





# INTRODUCTION

# Genesis of a project

---

- Using Twitter since 2008/2009
  - Mostly for academic purposes, when I started doing DH + Organizing DH conferences (DHLU 2009 / THATCamp Paris 2010)
- Collecting tweets since 2012
  - During conferences (Search API)
  - For my own interest: #ledebat/#manifpourtous (Streaming API)
- 11 November 2013  
launching of the Centenary in France
  - *Rendez-vous de l'Histoire de Blois*
  - Strong suggestion by two ww1 historians

# What is at stake?

---

- Memory/Past/History
- Memories of the 'historical' past is an important research field since the 1970s
  - See Pierre Nora (*Lieux de Mémoire*, 1980s/1990s)
  - Memory studies, including now *Digital Memory studies*
- Strong media exposure from time to time (French case)
  - Example: 1990s and Vichy (Rousso / Conan, *Vichy, un passé qui ne passe pas*)
  - 2000's and slavery/French colonial Empire / « Memory laws » / Competition between memories
- Memories of the past also depend on the nature of the media
  - Notion of 'régime d'historicité' (F. Hartog): presentism/memory of the past

# Hashtags?

---

- Hashtags: user-generated functionality of Twitter
  - A keyword with a # (#ww1)
  - Have several significations: emphasizing a concept, contributing to a global discussion, being a member of a community, etc.
- Popularity of #ww1 or #pgm
  - First use of ww1: 16 April 2007
  - First use of #ww1: 11 March 2009
  - Imperial War Museum: first Centenary-dedicated account (March 2011, first tweet: 8 July 2011)

# Collected hashtags

---

#100years, #11NOV, #11novembre, #1ereGuerreMondiale, #1gm, #1j1p, #1Weltkrieg, #1wk, #centenaire, #centenary, #firstworldwar, #fww, #greatwar, #Hartmannswillerkopf, #passchendaele, #poppies, #Somme, #Somme100, #Testamentsdepoilus, #Verdun, #wk1, #WomenHeroesofWWI, #womenofworldwarone, #womenofww1, #womenofwwi, #womenww1, #ww1, #ww1athome, #WW1centenary, #wwi, #WWIcentenary, 1418Centenary, arras100, Cambrai100, centenaire14, centenaire1914, centenaire2014, chemindesdames, CWGC100, GrandeGuerre, passchendaele100, PoilusVerdun, RemembranceisEveryday, RussianRevolution, RussianRevolution1917, Verdun2016, vimy100, wwiafrica

# The current state of the corpus

---

- 1 April 2014 - ...
  - Around 5 million tweets as of today
  - 2/3 of retweets
  - Around 1 million Twitter accounts
    - private individuals, institutions, project-based account, bots and many others
  - Around 200 000 hashtags
    - A couple of ten thousands used more than 10 times
- Not a lot of noise, except for : #11Nov/Verdun/Somme

# Tools for #ww1: harvesting data

---

- LAMP server
  - From a home-based server to a more professional one
- Until september 2017 : 140dev.com
  - Collects tweets from the public streaming API (json) and parse it to a MySQL database
    - Under the 1% of the firehose: no need to use the full API (commercial way)
  - Some data are not harvested
    - profile's icons for instance – only URLs to the image are stored
- Since September 2017 : DMI-TCAT
  - 140dev not developed anymore
  - Twitter is fastly changing / DMIT-TCAT follows this development



# Storing data: data-model

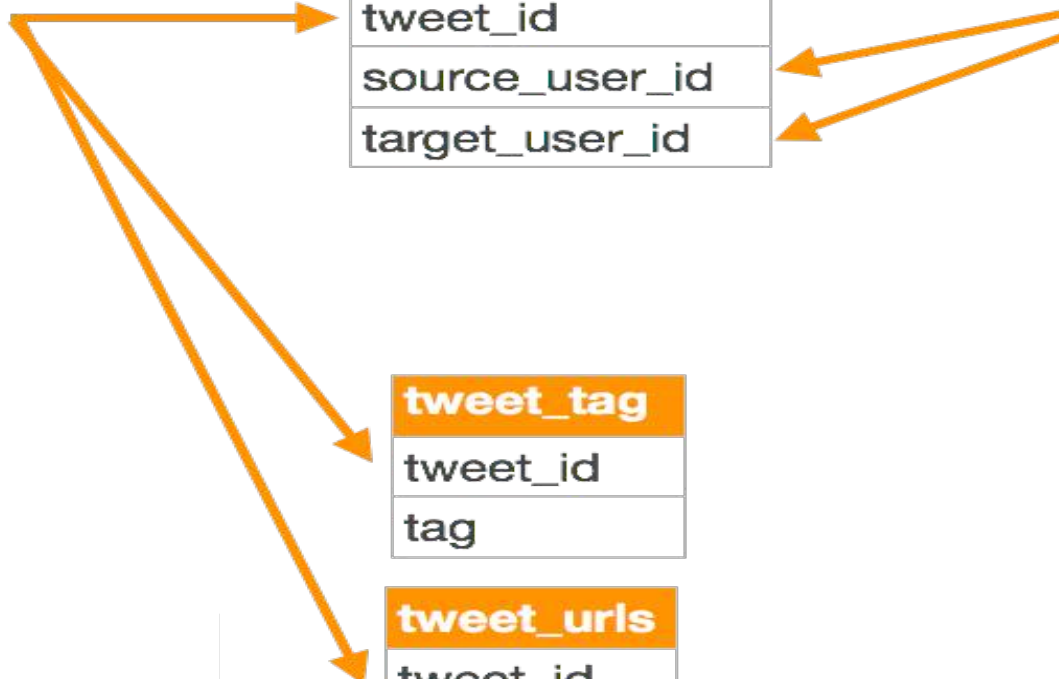
<b>tweets</b>
tweet_id
tweet_text
<b><i>created_at</i></b>
geo_lat
geo_long
user_id
screen_name
name
profile_image_ur
is_rt

