Twitter Challenges

november 2018

Jérôme Thièvre, Lead R&D engineer, jthievre@ina.fr

Legal context: 2011, Decree

Ina:

- Broadcasters’ online communication services
- Online communication services focusing on Radio and TV
- On demand audiovisual Media Services with broadcast content

Archive access restricted to Ina facilities

Twitter Collect at Ina

- tv/radio entities : channels, shows & media personalities
    - official publications and audience reactions
- events with high media echo
    - elections, sport, terrorist attacks, ...
- tweets embedded in archived web pages
- 700 000 tweets / day

Twitter Curation

- integrated with others social and video web platforms

⇒ link tv/radio entities to their web presences
  - website or part of a website
  - social account or keyword/hashtag (twitter, facebook, instagram, ...)
  - video account/channel (youtube, dailymotion, vimeo, ...)

- curation has been refactored the last 2 years
- managed internally in collaboration with a tv/radio documentalist

API motivations & usages

- Canonical form
- Rich structured data
    - easy to index and search
    - opportunity to experiment data visualizations
- 13500 accounts
    - timeline + search (mentions)
- 750 hashtags/keywords
    - streaming + search (7 day window sample retrieval)
- 2500 avg tweets from webpages every day
    - get statuses / hydrate

4 years of collect

- nearly 1 billion tweets collected
- APIs are stable, only one important change
- and reliable, no major collect interruption to report

4 years of collect

- exhaustiveness can’t be obtained, but attention is required to limit data ‘holes’

- 1% streaming cap was rarely reached
    - during mondial events : paris attacks, olympics, world cup, ...

- requires 'smart' use of multiple accounts

Linked materials

- tweet can contain images, videos or links
- images and video entities are collected too
  - declared as media entities by twitter
- but not external links for now
  - we miss pages, and external videos and images
  - but plan to do it
  - already do this kind of collect for linked materials from a webpage (embedded video & tweets)

Trends API experimentation

- idea : provide a safety net
- extract France Trends most popular hashtags/keywords
- collect related tweets with streaming API

Trends API experimentation

- idea : provide a safety net
- extract France Trends most popular hashtags/keywords
- collect related tweets with streaming API

- collected tweets are mostly irrelevant to our collection
- and contain a lot of spam/junk/porn messages
- this collection is kept apart from our official collection

Storage

- how to store tweets in DAFF (Digital Archive File Format)
  - keep our separation between metadata and data
  - new type of metadata web data not identified by an url

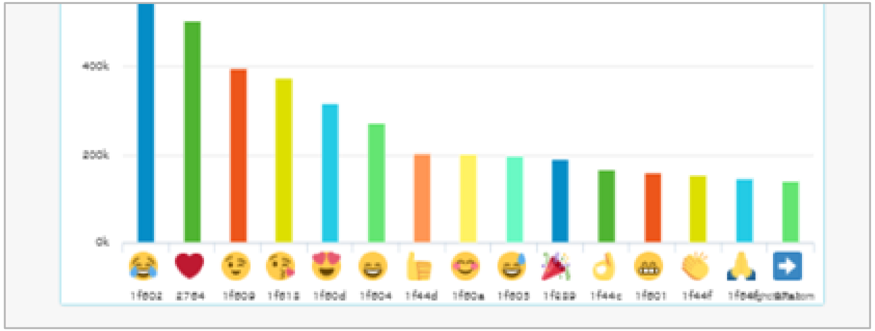
- records :
  - metadata : urn, status, origin, date, content type, collect method, collection name, content (sha256 link to data record), ...
  - data : raw json tweet in standard data DAFF record

