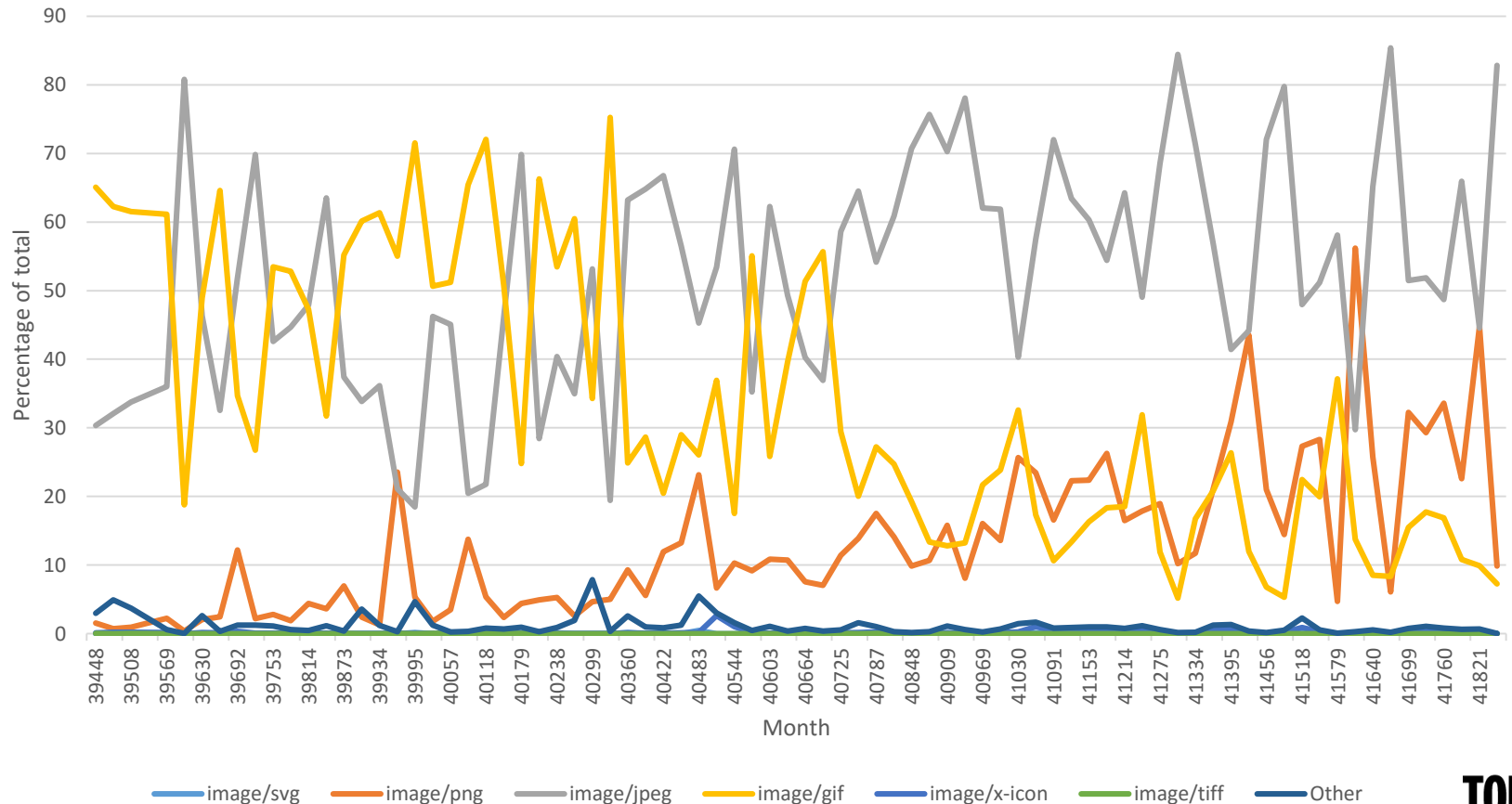
Frequencies of common image media types over time