<b>tweet_mentions</b>
tweet_id
source_user_id
target_user_id

<b>tweet_tag</b>
tweet_id
tag

<b>tweet_urls</b>
tweet_id
url

<b>users</b>
user_id
screen_name
name
profile_image_url
location
url
description
<b><i>created_at</i></b>
followers_count
friends_count
statuses_count
time_zone
<b><i>last_update</i></b>



# Exporting and preparing data


---

- SQL dump => non-dynamic database on laptop
  - Faster to deal with data
  - No real-time data treatment
- Export through SQL queries to CSV
  - Basic preparation with a combination of LibreOffice/OpenRefine/text editor
  - The magic of RegEx

# What kind of exports?

---

- Tweet-texts with metadata for text-mining
  - Original tweets only (No RTs)
  - Removal of hashtags, user names and URLs
- Different kinds of relations
  - Rts/mentions/hashtags...
- URLs
  - Lengthened through OpenRefine
  - Harvested, cleaned and text-mined
- Dates: Number of tweets/day
- Subparts of the corpus: Hashtags (#1j1p/#11novembre) - Iramuteq generated profiles



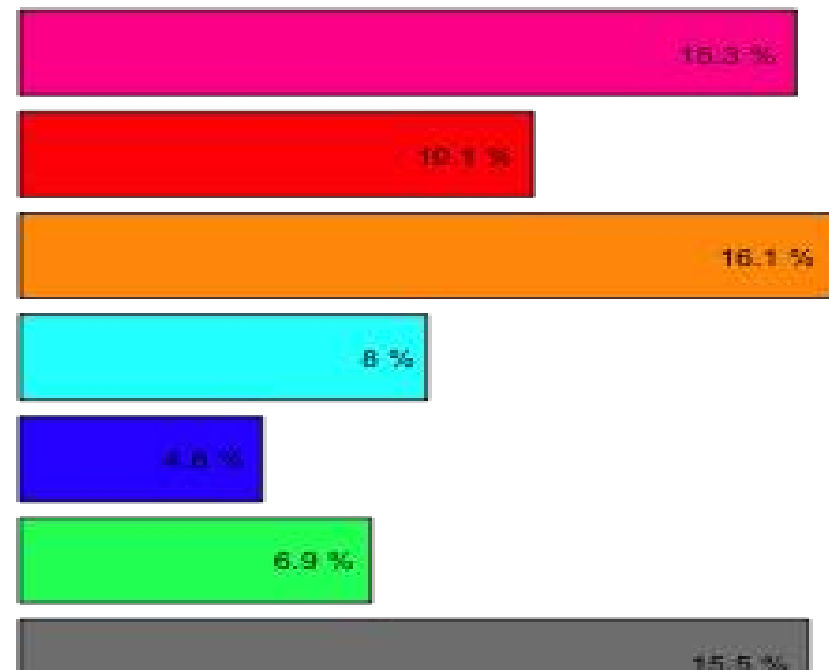
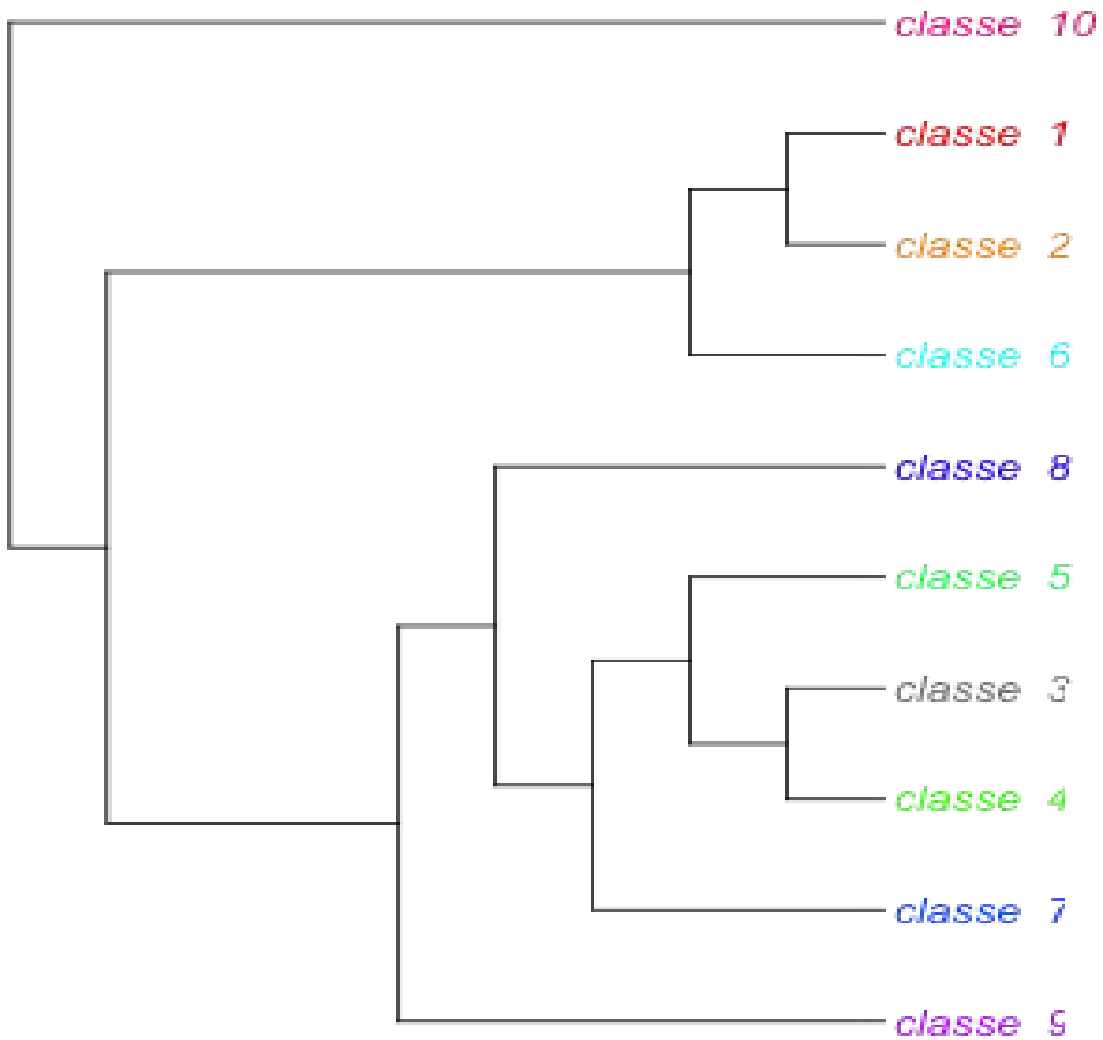
Historian facing a sea of data

