



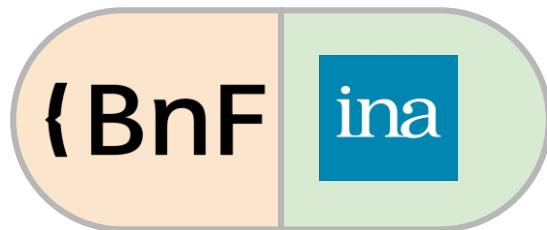
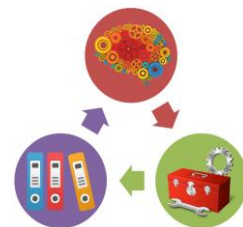
Twitter Challenges

november 2018



ina

Legal context: 2011, Decree



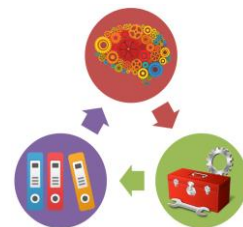
Ina:

- 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

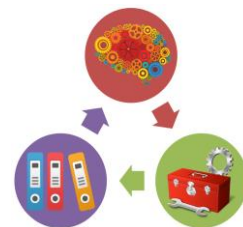


Twitter Collect at Ina



- 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

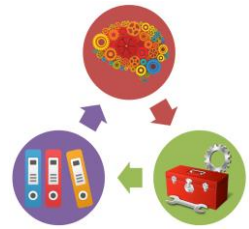
API motivations & usages

- 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

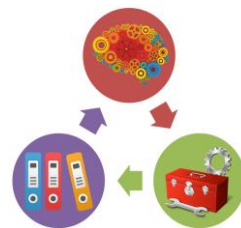


4 years of collect

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

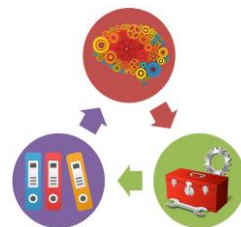


4 years of collect



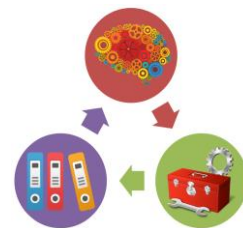
- exhaustiveness can't be obtained, but attention is required to limit data 'holes'
- 1% streaming cap was rarely reached
 - during mondial events : paris attacks, olympics, world cup, ...
- requires 'smart' use of multiple accounts

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

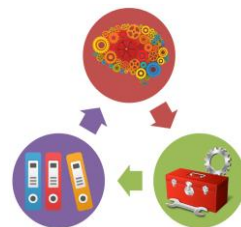
Trends API experimentation



- 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



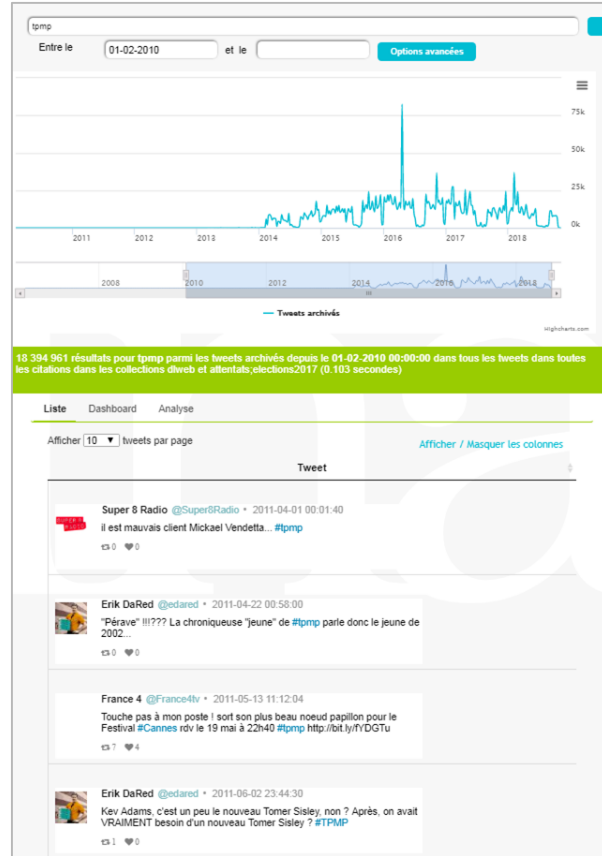
- how to store tweets in DAFF (Digital Archive File Format)
 - keep our separation between metadata and data
 - new type of metadata web data not identified by an url
- records :
 - metadata : urn, status, origin, date, content type, collect method, collection name, content (sha256 link to data record), ...
 - data : raw json tweet in standard data DAFF record