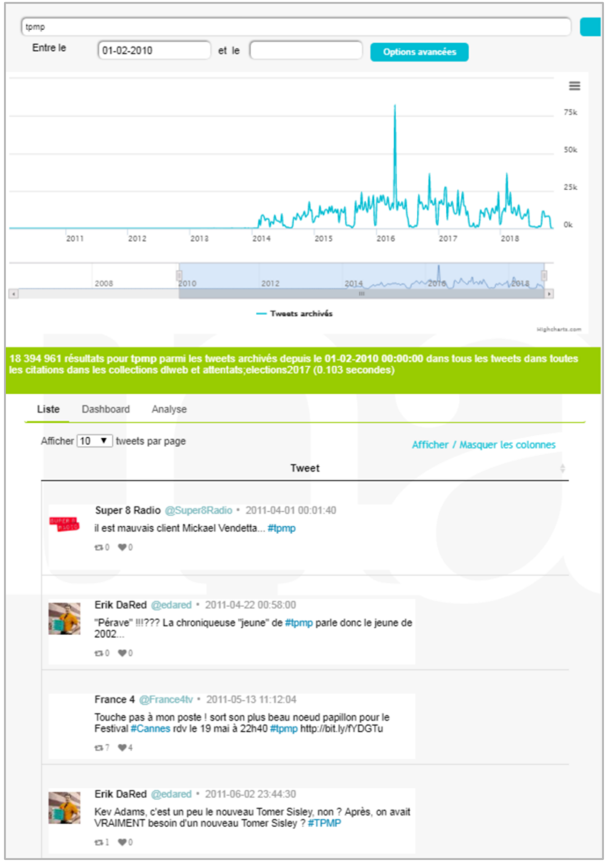
Archive storage and preservation is integrated in our current workflow

Access

Specific access application :

- fulltext search + specific fields
- timeline
- facets on relevant fields
- challenge to build
- good feedback from users





18 394 961 résultats pour tpmp parmi les tweets archivés depuis le 01-02-2010 00:00:00 dans tous les tweets dans toutes les citations dans les collections dlweb et attentats;elections2017 (0.103 secondes)

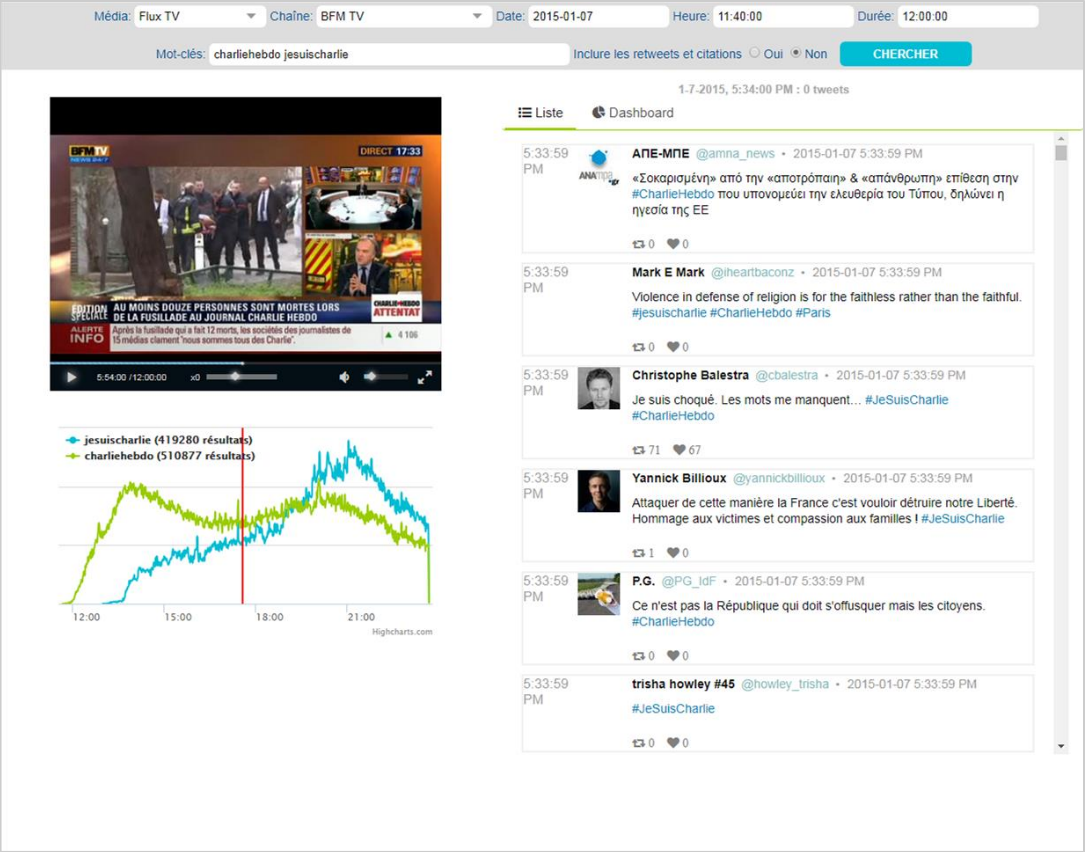
Super 8 Radio @Super8Radio • 2011-04-01 00:01:40
il est mauvais client Mickael Vendetta... #tpmp
0 ♥ 0