# Media types - image

Image media types as percentage of total over time

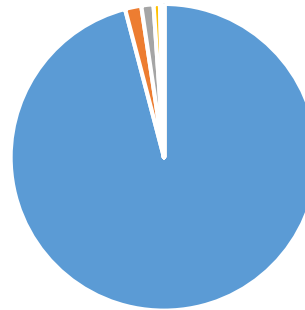


# Media types - text

- Plain text formats, including hypertext:
- text/plain
- text/html
- text/x-html
- Occasional use of non-standard media type labels was ignored for this analysis
- Entirely dominated by HTML

# Media types - text

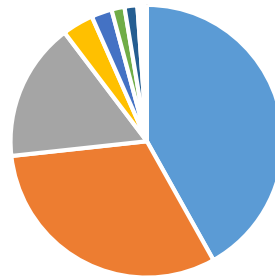
Text media types



- text/html
- text/xml
- text/calendar
- text/rdf+n3
- text/tab-separated-values
- text/x-cross-domain-policy
- text/HTML
- text/x-c
- text/Calendar
- text/x-csv
- text/plain
- text/javascript
- text/turtle
- text/vbscript
- text/rtf
- text/x-component
- text/enriched
- text/js
- text/vcard
- text/x-javascript
- text/css
- text/csv
- text/x-perl
- text/n3
- text/x-js
- text/x-vCalendar
- text/comma-separated-values
- text/x-vcard
- text/richtext
- text/x-ms-iqy

# Media types - text

Text media types (excluding HTML)



- |  |   |  |
|--|---|--|
| <span style="color: blue;">■</span> text/plain                         | <span style="color: orange;">■</span> text/css                        | <span style="color: grey;">■</span> text/xml                       |
| <span style="color: yellow;">■</span> text/javascript                  | <span style="color: blue;">■</span> text/csv                          | <span style="color: green;">■</span> text/calendar                 |
| <span style="color: darkblue;">■</span> text/turtle                    | <span style="color: brown;">■</span> text/x-perl                      | <span style="color: darkgrey;">■</span> text/rdf+n3                |
| <span style="color: darkred;">■</span> text/vbscript                   | <span style="color: darkblue;">■</span> text/n3                       | <span style="color: darkgreen;">■</span> text/tab-separated-values |
| <span style="color: lightblue;">■</span> text/rtf                      | <span style="color: orange;">■</span> text/x-js                       | <span style="color: grey;">■</span> text/x-cross-domain-policy     |
| <span style="color: yellow;">■</span> text/x-component                 | <span style="color: blue;">■</span> text/x-vCalendar                  | <span style="color: green;">■</span> text/HTML                     |
| <span style="color: blue;">■</span> text/enriched                      | <span style="color: orange;">■</span> text/comma-separated-values     | <span style="color: darkgrey;">■</span> text/x-c                   |
| <span style="color: darkred;">■</span> text/js                         | <span style="color: blue;">■</span> text/x-vcard                      | <span style="color: darkgreen;">■</span> text/Calendar             |
| <span style="color: lightblue;">■</span> text/vcard                    | <span style="color: orange;">■</span> text/richtext                   | <span style="color: grey;">■</span> text/x-csv                     |
| <span style="color: yellow;">■</span> text/x-javascript                | <span style="color: lightblue;">■</span> text/x-ms-iqy                | <span style="color: lightgreen;">■</span> text/x-patch             |
| <span style="color: darkblue;">■</span> text/json                      | <span style="color: brown;">■</span> text/x-c++                       | <span style="color: darkgrey;">■</span> text/x-vCard               |
| <span style="color: darkred;">■</span> text/directory                  | <span style="color: darkblue;">■</span> text/x-comma-separated-values | <span style="color: darkgreen;">■</span> text/rtf2                 |
| <span style="color: lightblue;">■</span> text/vnd.wap.wml              | <span style="color: orange;">■</span> text/htm                        | <span style="color: grey;">■</span> text/x-java                    |
| <span style="color: yellow;">■</span> text/x-json                      | <span style="color: blue;">■</span> text/JavaScript                   | <span style="color: lightgreen;">■</span> text/ecmascript          |
| <span style="color: darkblue;">■</span> text/plain,%20charset:%20UTF-8 | <span style="color: brown;">■</span> text/text                        | <span style="color: darkgrey;">■</span> text/x-python              |
| <span style="color: darkred;">■</span> text/x-calendar                 | <span style="color: blue;">■</span> text/html%20charset=iso-8859-1    | <span style="color: darkgreen;">■</span> text/troff                |
| <span style="color: lightblue;">■</span> text/XML                      | <span style="color: orange;">■</span> text/rdf                        | <span style="color: grey;">■</span> text/x-fortran                 |
| <span style="color: yellow;">■</span> text/dtd                         | <span style="color: lightblue;">■</span> text/css,%20charset:%20UTF-8 | <span style="color: lightgreen;">■</span> text/fragment            |
| <span style="color: blue;">■</span> text/x-handlebars-template         | <span style="color: orange;">■</span> text/lrc                        | <span style="color: darkgrey;">■</span> text/illegal               |