**How to read 5 million tweets when you are not a trained data-scientist?**

# Key concept of distant reading

---

- Franco Moretti, *Graphs, Maps and Trees*, Verso, 2007.
  - *Graphs* (Annales School)  
quantitative approach of literature
  - *Maps* (Geography)  
mapping literature
  - *Trees* (Evolution theory)  
families of novels
- Articulation of close reading/distant reading



# I. AVAILABLE TOOLS

# Distant reading of tweets

---

- What kind of distant reading techniques are required?
  - Very basic statistical operation
    - number of tweets per day for instance
  - Data / Text mining
  - Network analysis
- Imply to deal with not-that-structured data

# Numerous tools are at our disposal...

---

- ...to start a data-driven research
  - See *Digital Research Tools* (DiRT), maintained by Lisa Spiro
  - 86 tools in the 'Analyse Data' section alone
- How to choose them?
  - The good: reading research production (articles, etc.) that grounds the tool
  - The bad: choosing a tool because its results are easier to interpret
  - The ugly: choosing a tool because it's a tool we already know
- How to compare them?



# Text-mining : the tool

---

- IRaMuTeQ
  - Based on Max Reinert's *Théorie des mondes lexicaux*
  - Open source implementation vs commercial one (Alceste)
  - Can deal with quite a large amount of texts/segments of text
- <http://www.iramuteq.org>

# Text-mining: clustering

---

- Clustering: *classification hiérarchique descendante*
  - See: REINERT Max, « Les “mondes lexicaux” et leur “logique” à travers l’analyse statistique d’un corpus de récits de cauchemars » », *Langage et société* 66 (1), 1993, pp. 5-39
- *Mondes lexicaux*: « Il s’agit, non pas de comparer les distributions statistiques des “mots” dans différents corpus, mais d’étudier la structure formelle de leurs cooccurrences dans les “énoncés” d’un corpus donné. »
  - Analyse du discours/speech analysis
  - ‘Mondes’: to be understood as social representations
  - ‘Lexical worlds’ are opposed to one another

# Text-mining: similitude analysis

---

- How the words relate to each other? How are they connected?
  - Clustering is a way to see differences between different lexical worlds
  - Similitude analysis is a way to see how words are linked to each other

