The University Of Sheffield.

A temporal overview of TNA's CDX index

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

A WORLD TOP 100 UNIVERSITY

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)

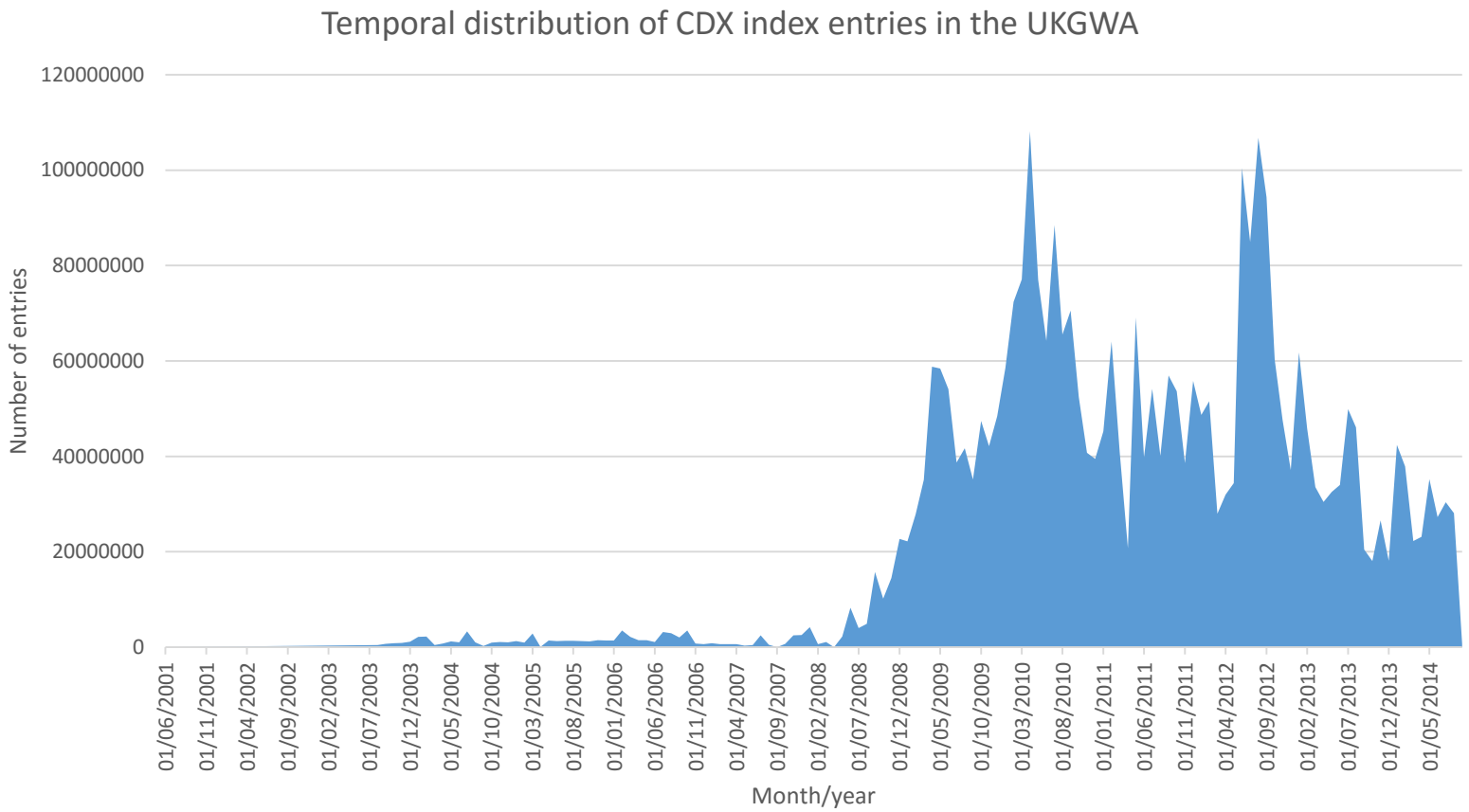
CDX processing

- 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

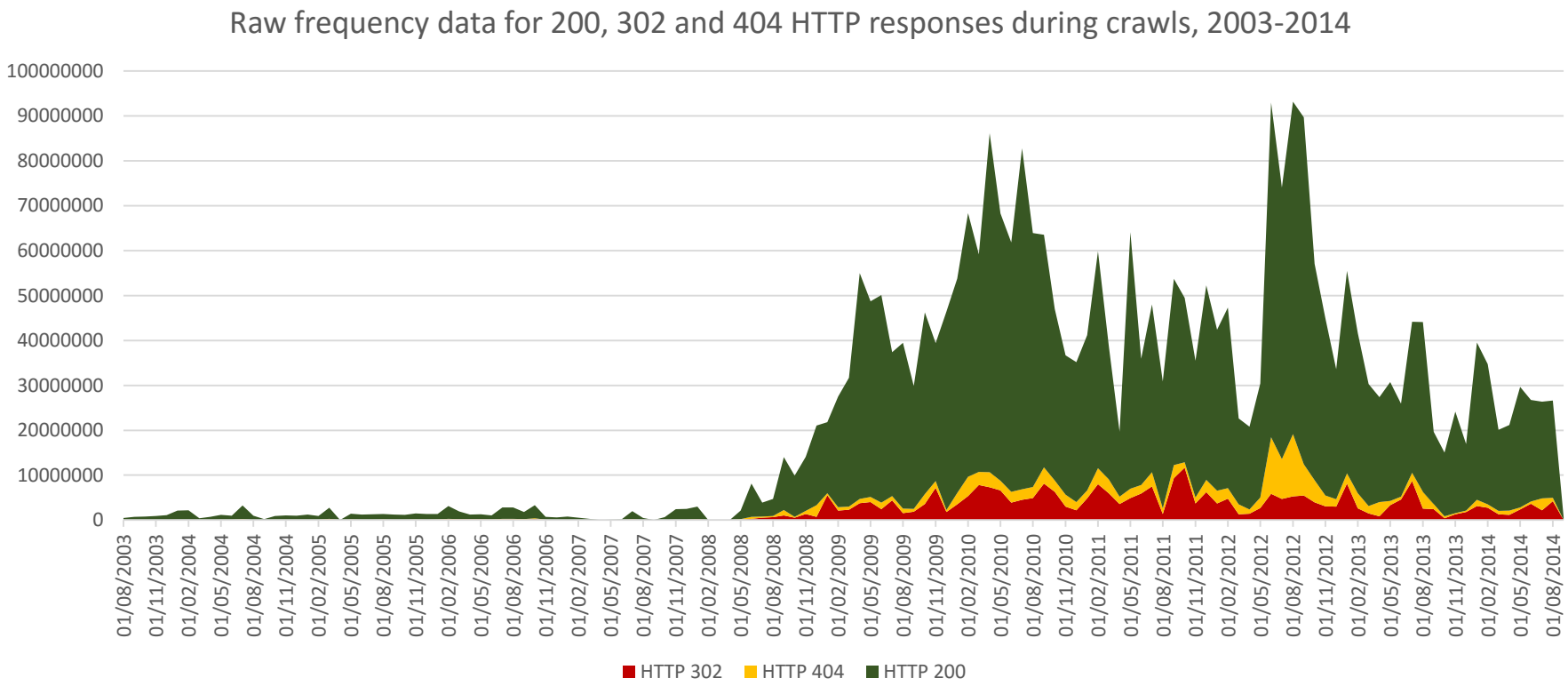
Archive overview



HTTP status codes

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

HTTP status codes (absolute)



