Accessing Archived Twitter Data @ina

Zeynep Pehlivan
Thomas Drugeon
Jérôme Thièvre

IIPC Web Archiving Conference 14th April 2016

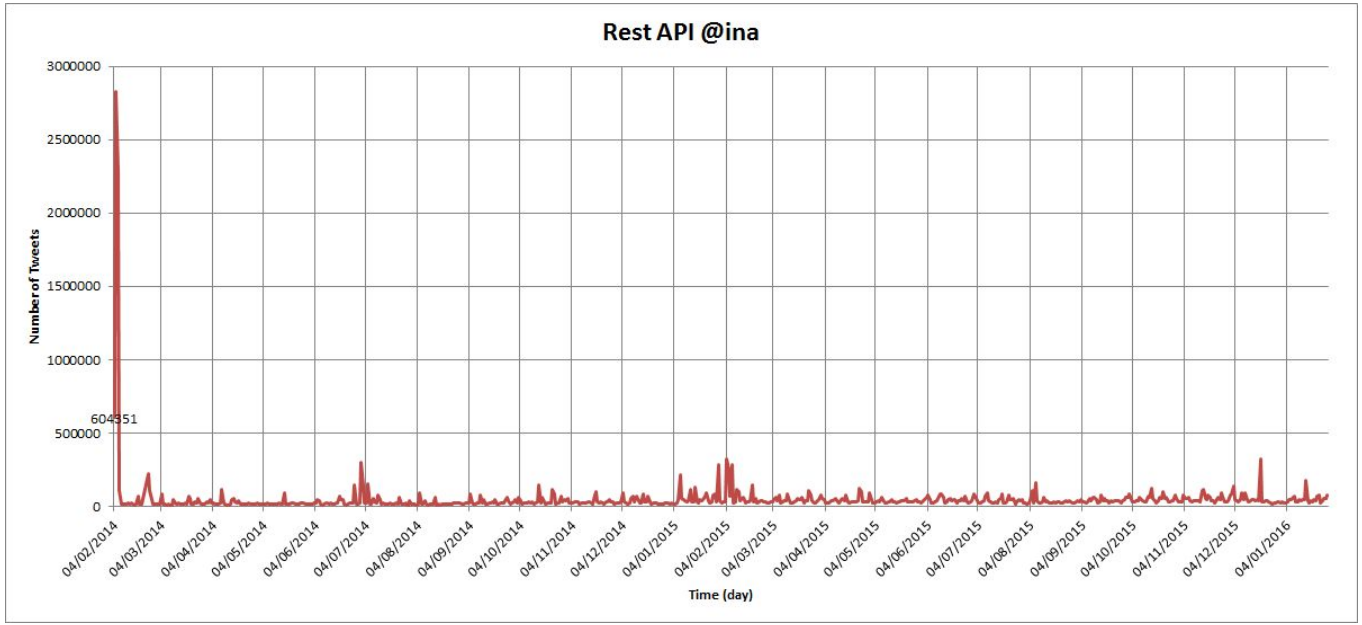
Archive Twitter @ina

- Radio, TV, web
- Crawling by using Twitter API (data)
- Since February 2014
- 11 000 users(timelines)
- 400 hashtags
- 250 millions of tweets
- Recontextualization(s)