Erik DaRed @edared • 2011-04-22 00:58:00
"Pérave" !!!??? La chroniqueuse "jeune" de #tpmp parle donc le jeune de 2002...
0 ♥ 0

France 4 @France4tv • 2011-05-13 11:12:04
Touche pas à mon poste ! sort son plus beau noeud papillon pour le Festival #Cannes rdv le 19 mai à 22h40 #tpmp http://bit.ly/fYDGTu
7 ♥ 4

Erik DaRed @edared • 2011-06-02 23:44:30
Kev Adams, c'est un peu le nouveau Tomer Sisley, non ? Après, on avait VRAIMENT besoin d'un nouveau Tomer Sisley ? #TPMP
1 ♥ 0

Access : social TV



Média: Flux TV
Chaîne: BFM TV
Date: 2015-01-07
Heure: 11:40:00
Durée: 12:00:00

Mot-clés: charliehebdo jesuischarlie

Inclure les retweets et citations ○ Oui ● Non

CHERCHER

BFM TV
NEWS 24/7
DIRECT 17:33
EDITION SPÉCIALE
AU MOINS DOUZE PERSONNES SONT MORTES LORS DE LA FUSILLADE AU JOURNAL CHARLIE HEBDO
CHARLIE HEBDO
ATTENTAT
ALERTE INFO
Après la fusillade qui a fait 12 morts, les sociétés des journalistes de 15 médias clament "nous sommes tous des Charlie".
▲ 4 106
5:54:00 /12:00:00
x0
jesuischarlie (419280 résultats)
charliehebdo (510877 résultats)
12:00
15:00
18:00
21:00
Highcharts.com

1-7-2015, 5:34:00 PM : 0 tweets

Liste | Dashboard

5:33:59 PM — **ΑΠΕ-ΜΠΕ** @amna_news • 2015-01-07 5:33:59 PM

«Σοκαρισμένη» από την «αποτρόπαιη» & «απάνθρωπη» επίθεση στην #CharlieHebdo που υπονομεύει την ελευθερία του Τύπου, δηλώνει η ηγεσία της ΕΕ

0 ♥ 0

5:33:59 PM — **Mark E Mark** @iheartbaconz • 2015-01-07 5:33:59 PM

Violence in defense of religion is for the faithless rather than the faithful. #jesuischarlie #CharlieHebdo #Paris

0 ♥ 0

5:33:59 PM — **Christophe Balestra** @cbalestra • 2015-01-07 5:33:59 PM

Je suis choqué. Les mots me manquent… #JeSuisCharlie #CharlieHebdo

71 ♥ 67

5:33:59 PM — **Yannick Billioux** @yannickbillioux • 2015-01-07 5:33:59 PM

Attaquer de cette manière la France c'est vouloir détruire notre Liberté. Hommage aux victimes et compassion aux familles ! #JeSuisCharlie

1 ♥ 0

5:33:59 PM — **P.G.** @PG_IdF • 2015-01-07 5:33:59 PM

Ce n'est pas la République qui doit s'offusquer mais les citoyens. #CharlieHebdo

0 ♥ 0

5:33:59 PM — **trisha howley #45** @howley_trisha • 2015-01-07 5:33:59 PM

#JeSuisCharlie

0 ♥ 0

Conclusion

- easy to collect from twitter API
- access complete & canonical data
- scalable & reliable

- storage & preservation integration
- require specific indexing and access

any Question ?