# Network visualization

---

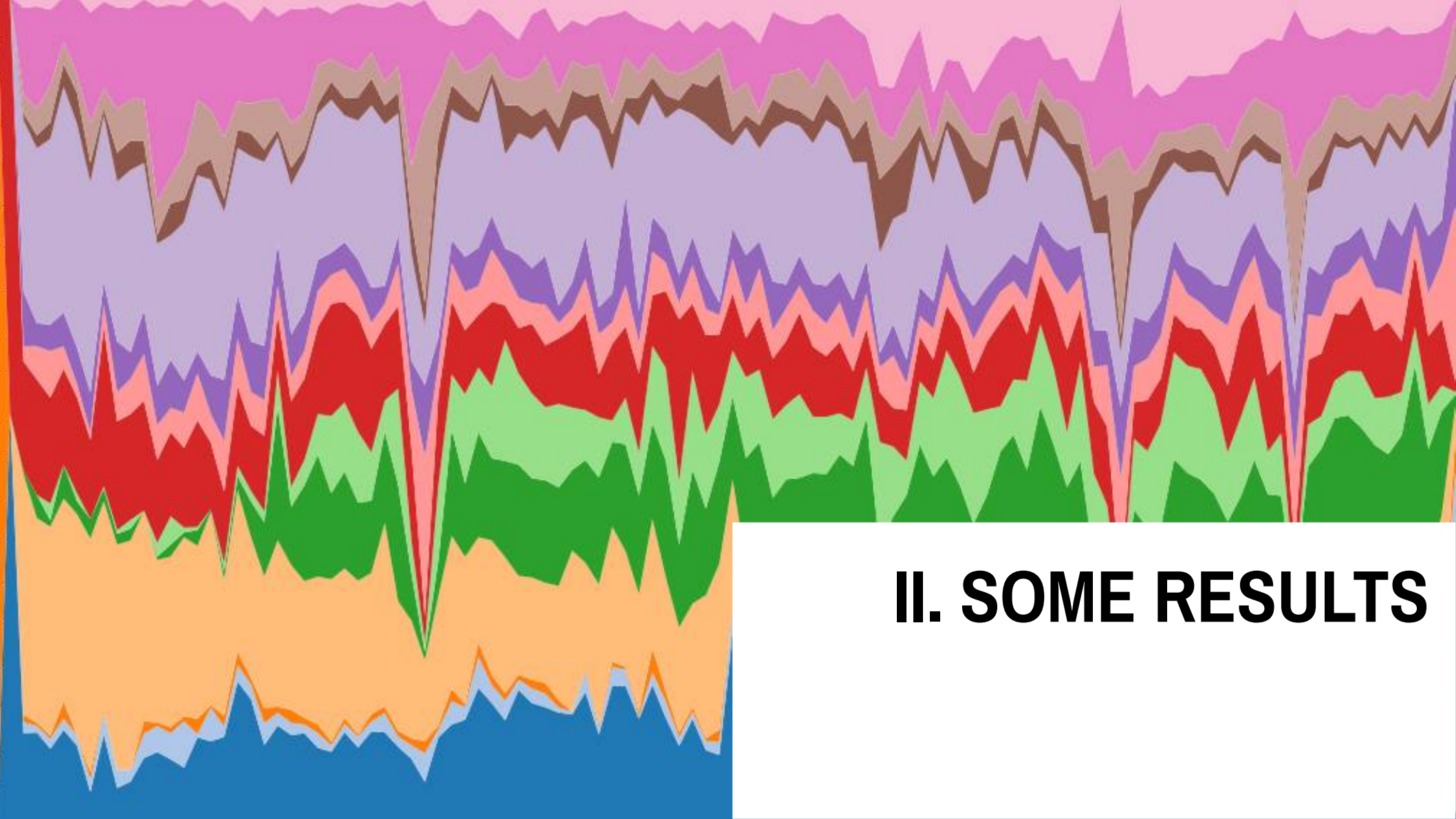
*Note : I am not pretending to do network analysis*

- Gephi
- In the case of network visualization, the difficulties are the following:
  - It requires a sense of aesthetics (not that important)
  - It requires to study sociological studies that are grounding it