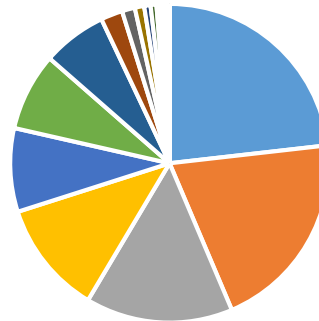


# Media types - video

- Compressed video formats:
- video/x-msvideo
- video/mpeg
- video/x-flv
- video/mp4
- Largely superseded by embedded YouTube links

# Media types - video

Video media types by total frequency



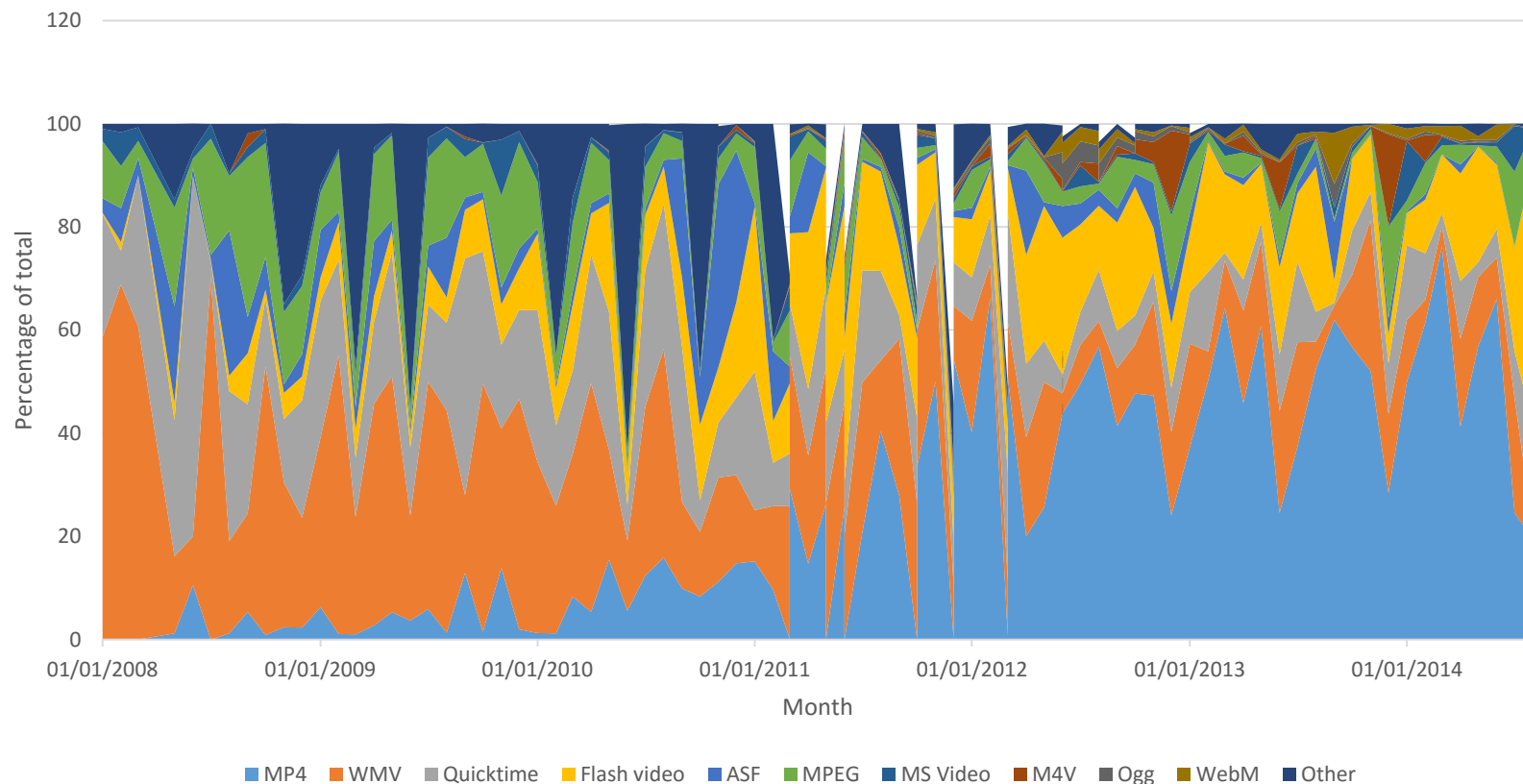
- video/mp4
- video/x-flv
- video/mpeg
- video/x-m4v
- video/3gpp
- video/x-frv
- video/mp4v-es
- video/m4v
- video/mpeg4
- video/x-flv%20.flv
- video/mp4v
- video/x-FLV
- video/asf
- video/x-download-quicktime
- video/x-ms-wm
- video/mpg4

- video/x-ms-wmv
- video/x-ms-asf
- video/x-ms-wvx
- video/ogg
- video/x-ms-asx
- video/unknown
- video/avi
- video/x-f4v
- video/x-mpeg
- video/wmv
- video/x-ms-wmv%20video/quicktime
- video/ogv
- video/MP4
- video/x-unknown
- video/f4m
- video/shockwave

- video/quicktime
- video/x-ms-wmx
- video/x-msvideo
- video/webm
- video/x-mp4
- video/flv
- video/vnd.objectvideo
- video/mpv
- video/3gpp,%20audio/3gpp
- video/x-sgi-movie
- video/msvideo
- video/dl
- video/swf
- video/mpg
- video/h264
- video/f4v