Thank you!

dlweb@ina.fr

Annex : Twitter APIs

- timeline : last 3200 tweets + 200 tweets / request + 1500 requests / 15 min
- streaming : realtime with 1% global cap + 400 filters max
- search : sample of tweets from last 7 days + 100 tweets / requests + 450 requests / min)
- get statuses : 100 tweets / request + 900 requests / 15 min

Annex : DAFF

```
record ==
             <"R">
             <1Byte type flag>
             <1Byte compression flag>
             <256bit SHA-256 signature>
             <64bit content size>
             <content>
type flag
             0 : format description record
             1 : purpose description
record
             10 : metadata record in YAML
format
             20 : metadata record in JSON
format
             11 : data record
             can be extended
compression flag
             0 : raw
             1 : deflate
             2 : gzip
             3 : bzip2
             can be extended
```

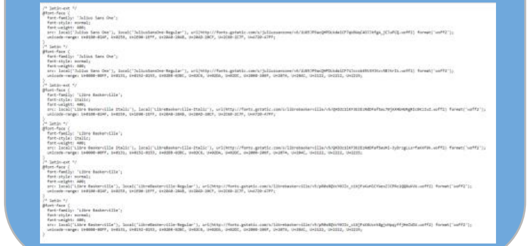
```
Metadata record
             Type flag: 10 or 20
             YAML (10) or JSON (20) encoded.

             Mandatory keys
                          - url, in a normalized form
                          - date, ISO 8601 string
                          - content: SHA-256 of the content in hexadecimal
format
                          - status: 'ok' | 'redirection' | 'request_error' |
'server_error' | 'info' | 'unchanged' | 'ignored'
                          - location: url to redirect to, in case of
protocol redirection (status = 'redirection')
             Optional keys
                          - type (MIME Type)
                          - page, 0/1, reflects if the url was a page
                          - last_modified, reflecting the Last-Modified HTTP
header
                          - ip, the ip of the peer server (if not a proxy)
                          - comment, plain message
                          - ...
```

Annex : DAFF for http(s) data

daff data records

sha_256: e4ba78b2c0034f

sha_256: babc4e3130004def6



```
/* latin-ext */
@font-face {
  font-family: 'Julius Sans One';
  font-style: normal;
  font-weight: 400;
  src: local('Julius Sans One'), local('JuliusSansOne-Regular'), url(http://fonts.gstatic.com/s/juliussansone/v6/[illegible].woff2) format('woff2');
  unicode-range: [illegible];
}
/* latin */
@font-face {
  font-family: 'Julius Sans One';
  font-style: normal;
  font-weight: 400;
  src: local('Julius Sans One'), local('JuliusSansOne-Regular'), url(http://fonts.gstatic.com/s/juliussansone/v6/[illegible].woff2) format('woff2');
  unicode-range: [illegible];
}
/* latin-ext */
@font-face {
  font-family: 'Libre Baskerville';
  font-style: italic;
  font-weight: 400;
  src: local('Libre Baskerville Italic'), local('LibreBaskerville-Italic'), url(http://fonts.gstatic.com/s/librebaskerville/v5/[illegible].woff2) format('woff2');
  unicode-range: [illegible];
}
/* latin */
@font-face {
  font-family: 'Libre Baskerville';
  font-style: italic;
  font-weight: 400;
  src: local('Libre Baskerville Italic'), local('LibreBaskerville-Italic'), url(http://fonts.gstatic.com/s/librebaskerville/v5/[illegible].woff2) format('woff2');
  unicode-range: [illegible];
}
/* latin-ext */
@font-face {
  font-family: 'Libre Baskerville';
  font-style: normal;
  font-weight: 400;
  src: [illegible]
  unicode-range: [illegible];
}
/* latin */
@font-face {
  font-family: 'Libre Baskerville';
  font-style: normal;
  font-weight: 400;
  src: [illegible]
  unicode-range: [illegible];
}
```

daff metadata records

```
url:     http://prog-tv.fr/images/goldorak.jpg
date:    2018-01-02 08:51:00Z
sha_256: e4ba78b2c0034f
```

```
url:     http://prog-tv.fr/images/goldorak.jpg
date:    2018-01-13 10:02:00Z
sha_256: e4ba78b2c0034f
```

```
url:     http://anime-fan.fr/img/grendizer.jpg
date:    2018-02-02 18:33:00Z
sha_256: e4ba78b2c0034f
```

```
url:     http://prog-tv.fr/style/main.css
date:    2018-01-02 08:50:00Z
sha_256: babc4e3130004def6
```

```
url:     http://prog-tv.fr/style/main.css
date:    2018-01-13 10:01:00Z
sha_256: babc4e3130004def6
```