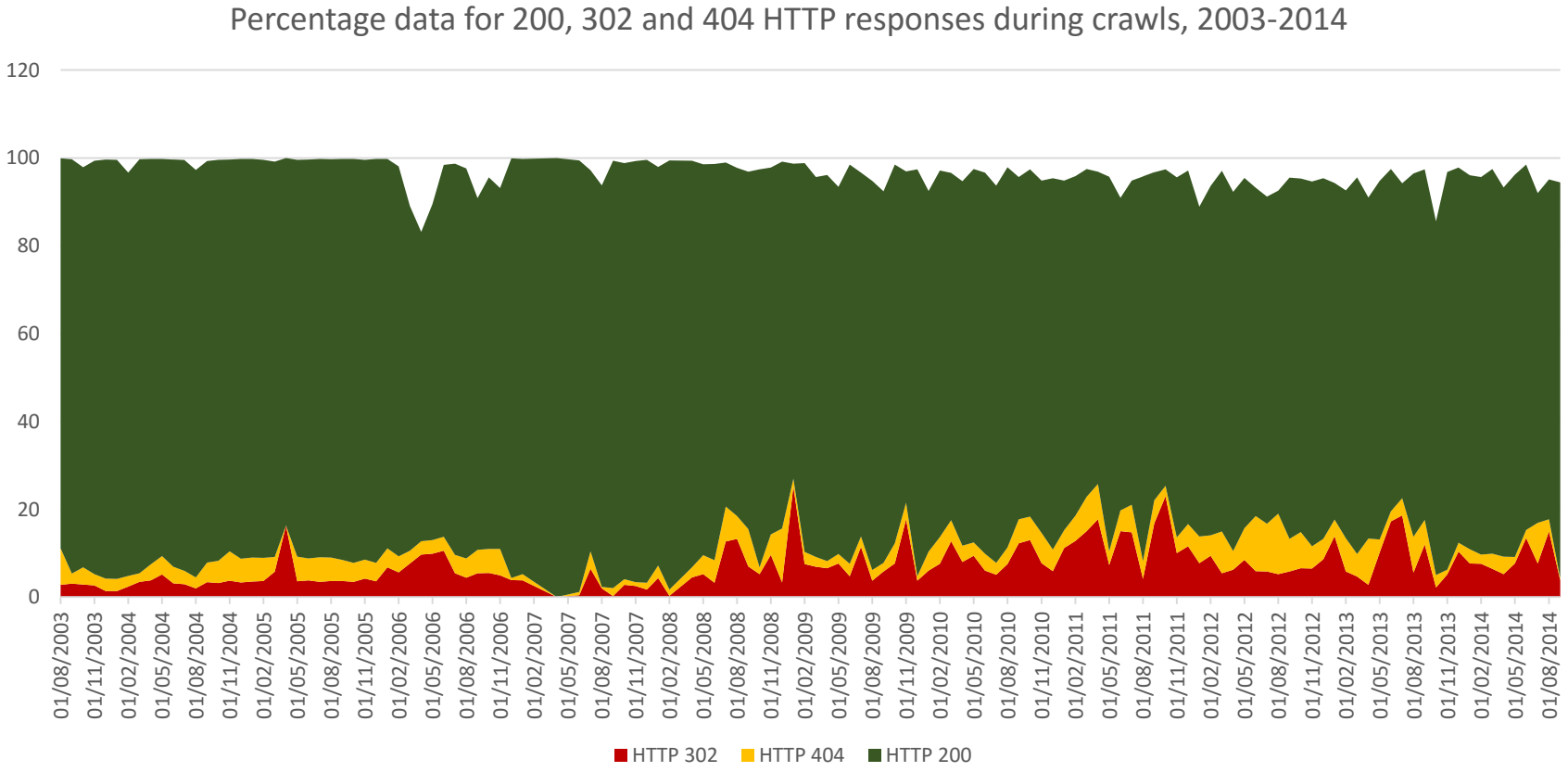
HTTP status codes (absolute)

- 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

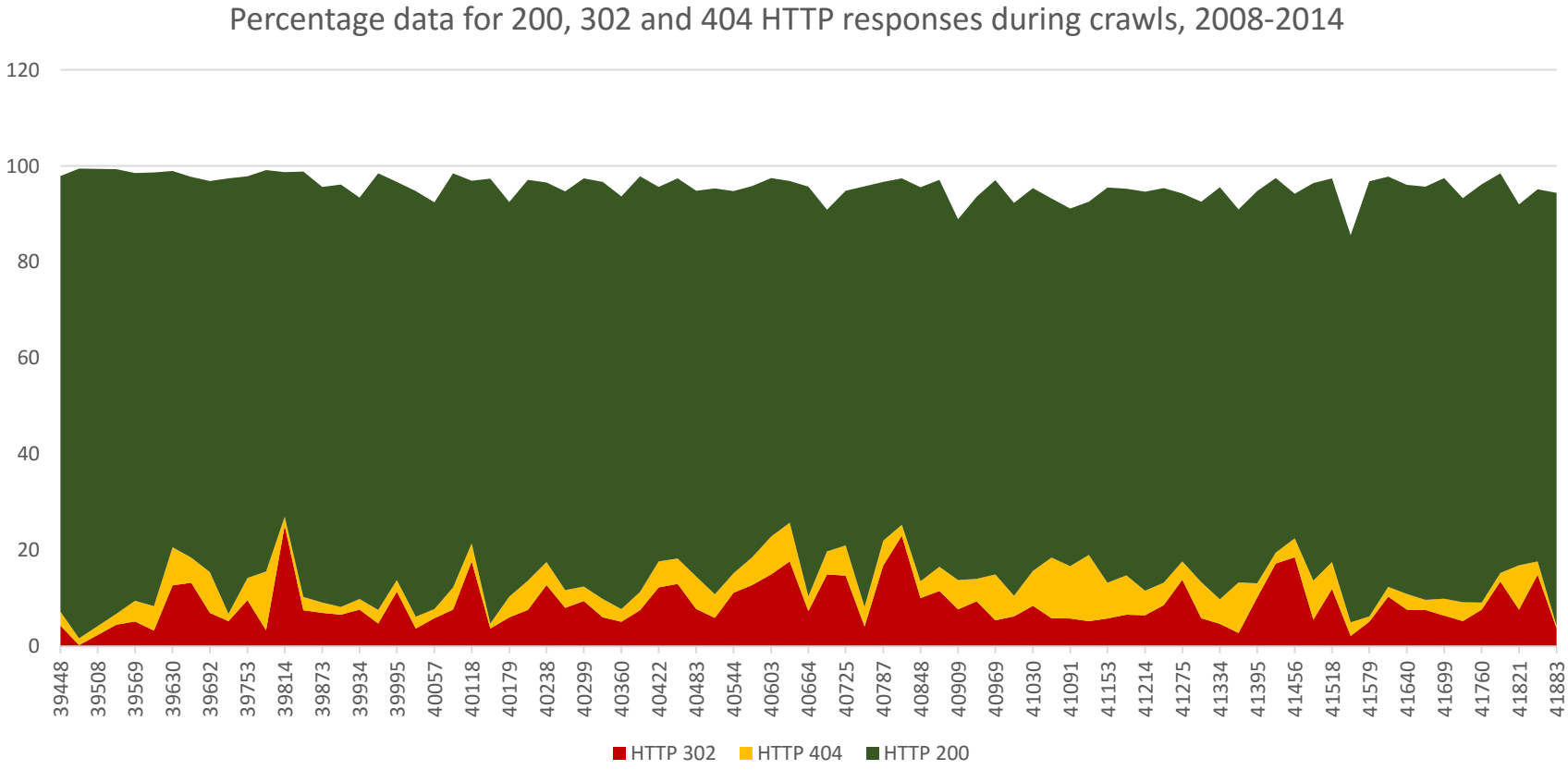
HTTP status codes (relative)

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

HTTP status codes (relative)



Post-2008 HTTP status codes



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.

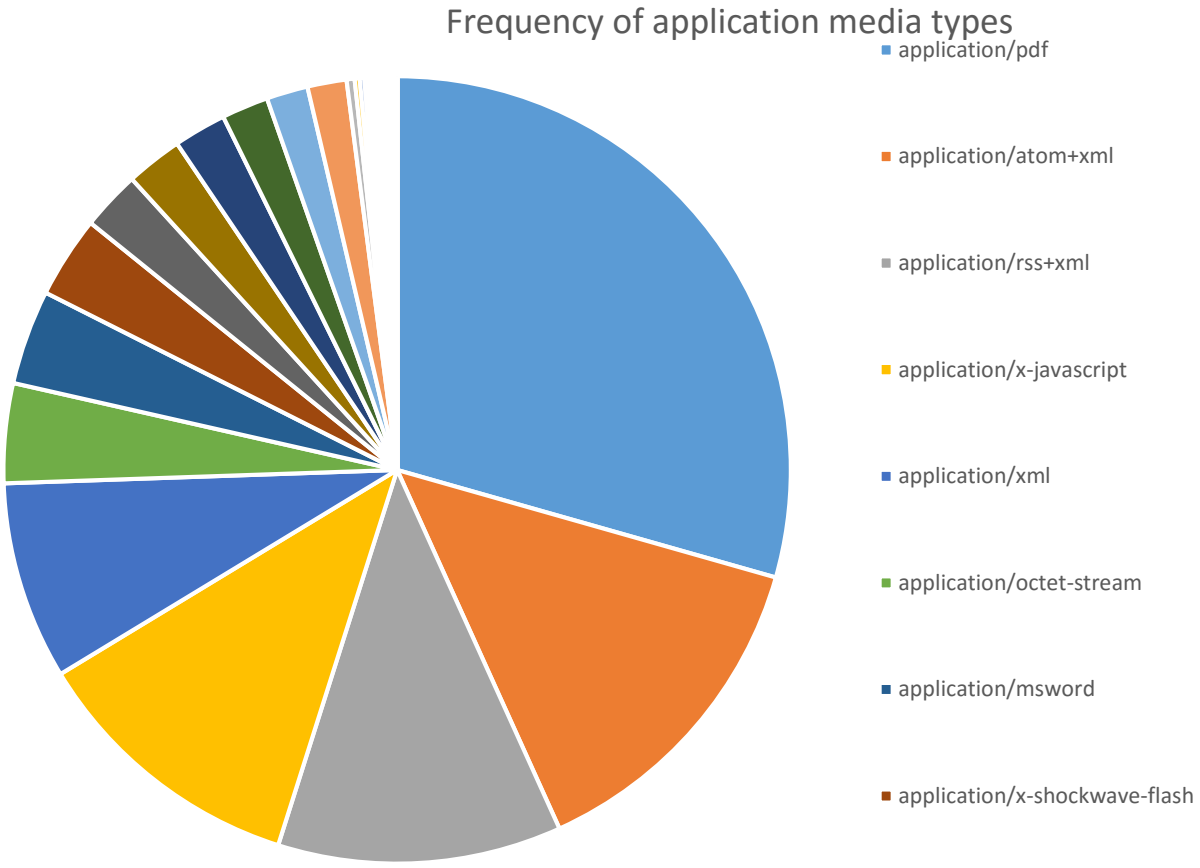
Media types (MIME types)