# Other and less used software

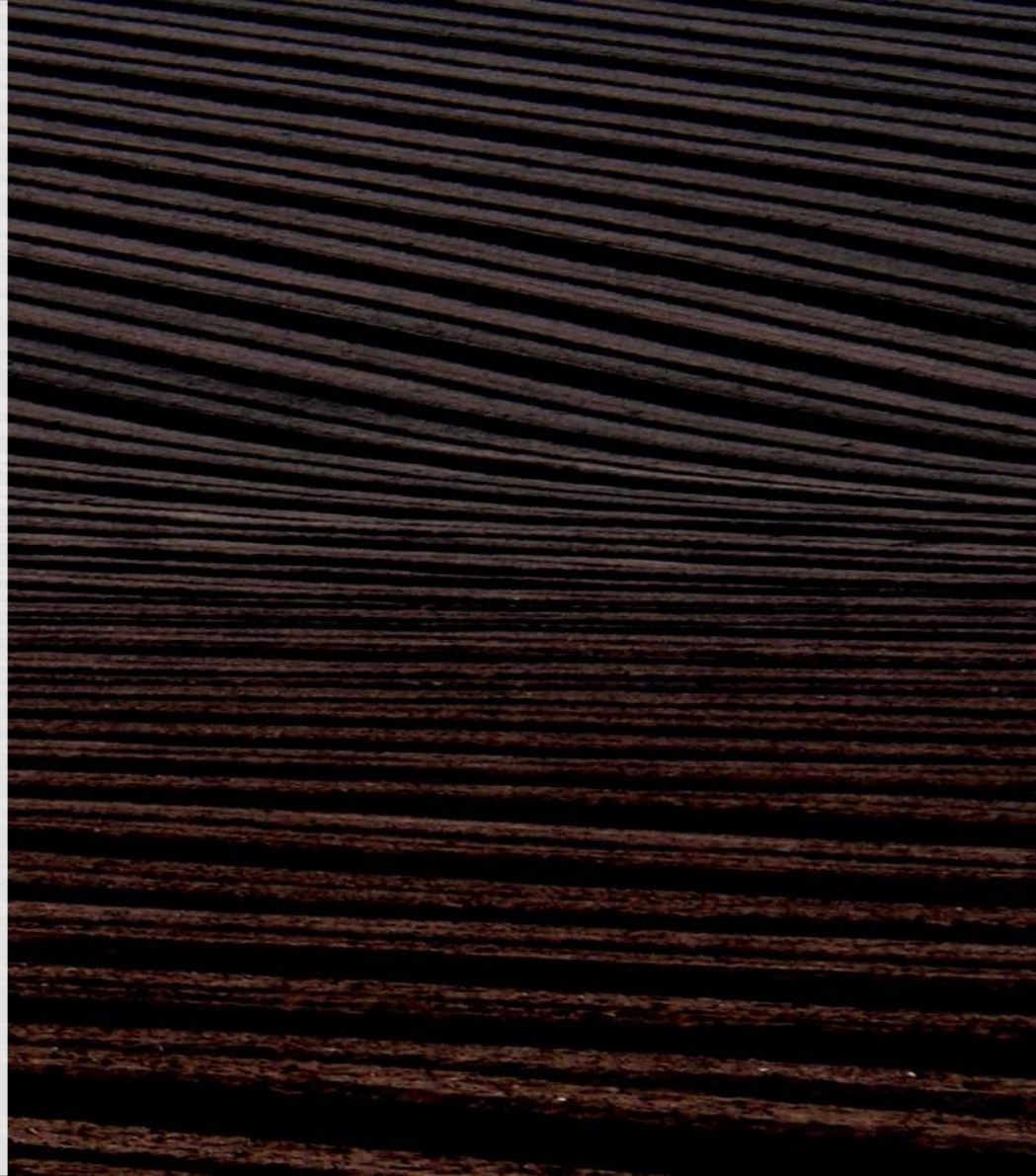
---

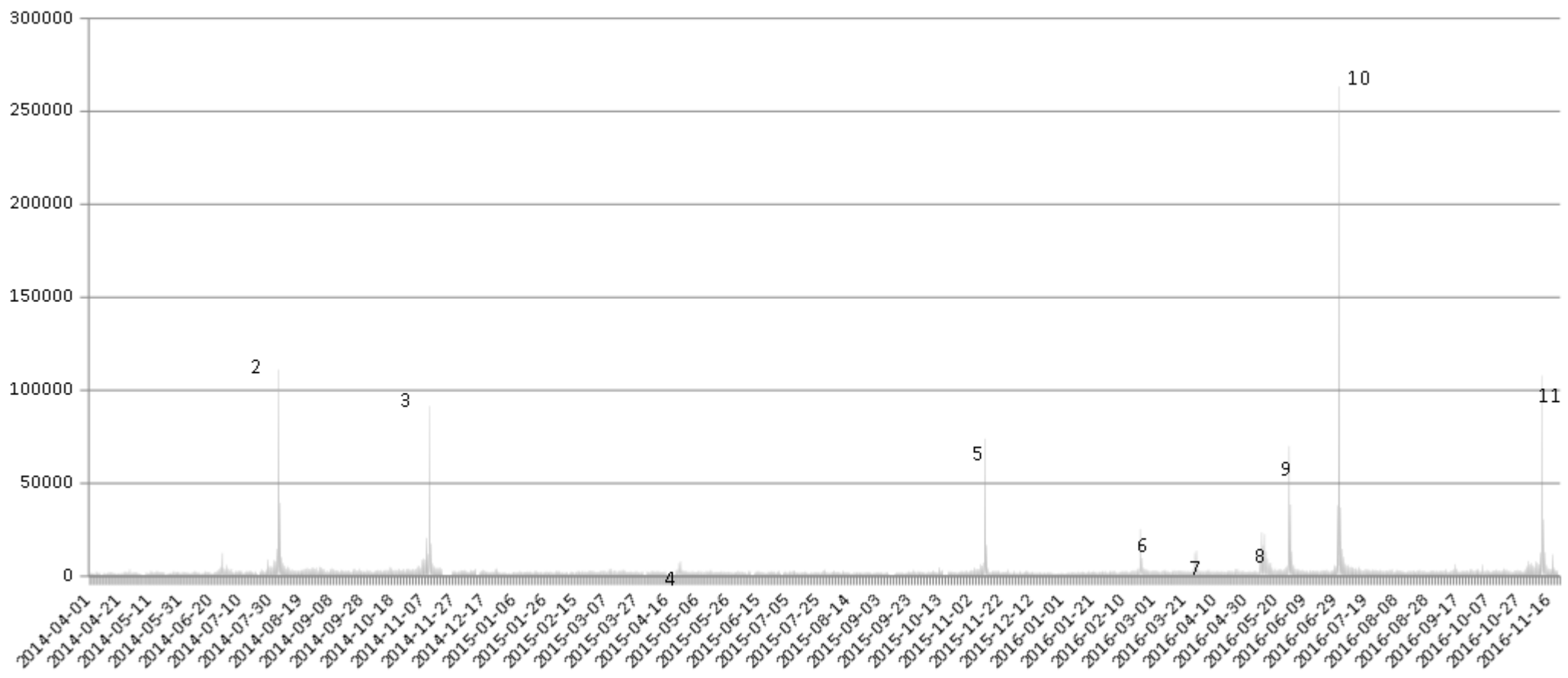
- Tableau Public
- Tropes
- VoyantTools
  - No lemmatisation, a lot of gadgets' dataviz
  - Better with structured data (XML-TEI)
- Some tries with MALLET
  - Difficult to interpret