# Media types - video

Video media types as percentage of category total

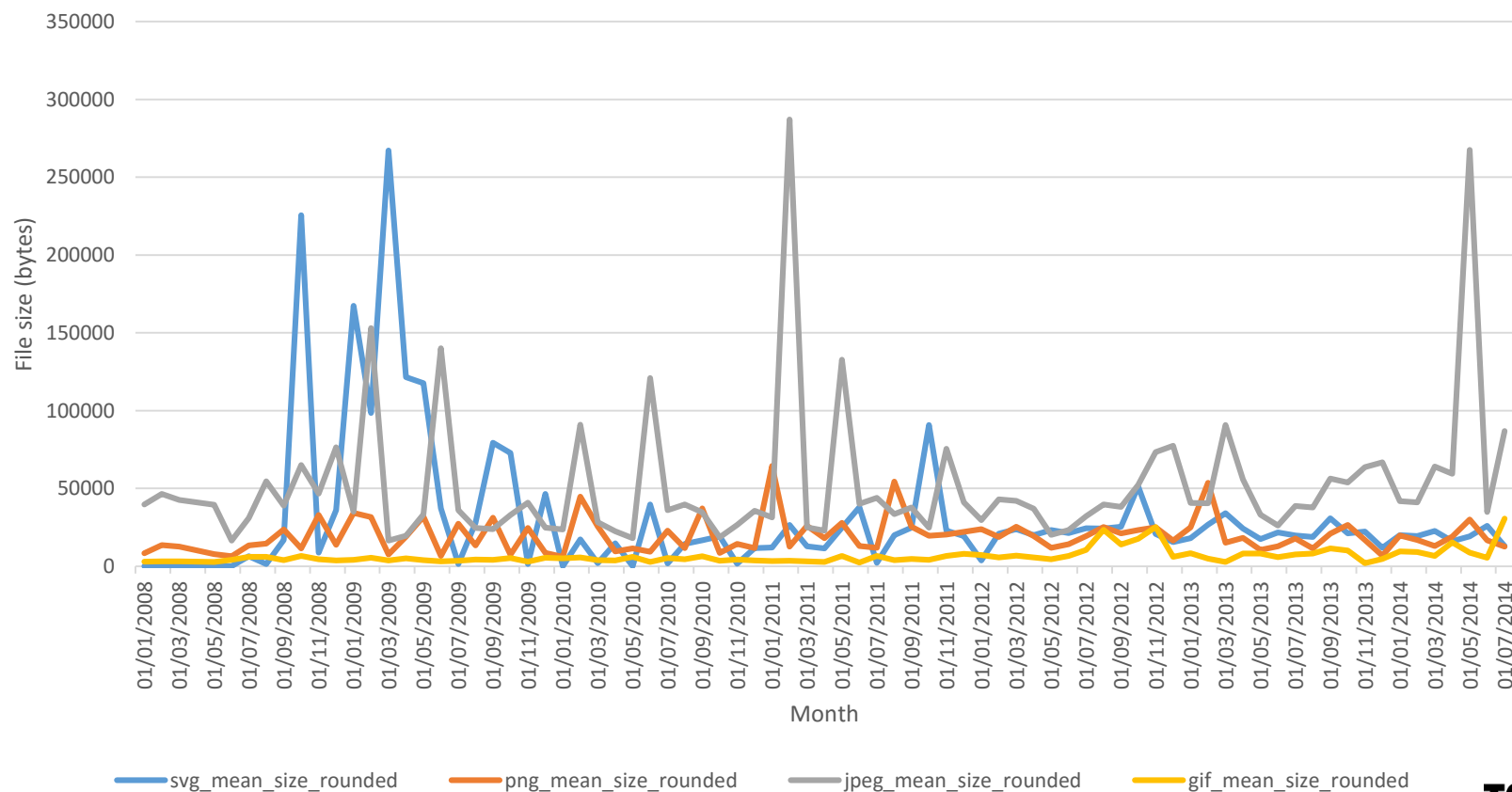


# File size over time

- Compressed image formats only
- This is due to variable compressibility of uncompressed images, documents, text, etc.
- CDX index only contains compressed size data and therefore is not a true representation of file size trends