Archive storage and preservation is integrated in our current workflow

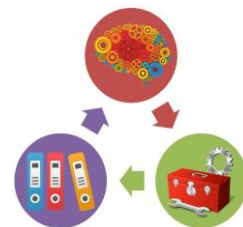
Access

Specific access application :

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



Access : social TV



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

Mot-clés: charlieheβδο jesuischarlie Inclure les retweets et citations Oui Non **CHERCHER**

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

Liste Dashboard

◆ jesuischarlie (419280 résultats)
◆ charlieheβδο (510877 résultats)

Highcharts.com

5:33:59 PM **ΑΠΕ-ΜΠΕ** @amna_news · 2015-01-07 5:33:59 PM
«Σοκαρισμένη» από την «αποτρόπαιη» & «απάνθρωπη» επίθεση στην #CharlieHeβδο που υπονομεύει την ελευθέρια του Τύπου, δηλώνει η ηγεσία της ΕΕ

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 #CharlieHeβδο #Paris

5:33:59 PM **Christophe Balestra** @cbalestra · 2015-01-07 5:33:59 PM
Je suis choqué. Les mots me manquent... #JeSuisCharlie #CharlieHeβδο

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

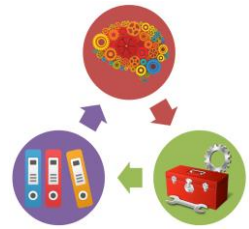
5:33:59 PM **P.G.** @PG_Idf · 2015-01-07 5:33:59 PM
Ce n'est pas la République qui doit s'offusquer mais les citoyens. #CharlieHeβδο

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

Conclusion

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

- storage & preservation integration
- require specific indexing and access





any Question ?



ina



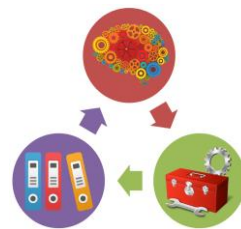
Thank you!