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

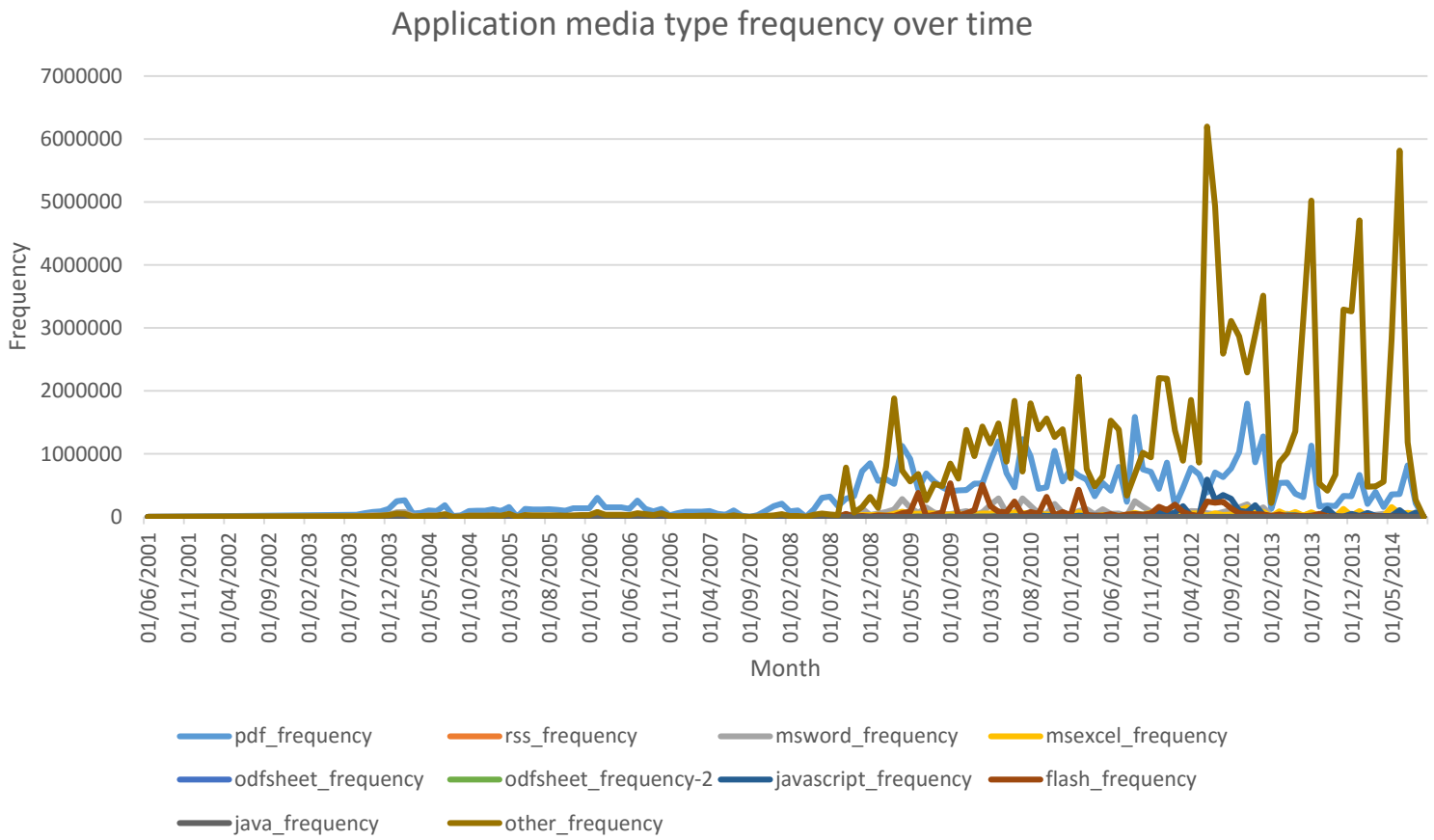
Media types - application

- 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

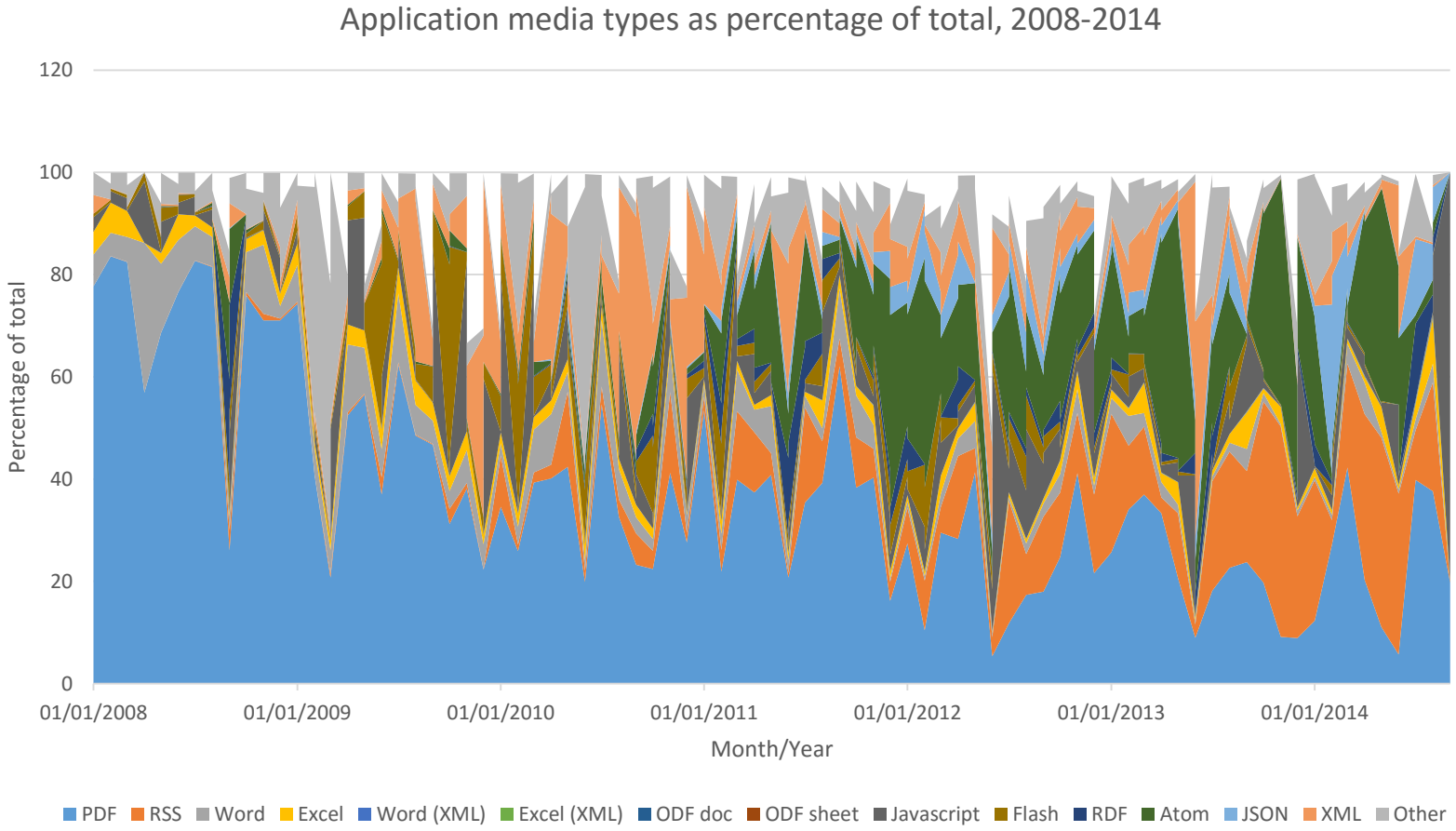
Media types - application



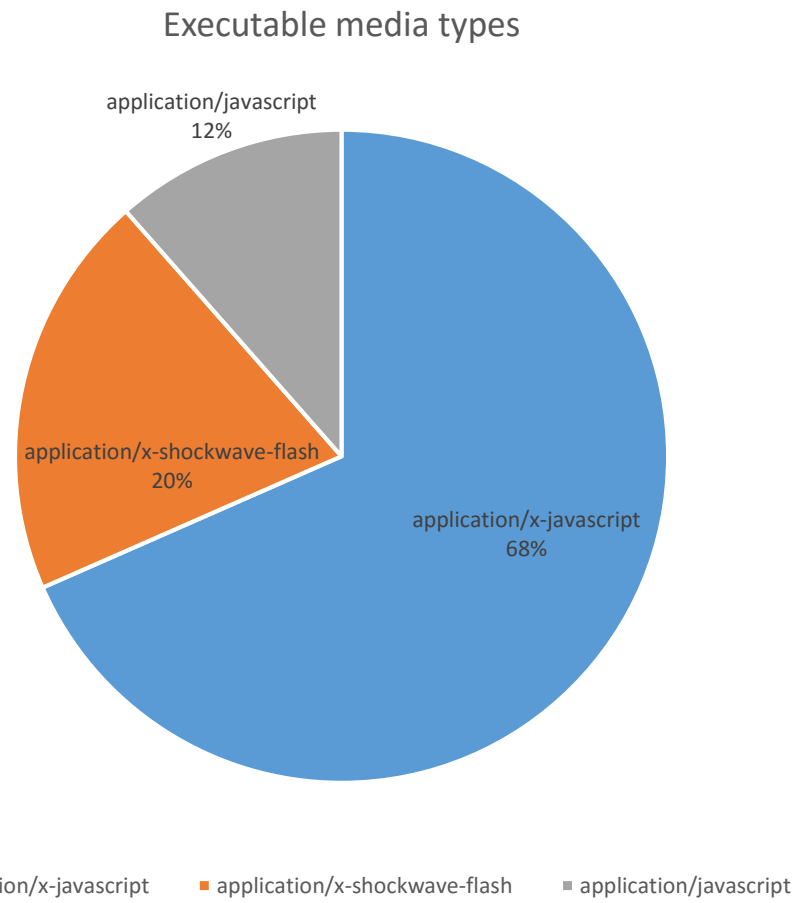