# File size over time - images

Mean image file size over time

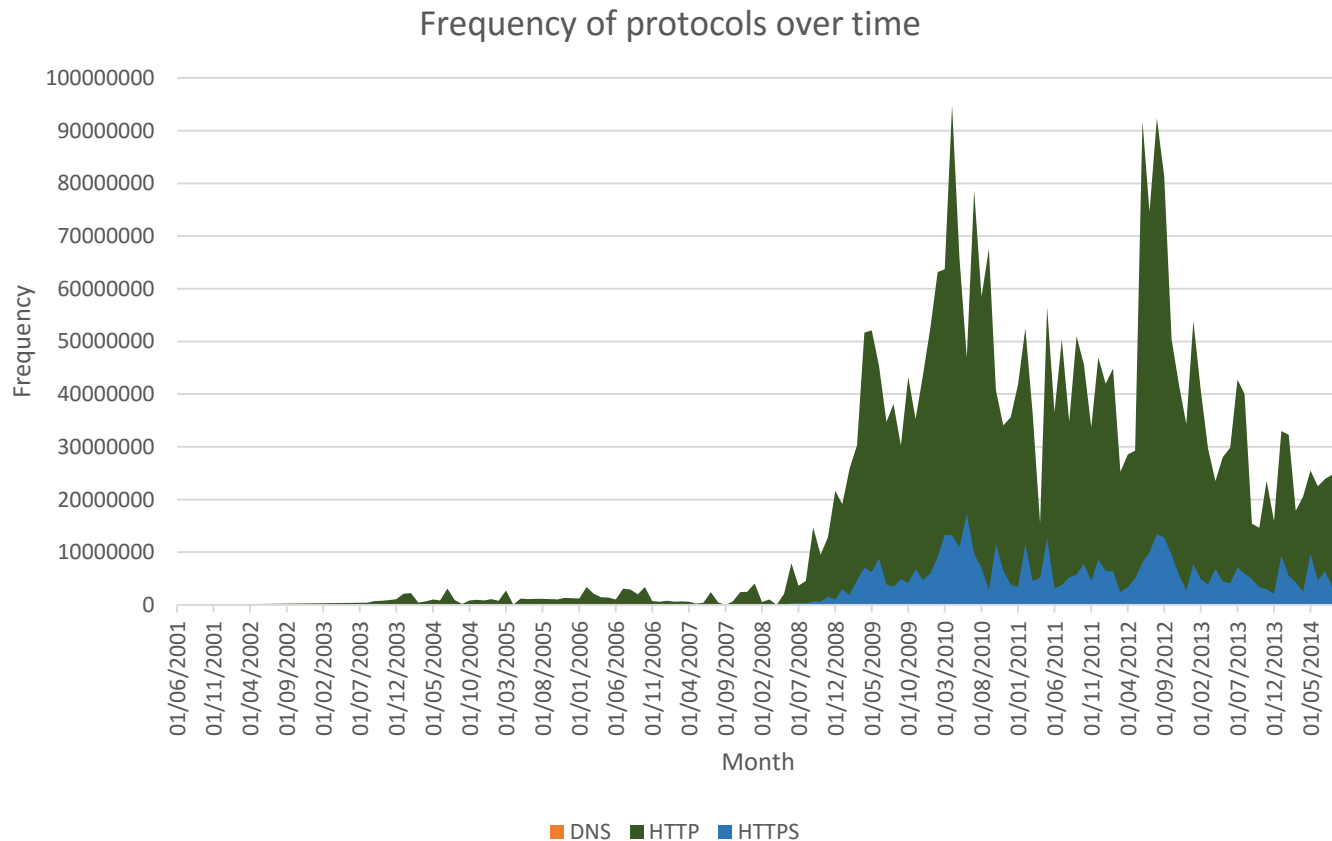


# Protocol changes

- CDX contains protocol information within URL parameters
- Protocol can be extracted from this parameter and aggregated temporally
- This reveals popularity trends for protocols
- HTTP vs. HTTPS