dlweb@ina.fr



ina

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

```
1 sha_256: 01
2 sha_256: 02
3 sha_256: 03
4 sha_256: 04
5 sha_256: 05
6 sha_256: 06
7 sha_256: 07
8 sha_256: 08
9 sha_256: 09
10 sha_256: 0a
11 sha_256: 0b
12 sha_256: 0c
13 sha_256: 0d
14 sha_256: 0e
15 sha_256: 0f
16 sha_256: 10
17 sha_256: 11
18 sha_256: 12
19 sha_256: 13
20 sha_256: 14
21 sha_256: 15
22 sha_256: 16
23 sha_256: 17
24 sha_256: 18
25 sha_256: 19
26 sha_256: 1a
27 sha_256: 1b
28 sha_256: 1c
29 sha_256: 1d
30 sha_256: 1e
31 sha_256: 1f
32 sha_256: 20
33 sha_256: 21
34 sha_256: 22
35 sha_256: 23
36 sha_256: 24
37 sha_256: 25
38 sha_256: 26
39 sha_256: 27
40 sha_256: 28
41 sha_256: 29
42 sha_256: 2a
43 sha_256: 2b
44 sha_256: 2c
45 sha_256: 2d
46 sha_256: 2e
47 sha_256: 2f
48 sha_256: 30
49 sha_256: 31
50 sha_256: 32
51 sha_256: 33
52 sha_256: 34
53 sha_256: 35
54 sha_256: 36
55 sha_256: 37
56 sha_256: 38
57 sha_256: 39
58 sha_256: 3a
59 sha_256: 3b
60 sha_256: 3c
61 sha_256: 3d
62 sha_256: 3e
63 sha_256: 3f
64 sha_256: 40
65 sha_256: 41
66 sha_256: 42
67 sha_256: 43
68 sha_256: 44
69 sha_256: 45
70 sha_256: 46
71 sha_256: 47
72 sha_256: 48
73 sha_256: 49
74 sha_256: 4a
75 sha_256: 4b
76 sha_256: 4c
77 sha_256: 4d
78 sha_256: 4e
79 sha_256: 4f
80 sha_256: 50
81 sha_256: 51
82 sha_256: 52
83 sha_256: 53
84 sha_256: 54
85 sha_256: 55
86 sha_256: 56
87 sha_256: 57
88 sha_256: 58
89 sha_256: 59
90 sha_256: 5a
91 sha_256: 5b
92 sha_256: 5c
93 sha_256: 5d
94 sha_256: 5e
95 sha_256: 5f
96 sha_256: 60
97 sha_256: 61
98 sha_256: 62
99 sha_256: 63
100 sha_256: 64
101 sha_256: 65
102 sha_256: 66
103 sha_256: 67
104 sha_256: 68
105 sha_256: 69
106 sha_256: 6a
107 sha_256: 6b
108 sha_256: 6c
109 sha_256: 6d
110 sha_256: 6e
111 sha_256: 6f
112 sha_256: 70
113 sha_256: 71
114 sha_256: 72
115 sha_256: 73
116 sha_256: 74
117 sha_256: 75
118 sha_256: 76
119 sha_256: 77
120 sha_256: 78
121 sha_256: 79
122 sha_256: 7a
123 sha_256: 7b
124 sha_256: 7c
125 sha_256: 7d
126 sha_256: 7e
127 sha_256: 7f
128 sha_256: 80
129 sha_256: 81
130 sha_256: 82
131 sha_256: 83
132 sha_256: 84
133 sha_256: 85
134 sha_256: 86
135 sha_256: 87
136 sha_256: 88
137 sha_256: 89
138 sha_256: 8a
139 sha_256: 8b
140 sha_256: 8c
141 sha_256: 8d
142 sha_256: 8e
143 sha_256: 8f
144 sha_256: 90
145 sha_256: 91
146 sha_256: 92
147 sha_256: 93
148 sha_256: 94
149 sha_256: 95
150 sha_256: 96
151 sha_256: 97
152 sha_256: 98
153 sha_256: 99
154 sha_256: 9a
155 sha_256: 9b
156 sha_256: 9c
157 sha_256: 9d
158 sha_256: 9e
159 sha_256: 9f
160 sha_256: a0
161 sha_256: a1
162 sha_256: a2
163 sha_256: a3
164 sha_256: a4
165 sha_256: a5
166 sha_256: a6
167 sha_256: a7
168 sha_256: a8
169 sha_256: a9
170 sha_256: aa
171 sha_256: ab
172 sha_256: ac
173 sha_256: ad
174 sha_256: ae
175 sha_256: af
176 sha_256: b0
177 sha_256: b1
178 sha_256: b2
179 sha_256: b3
180 sha_256: b4
181 sha_256: b5
182 sha_256: b6
183 sha_256: b7
184 sha_256: b8
185 sha_256: b9
186 sha_256: ba
187 sha_256: bb
188 sha_256: bc
189 sha_256: bd
190 sha_256: be
191 sha_256: bf
192 sha_256: c0
193 sha_256: c1
194 sha_256: c2
195 sha_256: c3
196 sha_256: c4
197 sha_256: c5
198 sha_256: c6
199 sha_256: c7
200 sha_256: c8
201 sha_256: c9
202 sha_256: ca
203 sha_256: cb
204 sha_256: cc
205 sha_256: cd
206 sha_256: ce
207 sha_256: cf
208 sha_256: d0
209 sha_256: d1
210 sha_256: d2
211 sha_256: d3
212 sha_256: d4
213 sha_256: d5
214 sha_256: d6
215 sha_256: d7
216 sha_256: d8
217 sha_256: d9
218 sha_256: da
219 sha_256: db
220 sha_256: dc
221 sha_256: dd
222 sha_256: de
223 sha_256: df
224 sha_256: e0
225 sha_256: e1
226 sha_256: e2
227 sha_256: e3
228 sha_256: e4
229 sha_256: e5
230 sha_256: e6
231 sha_256: e7
232 sha_256: e8
233 sha_256: e9
234 sha_256: ea
235 sha_256: eb
236 sha_256: ec
237 sha_256: ed
238 sha_256: ee
239 sha_256: ef
240 sha_256: f0
241 sha_256: f1
242 sha_256: f2
243 sha_256: f3
244 sha_256: f4
245 sha_256: f5
246 sha_256: f6
247 sha_256: f7
248 sha_256: f8
249 sha_256: f9
250 sha_256: fa
251 sha_256: fb
252 sha_256: fc
253 sha_256: fd
254 sha_256: fe
255 sha_256: ff
```

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