Media types - application



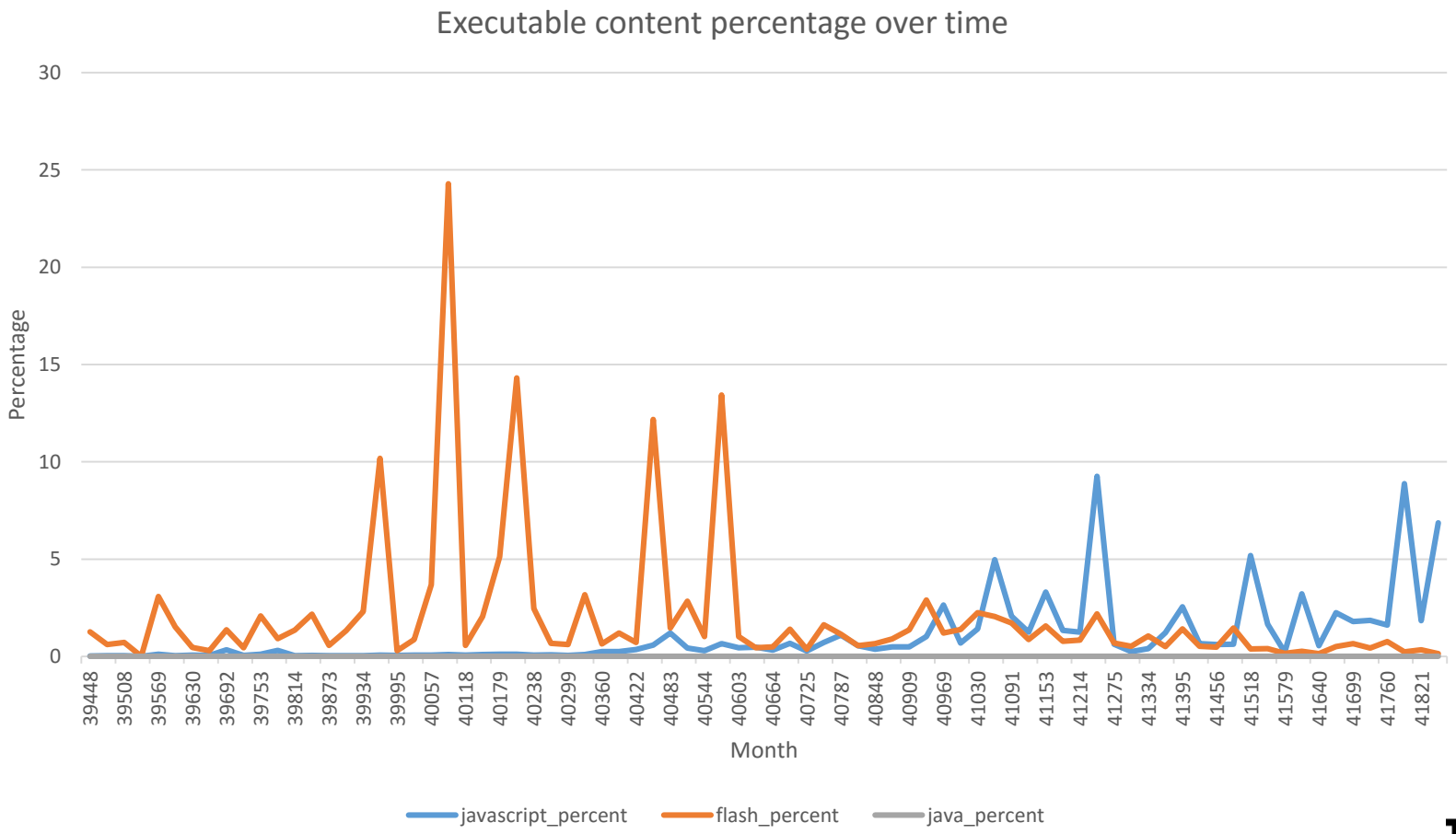
Media types - application



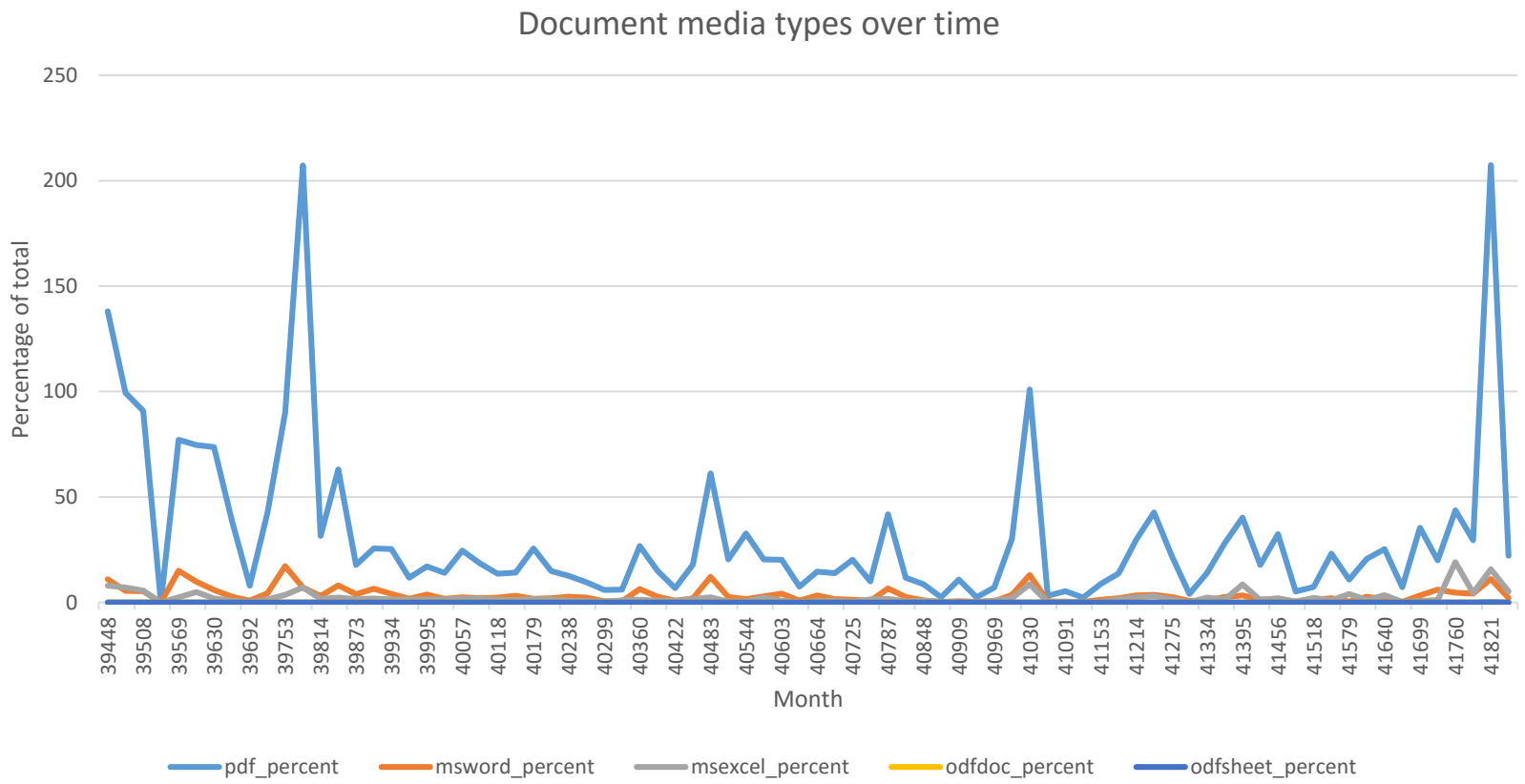
Media types - executable



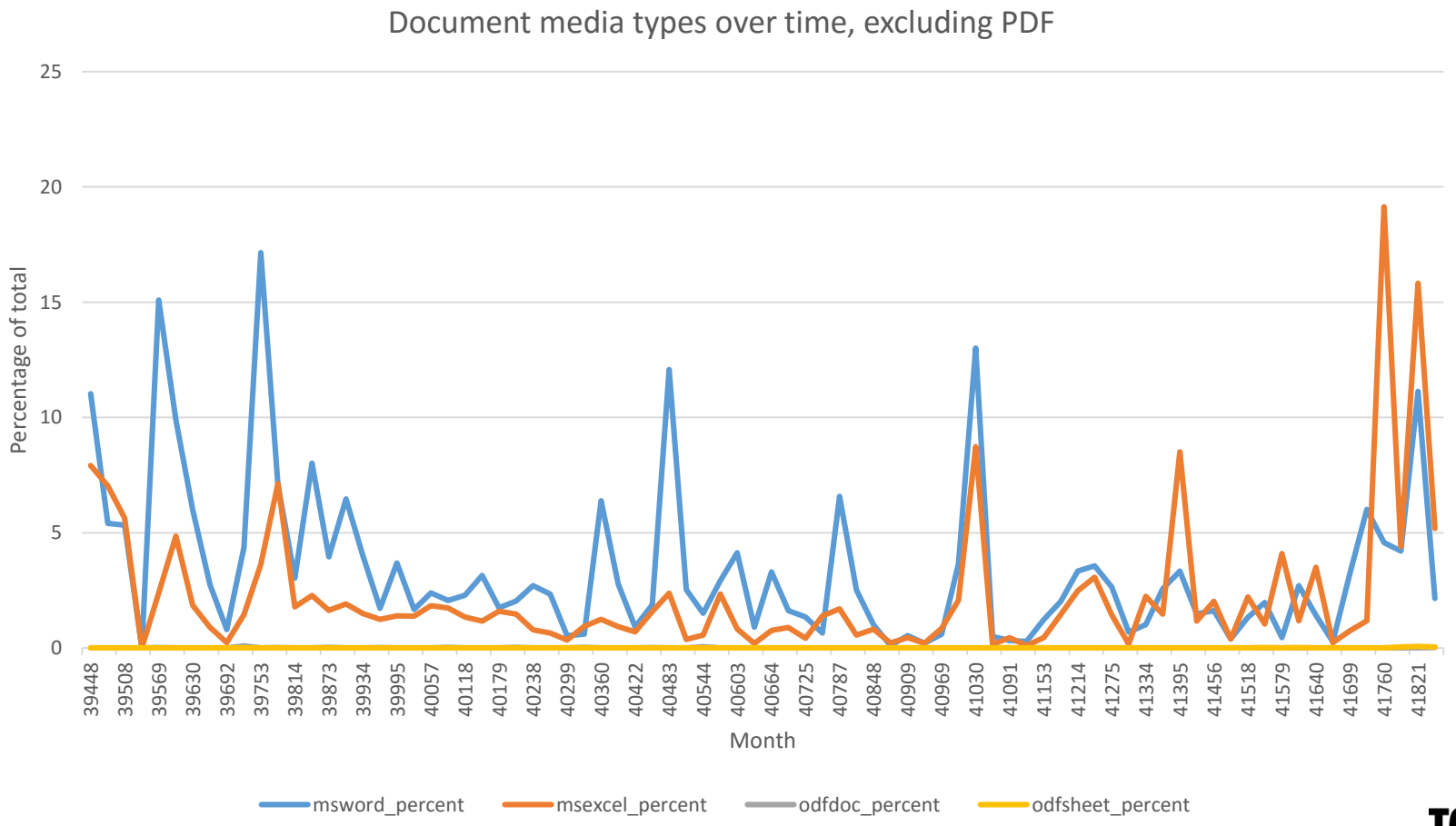
Media types - executable



Media types - document