ina

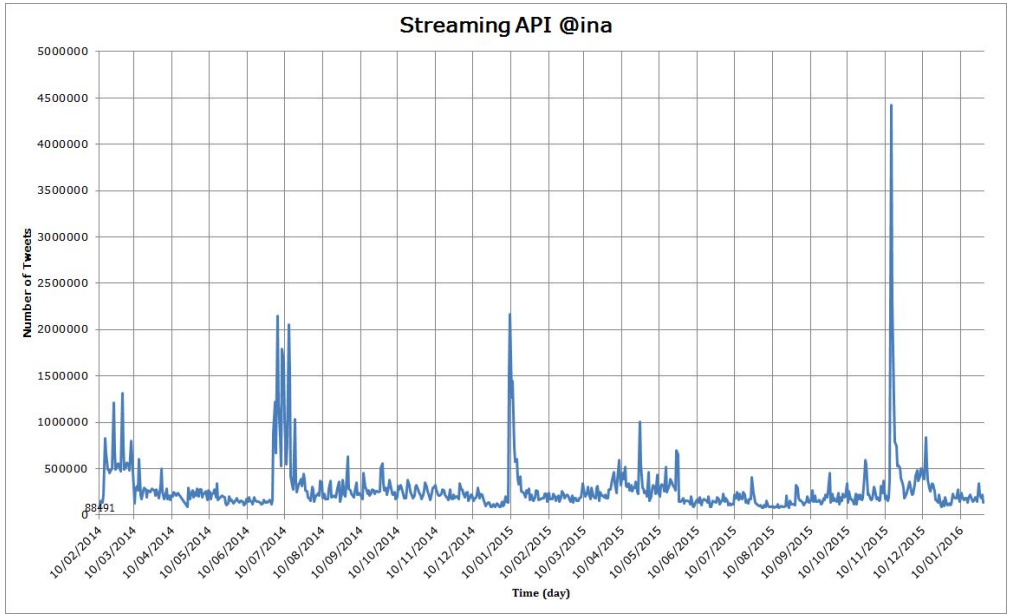
Rest API : timelines

- Total : 40 millions
- Average per day: 48 000



Streaming API: hashtags

- Total : 210 millions
- Average per day : 270 000



Restrictions

- Streaming API : 400 hashtags, 5000 users
- 1% of tweets published at time t
- REST API : 3200 old tweets per user
- Search API : 15 minute window of 180 for user and 450 for app

ina

Additional Sources

- Linkfluence
  - `#JeSuisCharlie, #CharlieHebdo`
- Nick Ruest, 2015-01-12
  - `#JeSuisCharlie, #JeSuisAhmed, #JeSuisJuif, #CharlieHebdo`
- Nick Ruest, 2015-12-14
  - `#paris, #Bataclan, #parisattacks, #porteouverte`

ina

Recontextualization

- Canonical version?
  - Twitter page?
  - Data
  - Authenticity, integrity
- Second Screen
  - TV Sync
  - Indexing
- Search / data mining
  - Data coverage
  - Generic or specific tools
  - Open data?

ina

Missing data

- 1% = 100% most of the time
- 1% is representative for analytics
- Coverage of our collection
- Possible to estimate?
    - For the hashtags we track
    - Based on the rate limit information

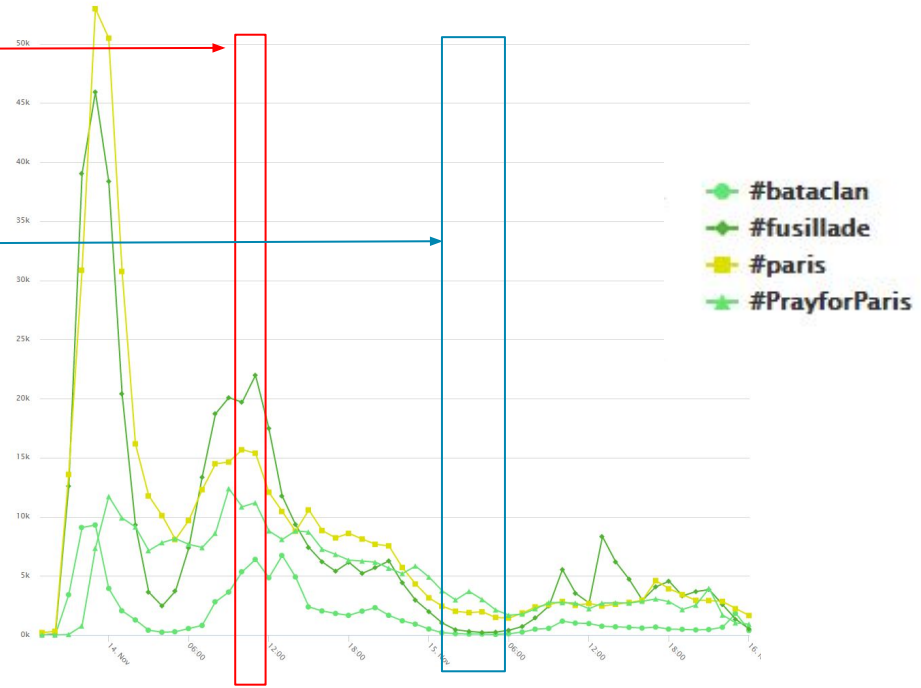

```
{"follow":[],"track":["#PrayForParis","#fusillade",
…
{"limit":{"track":2058,"timestamp_ms":"1447456084764"}}
…
{"limit":{"track":2159,"timestamp_ms":"1447456084783"}}
```


ina

Estimations

"Track":2058
"timestamp_ms":"1447456084764"

"Track":2159
"timestamp_ms":"1447456084783"



Issues

- No ground truth (Firehose data)
- Additional sources
- Streaming API is biased
  - Hashtags overrepresented or underrepresented
  - Window size, shape, dynamicity
- Recent research (F. Morsttatter) propose to use Sample API

ina

Demo

Thank you
Questions?