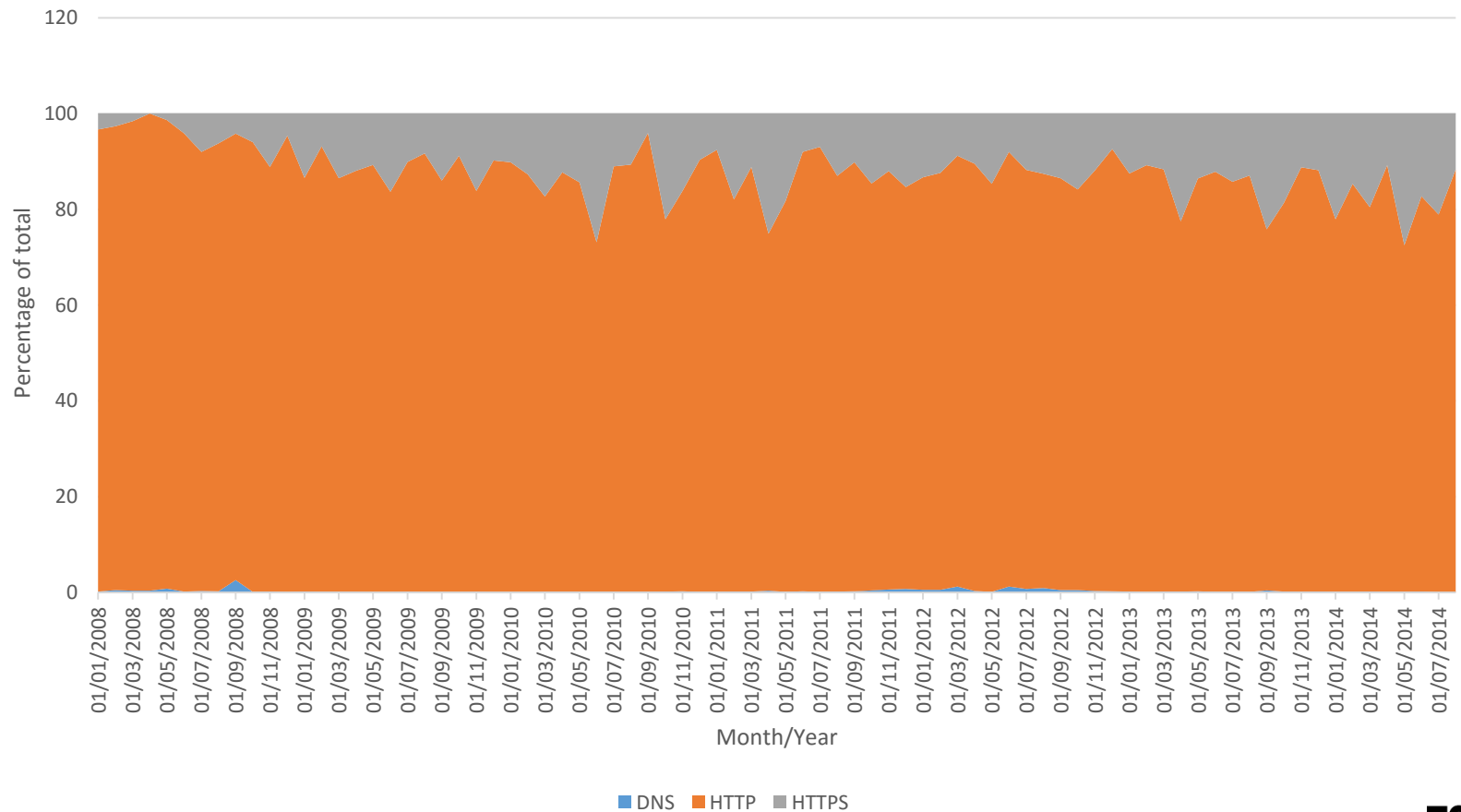
# HTTP vs HTTPS (absolute)





# HTTP vs HTTPS (relative)

Protocols as percentage of total, 2008-2014





# Conclusions

- It is possible to perform useful temporal analysis of CDX index data
- Transformation is necessary – SQL is feasible, commonly available and low cost
- Archive data has particular weaknesses – data cannot be assumed to be fully representative of the content of the target Web subset
- Even with this noise, trends can be identified