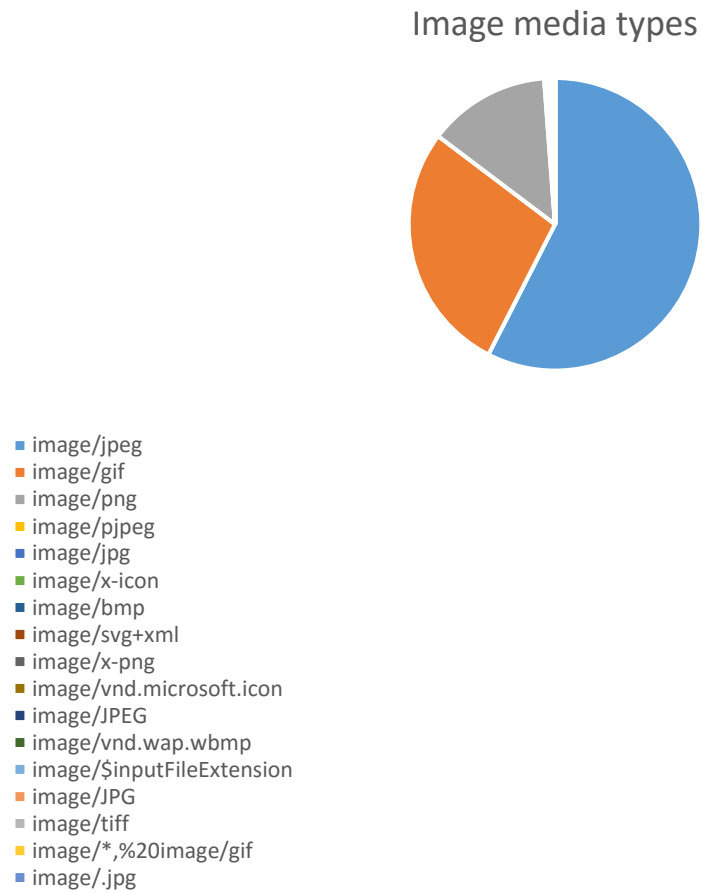
Media types - document



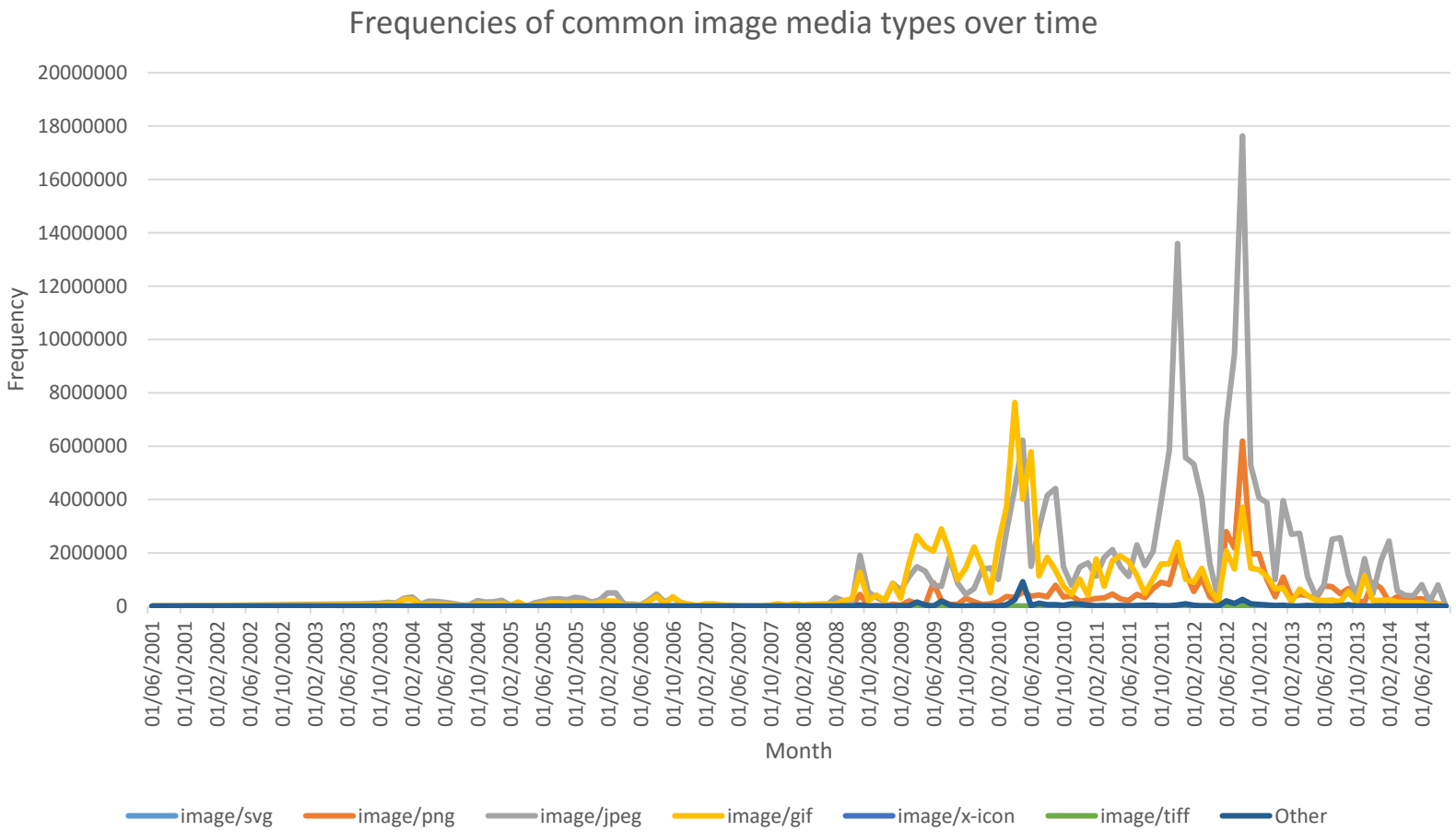
Media types - image

- 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

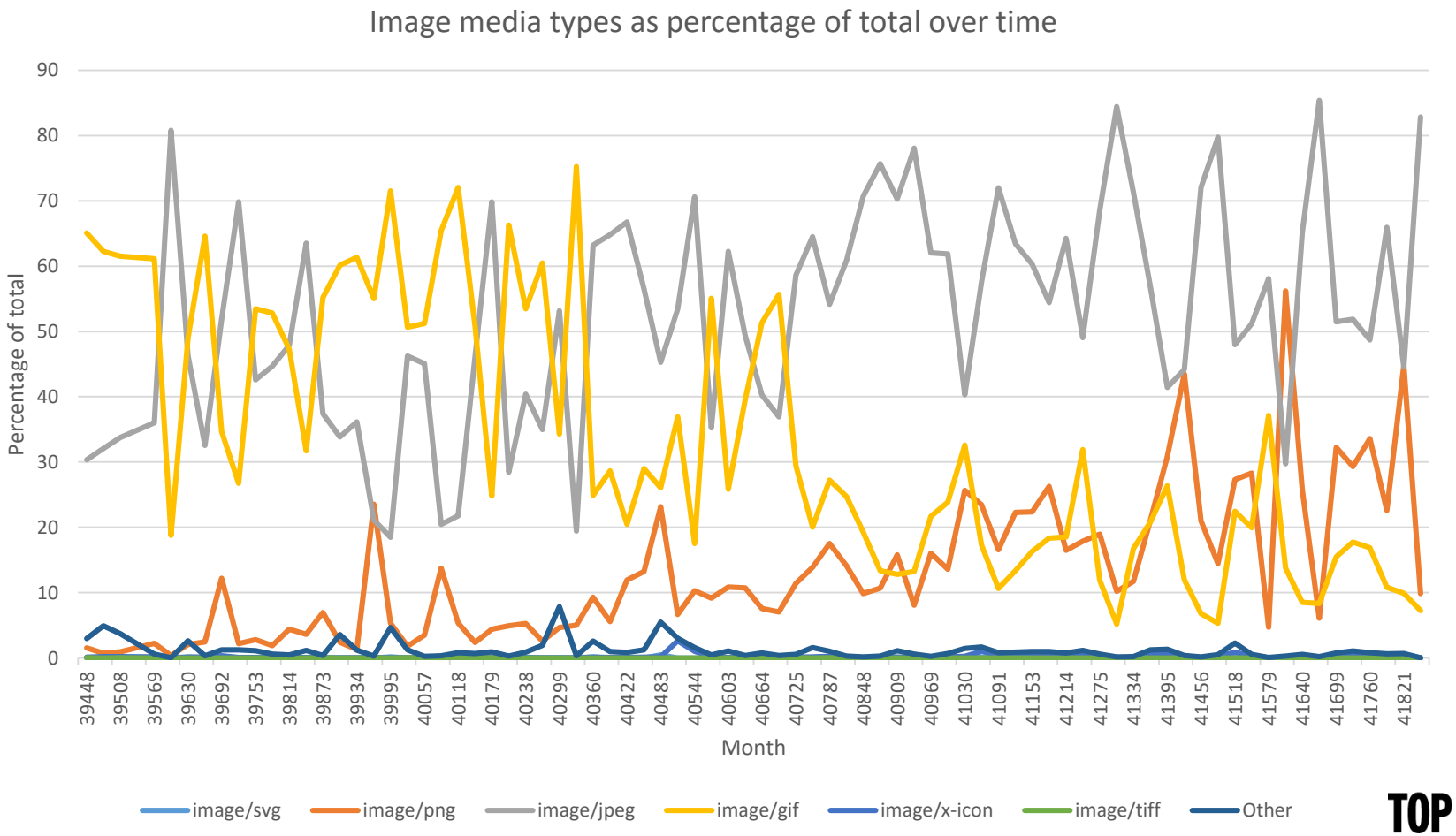
Media types - image



Media types - image



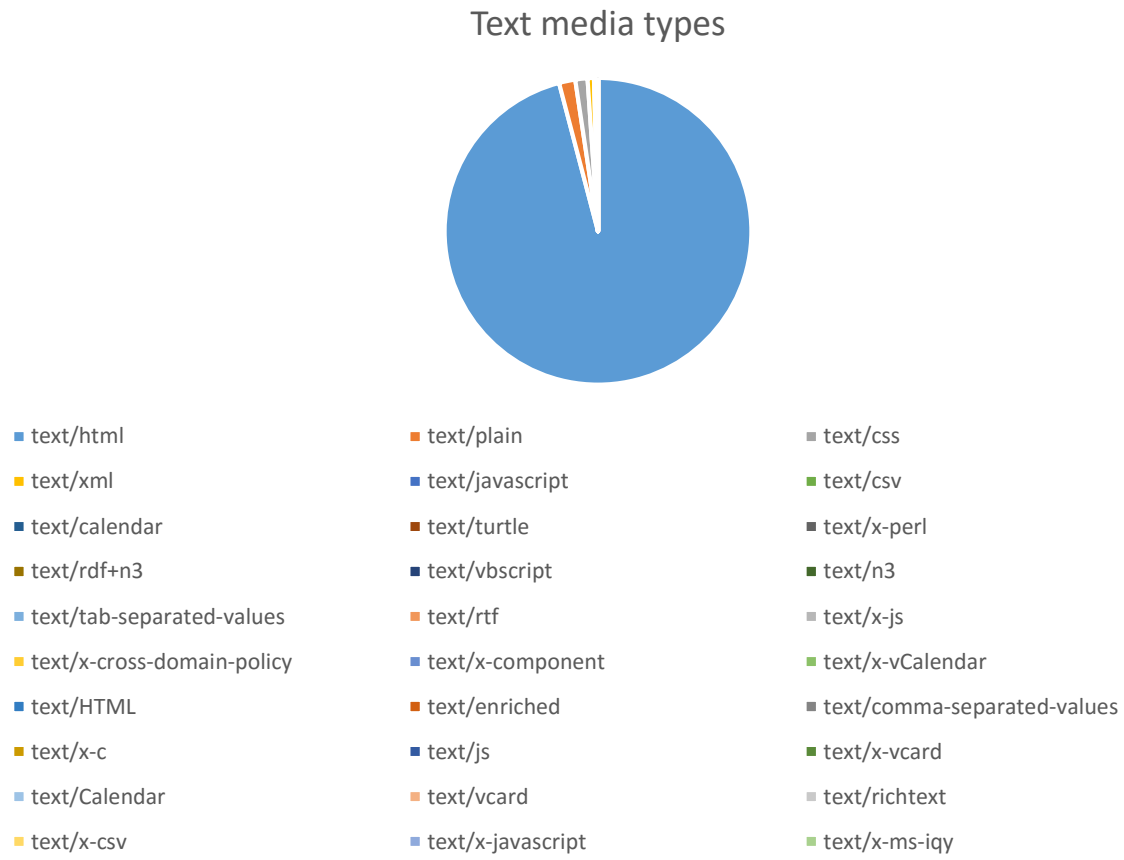
