The challenges of collecting social media

Géraldine Camile – Peter Stirling
Bibliothèque nationale de France

From 2011 to today

- Boom in number of Twitter accounts collected, especially for elections

Top 10 des domaines en 2012	Nb de sites sélectionnés	%
twitter	1 033	10,34%
facebook	997	9,98%
over-blog	551	5,51%

Top 10 des domaines en 2017	Nb de sites sélectionnés	%
twitter	3 689	45,38%
wikipedia	401	4,93%
over-blog	120	1,48%
wordpress	108	1,33%
blogspot	86	1,06%

⇒ web crawls since 2011 for specific projects such as elections or Olympic games

- Today 84 % of the accounts selected in 2013 are still

⇒ Until 2018: bi-weekly daily crawl of accounts or

Web crawl versus API crawl

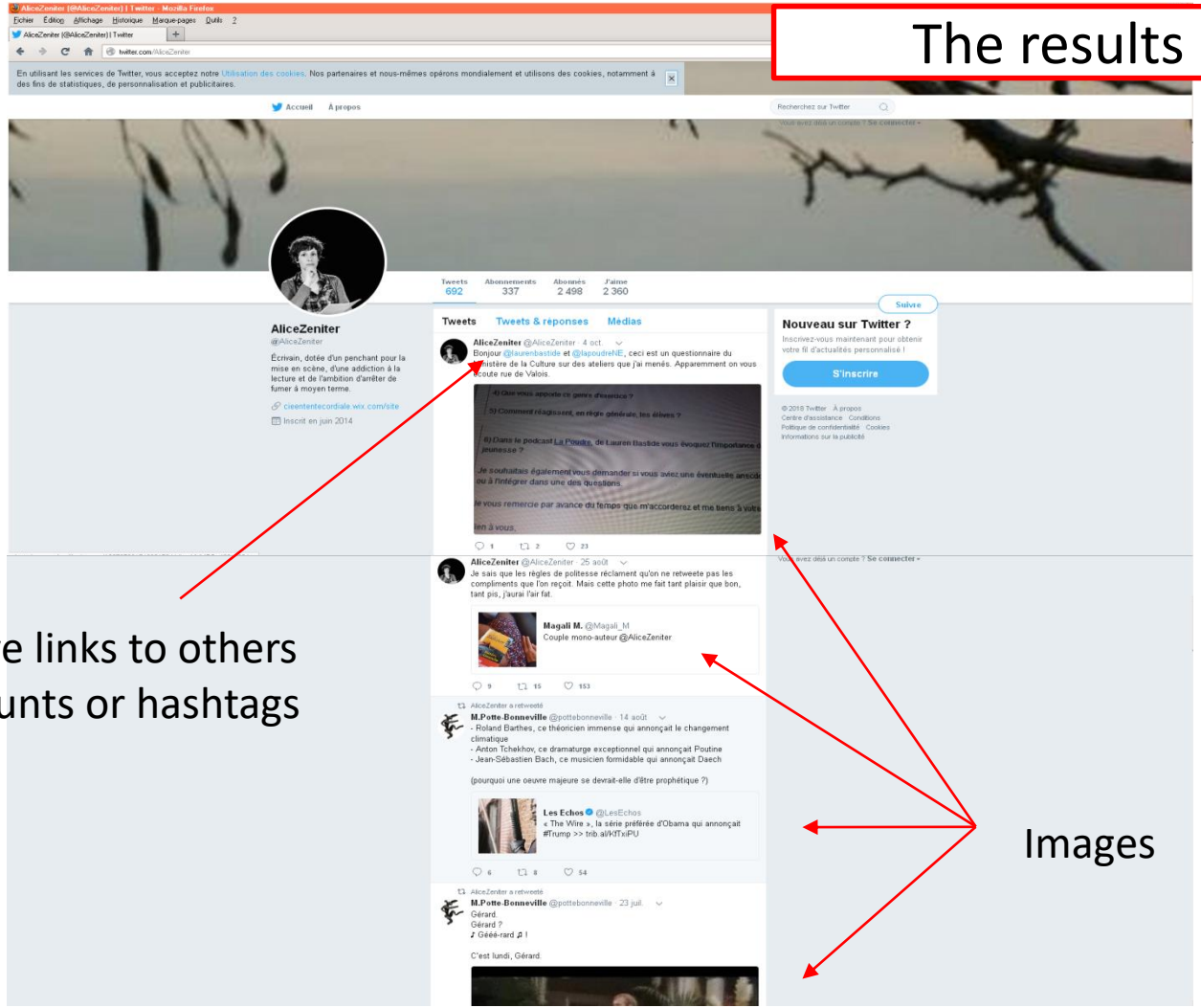
- Our main goals
  - Build **representative** collections ≠ all the tweets
    - Choice of accounts and hashtags according to a collection policy
  - Render them **as they were published** on the live web
    - The links between the accounts and hashtags
    - The media (images, videos…)
- Our constraints
  - No funding for an API service subscription
  - Less development on crawl, access and preservation

depth = Page +1

HERITRIX

WayBackMachine

The results



Active links to others accounts or hashtags

Images

The advantages

- Reproduce tweets as published on the live web

- Process uses the existing BnF workflow
  - Crawl using NetarchiveSuite and Heritrix 3
    - "websocial" harvest template: page + 1, duration 11 hours, specific filters
  - Access in the BnF web archives

The limits

- The timeline : 20 tweets per capture (40 tweets daily)

=> Representativeness ?

- The depth of the capture depends on the number of accounts selected and the duration of the crawl
  - 3 690 accounts during the elections

Two complementary approaches

	Web crawl	API crawl
Tools for crawl, access and preservation	BCWeb Heritrix 3 OpenWayback SPAR	Open source toole and libraries Specific access and preservation development
Advantages	Render tweets as online Keep the links between accounts Keep the existing workflow (homogeneity and interoperability of collections, preservation)	Developer friendly and easy to use Reliable (until now) Scalability Full metadata
Limits	Number of accounts (3000?) Number of tweets Missing videos Missing short links Few metadata	Public API limits Develop and maintain specific access solution