Media types - image



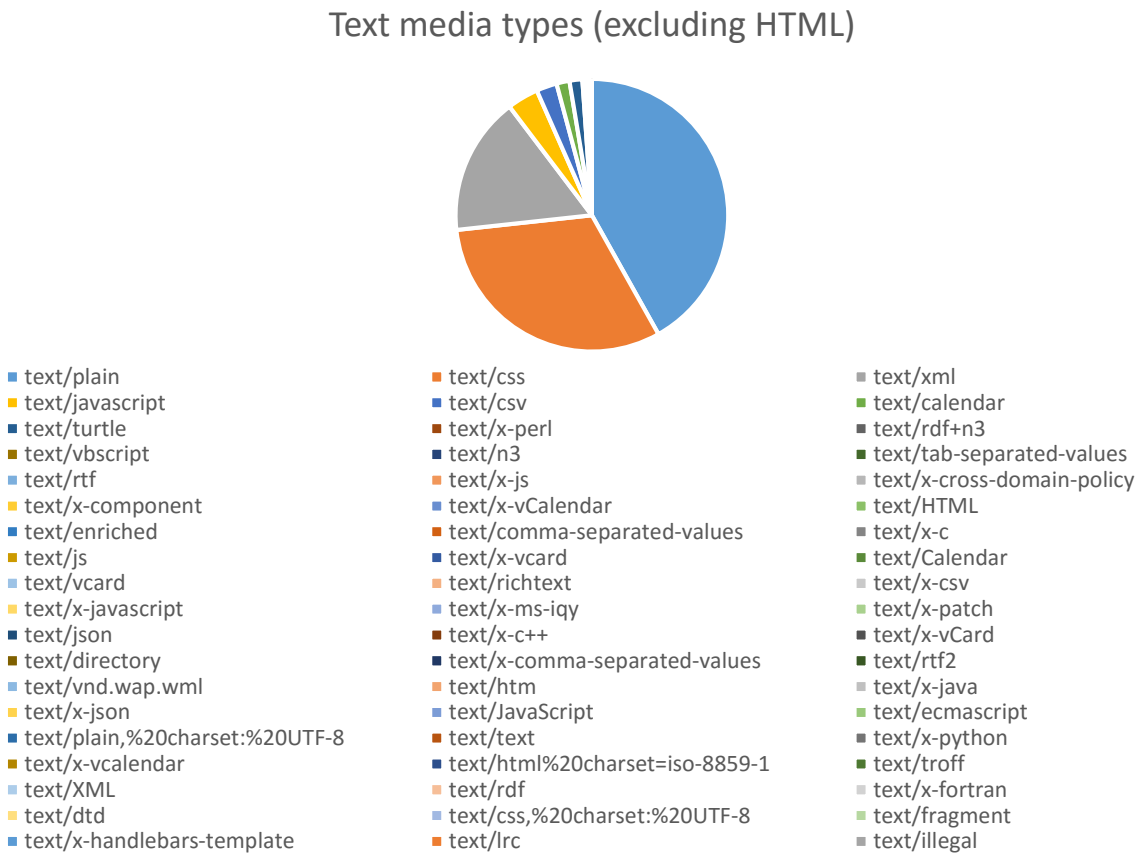
Media types - text

- 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



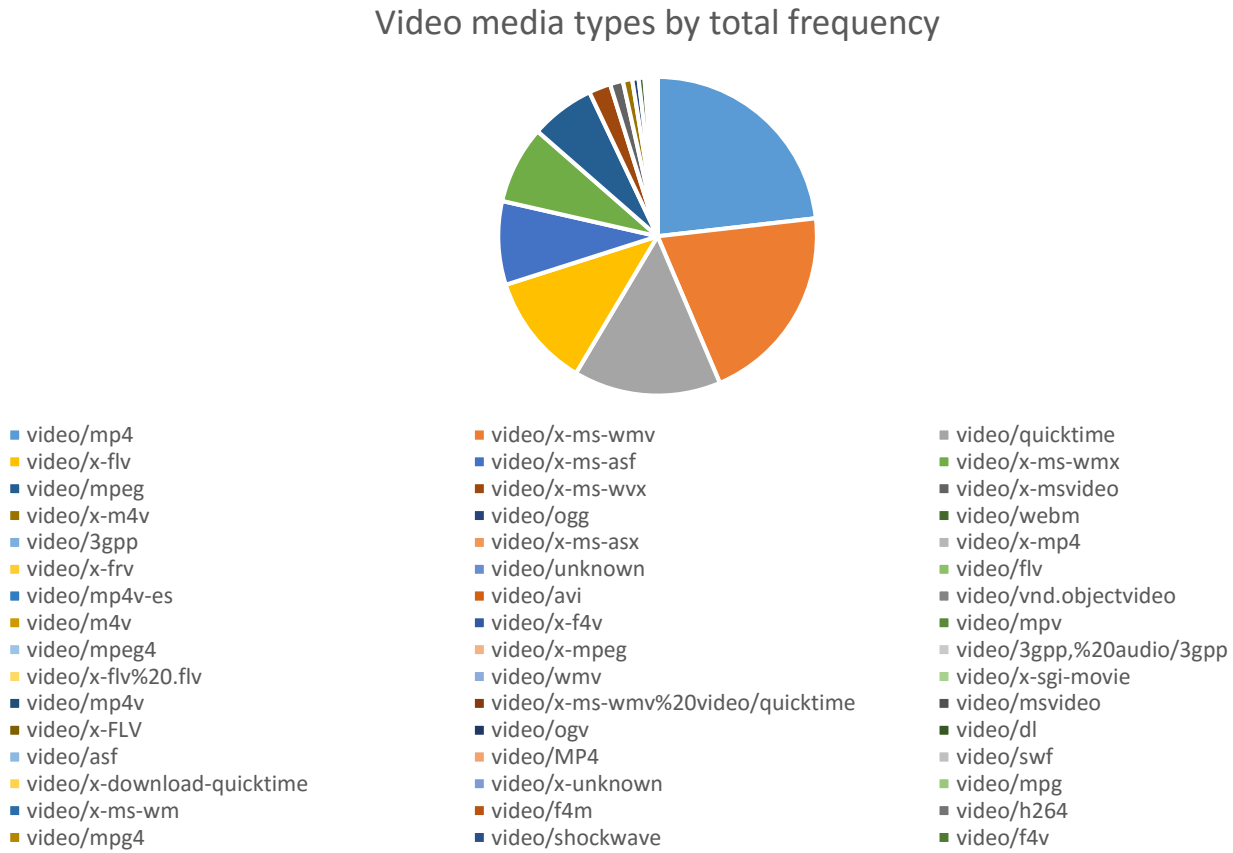
Media types - text



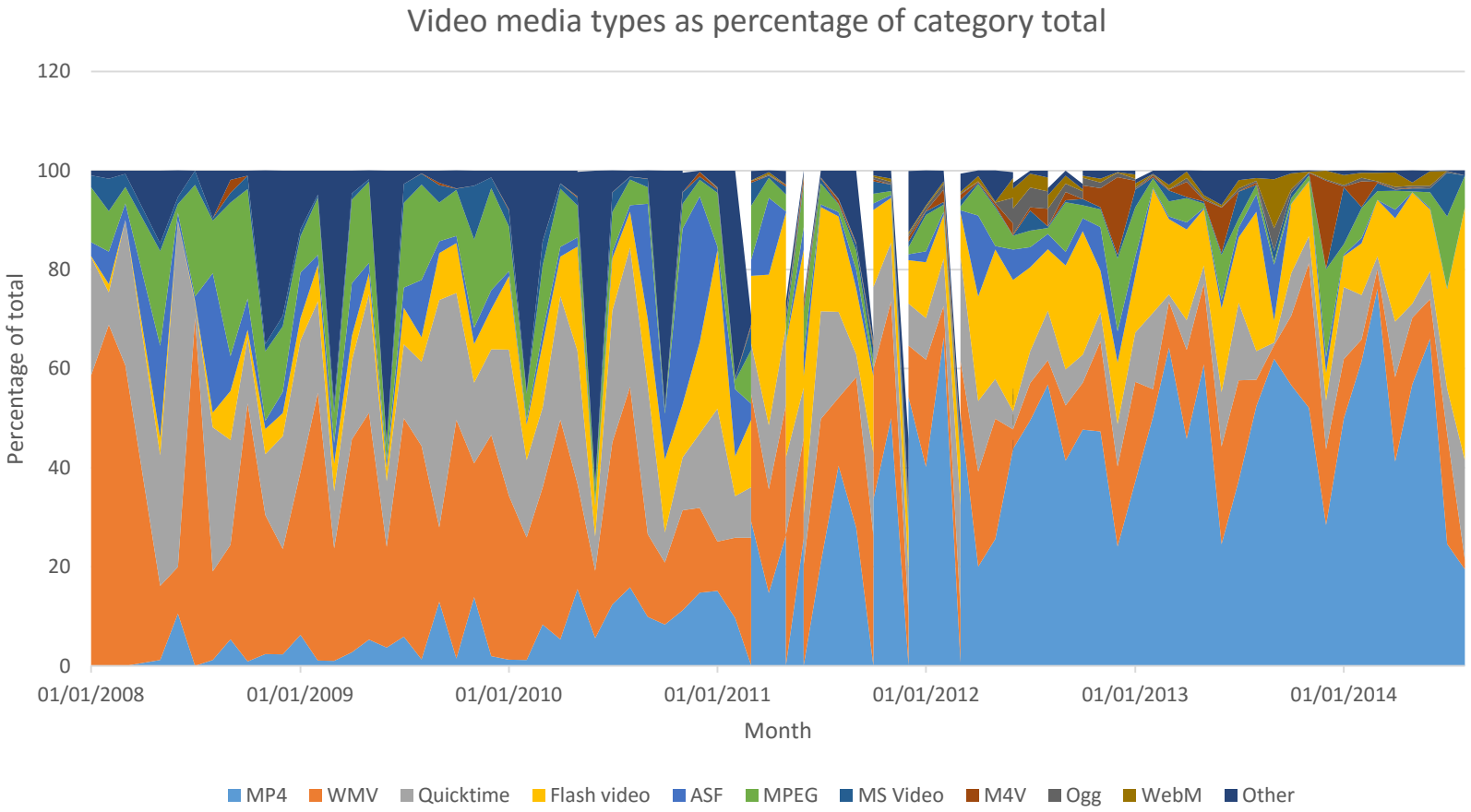
Media types - video

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

Media types - video



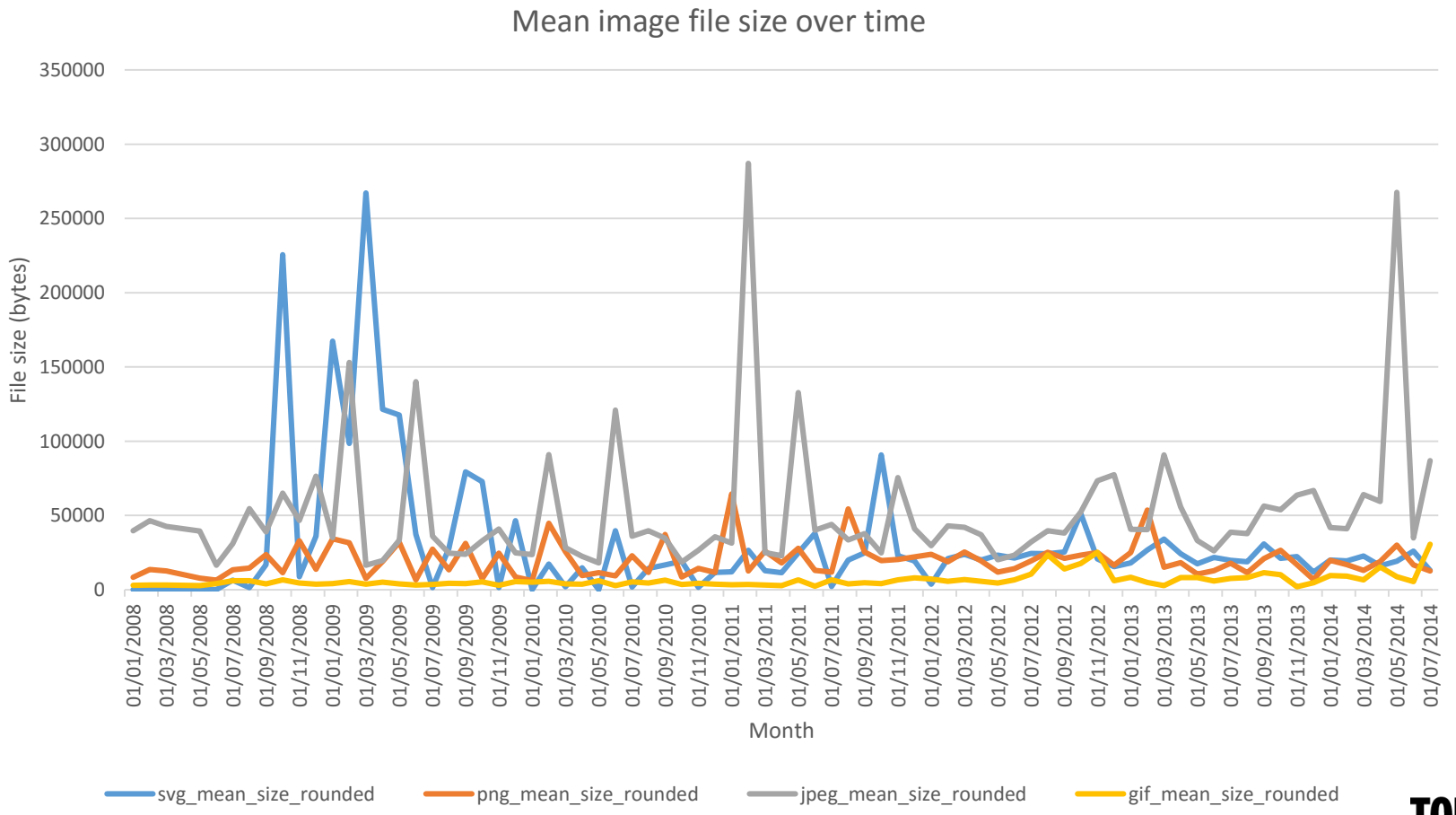
Media types - video



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

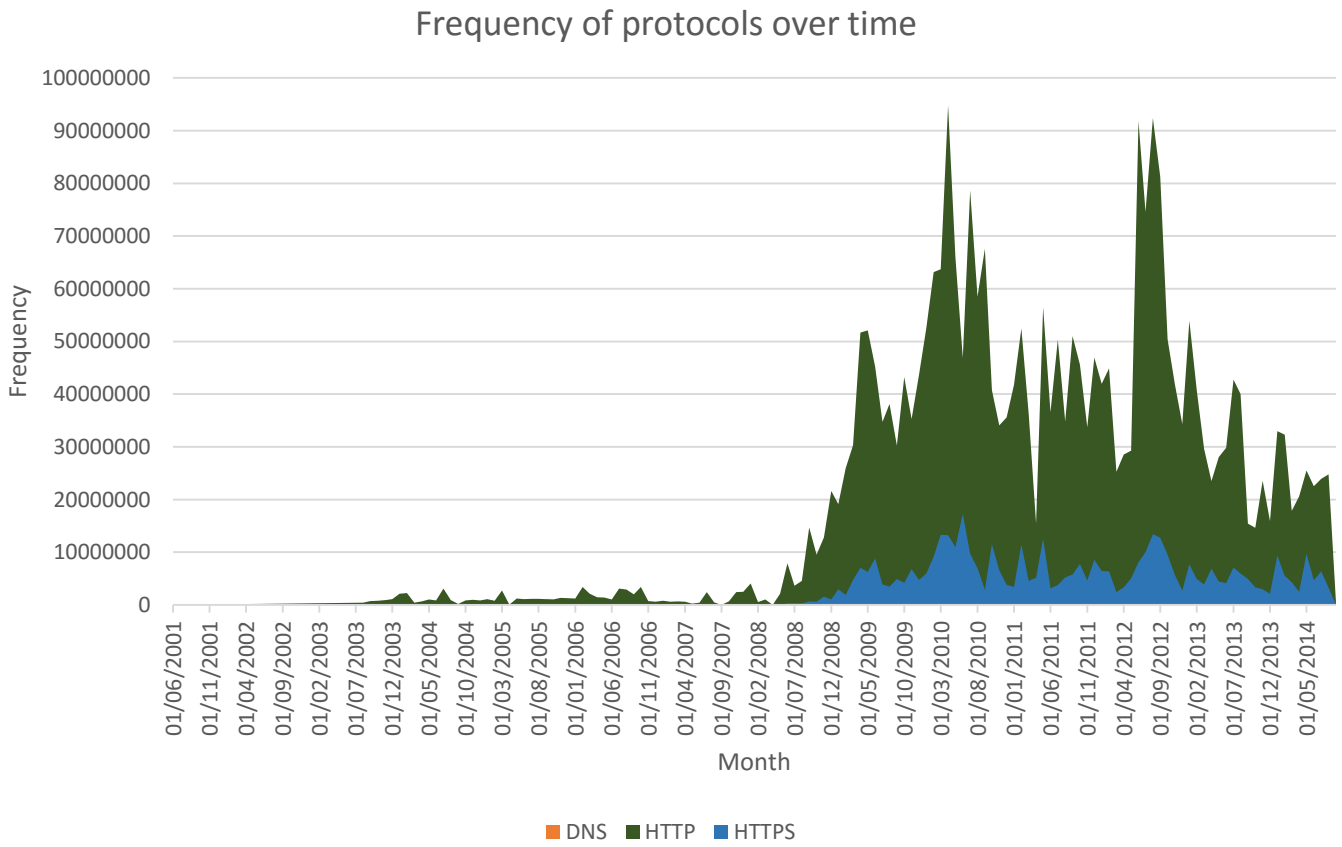
File size over time - images



Protocol changes

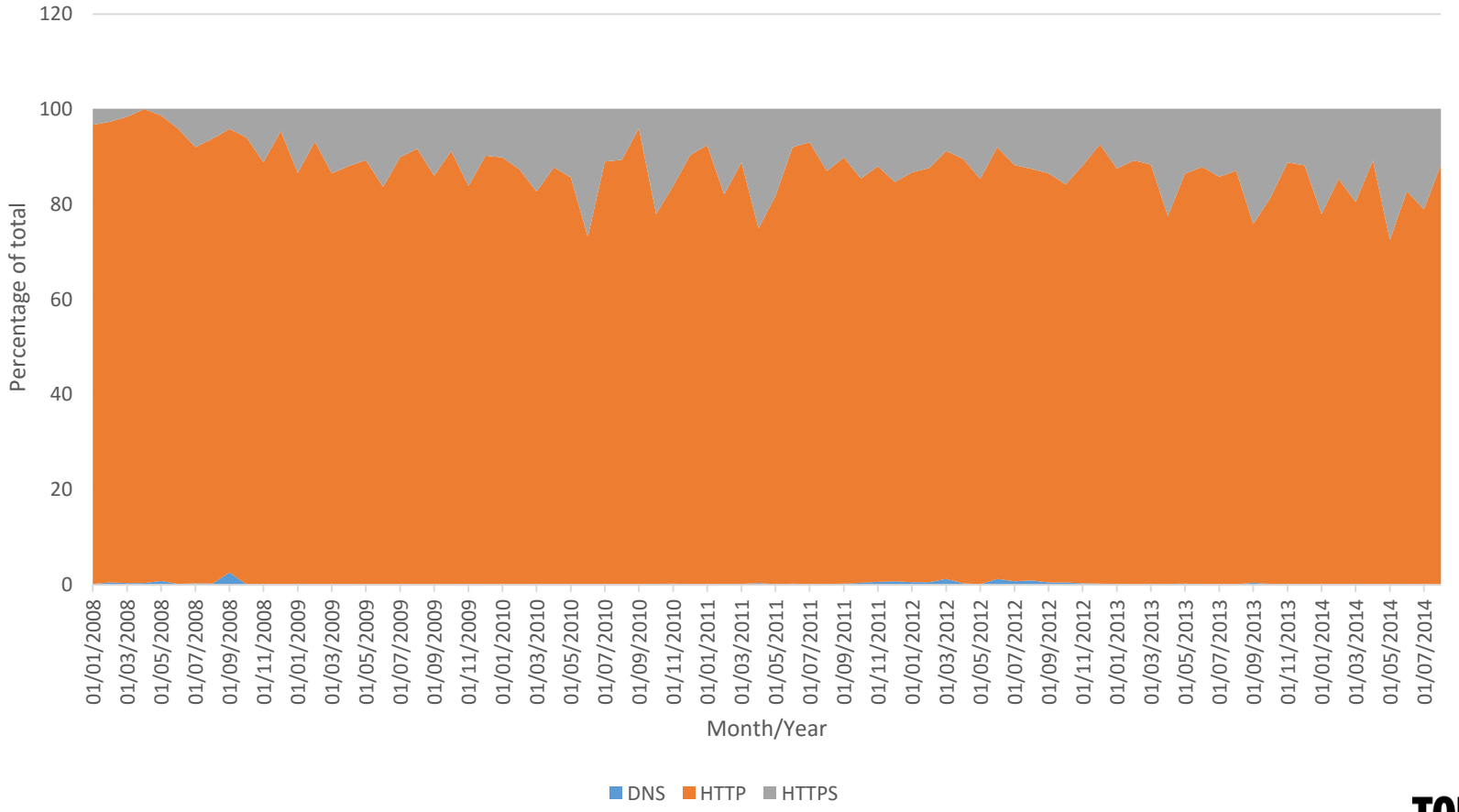
- CDX contains protocol information within URL parameters
- Protocol can be extracted from this parameter and aggregated temporally
- This reveals popularity trends for protocols
- HTTP vs. HTTPS

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