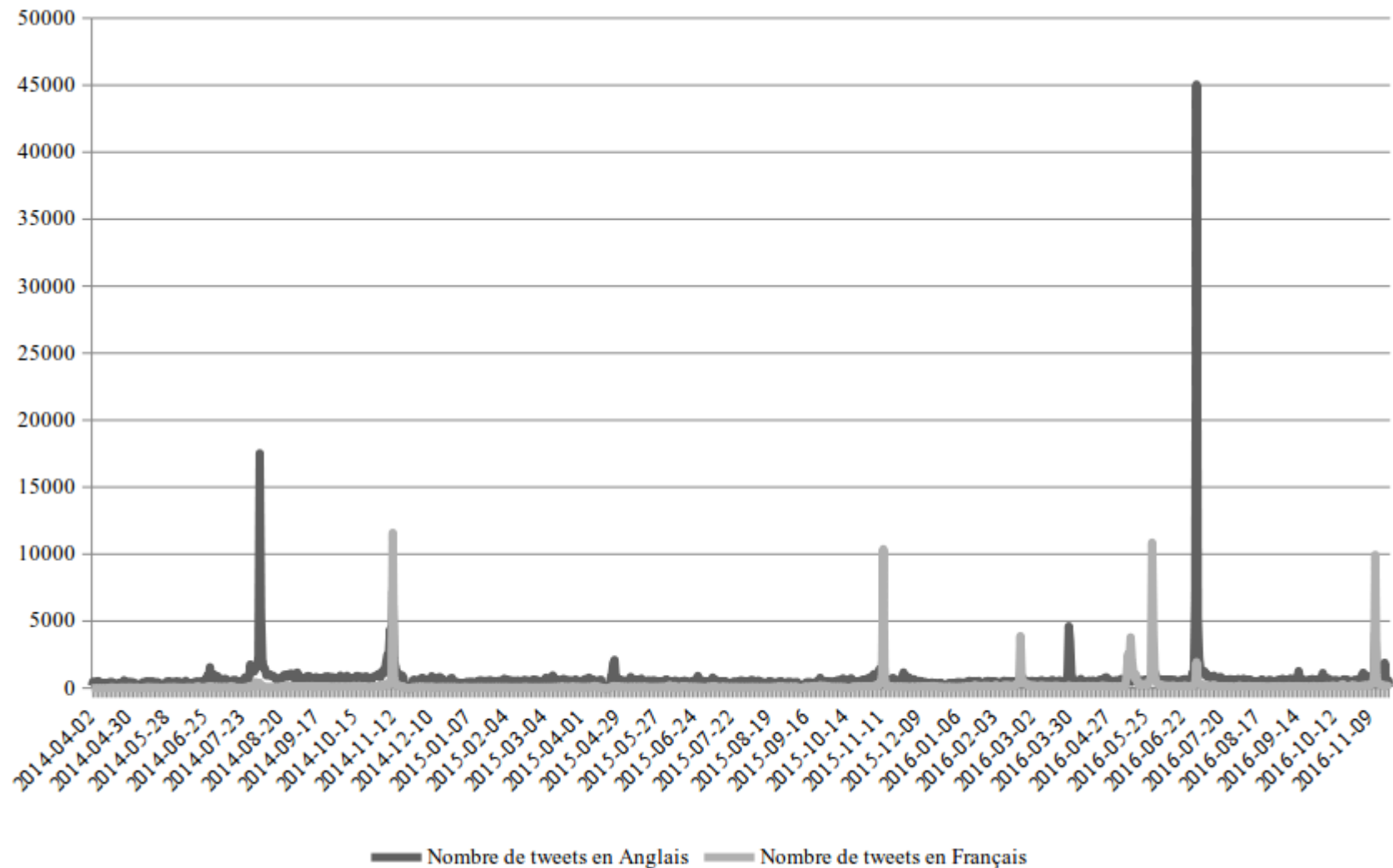
## II. SOME RESULTS

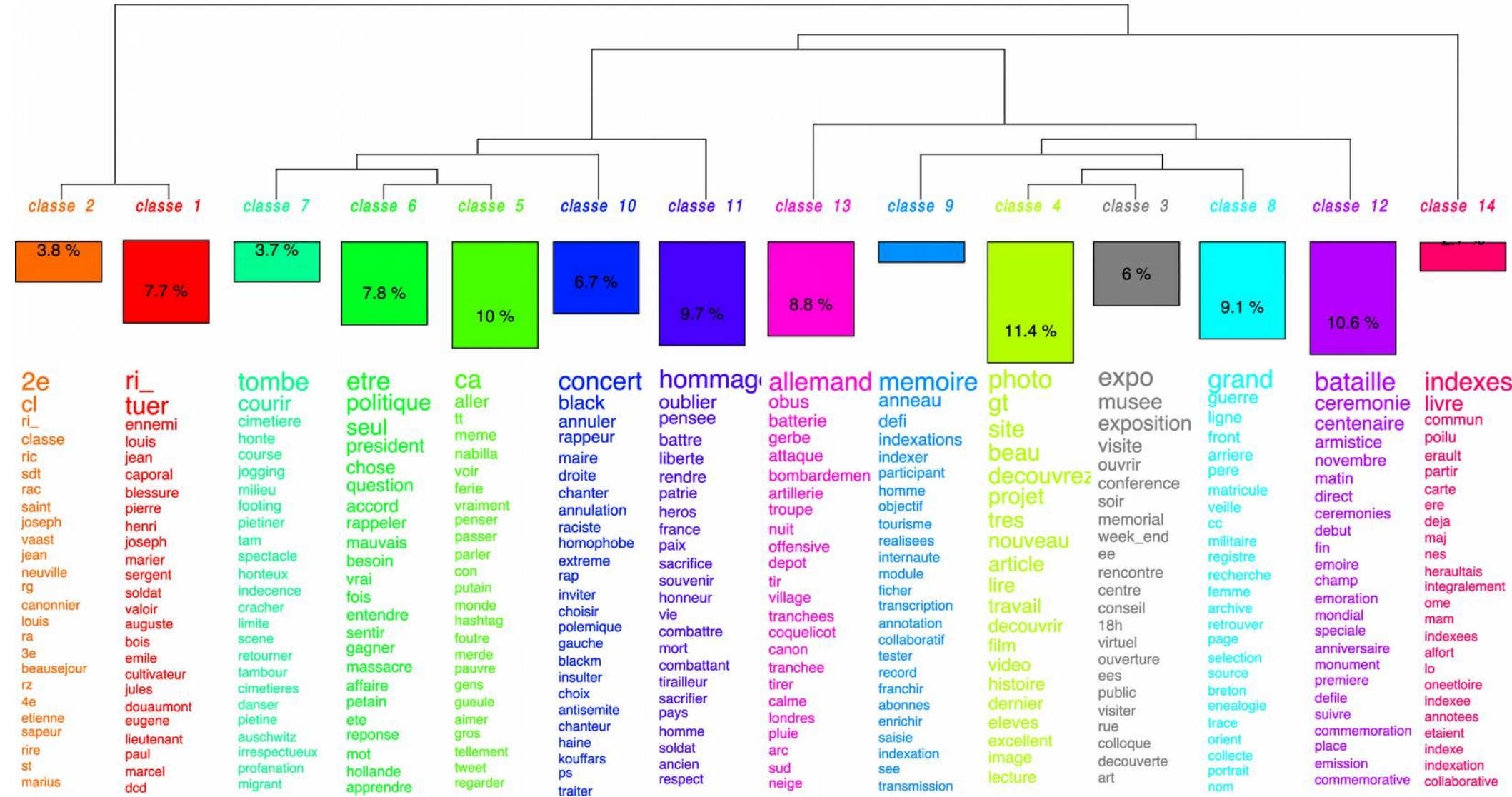
# A. Topics & Temporalities

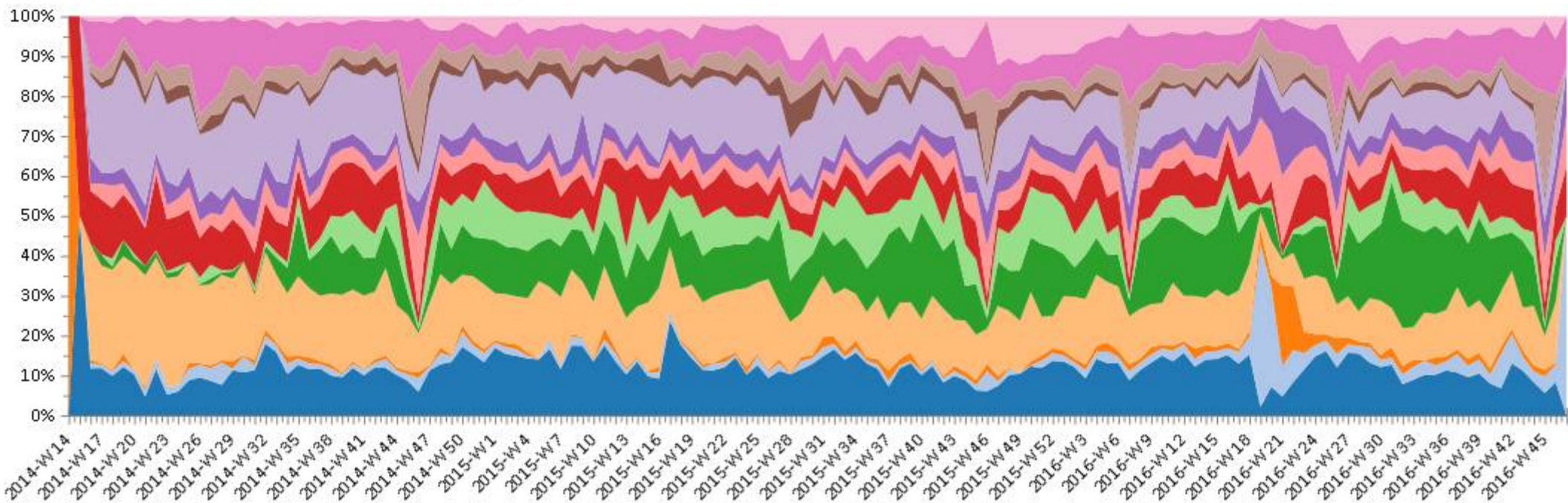






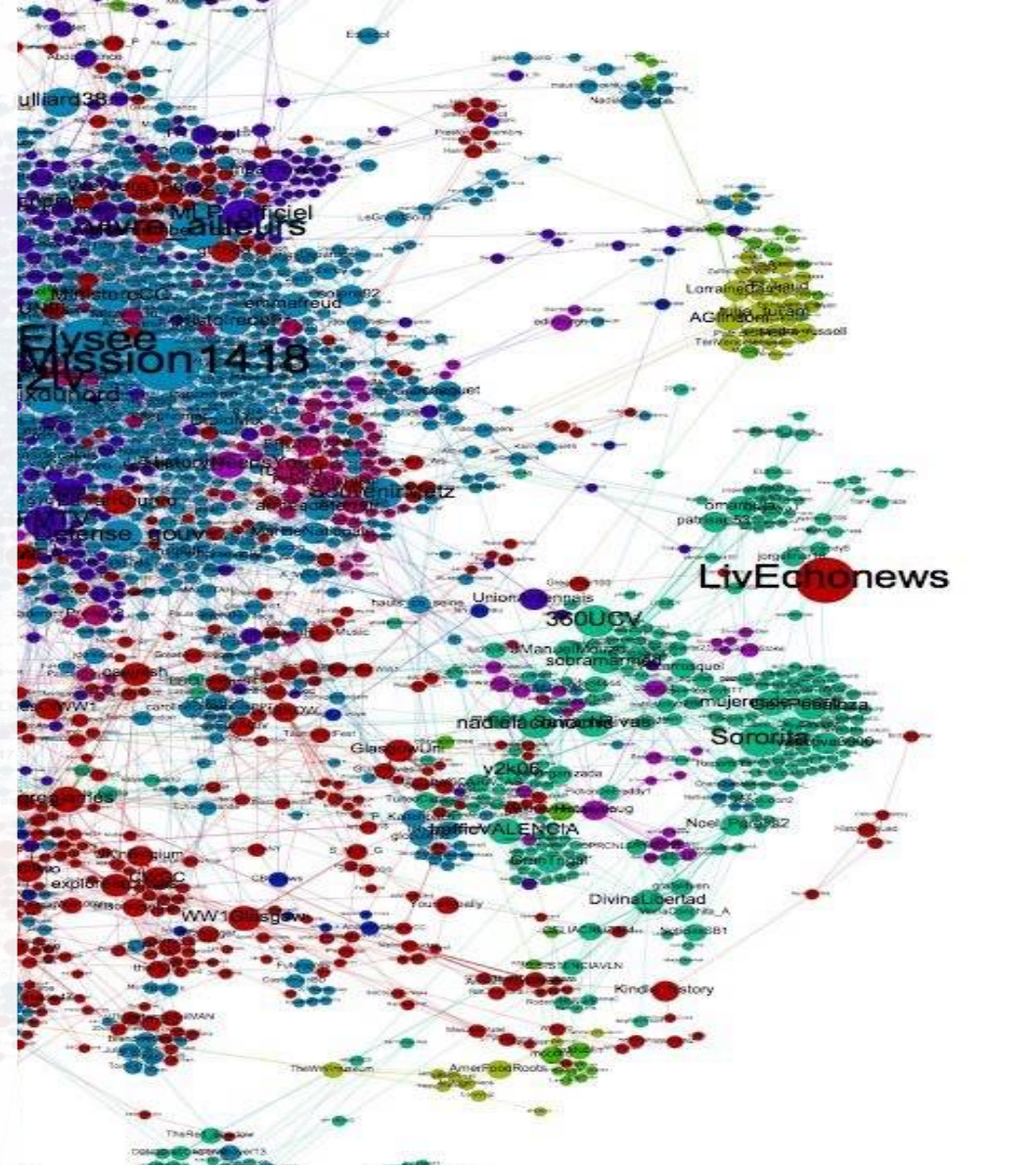






- allemand, obus, batterie (classe 13)
- RI, tuer, ennemi (classe 1)
- être, politique, seul (classe 6)
- bataille, cérémonie, centenaire (classe 12)
- concert, black, annulé (classe 10)
- 2e, cl, RI (classe 2)
- grand, guerre, ligne front (classe 8)
- indexé, livre, commun (classe 14)
- tombe, courir, cimetière (classe 7)
- Expo, musée, exposition (classe 3)
- mémoire, défi, indexation (classe 9)
- Photo, site, beau, projet (classe 4)
- férié (classe 5)
- hommage, oublier, penser (classe 11)

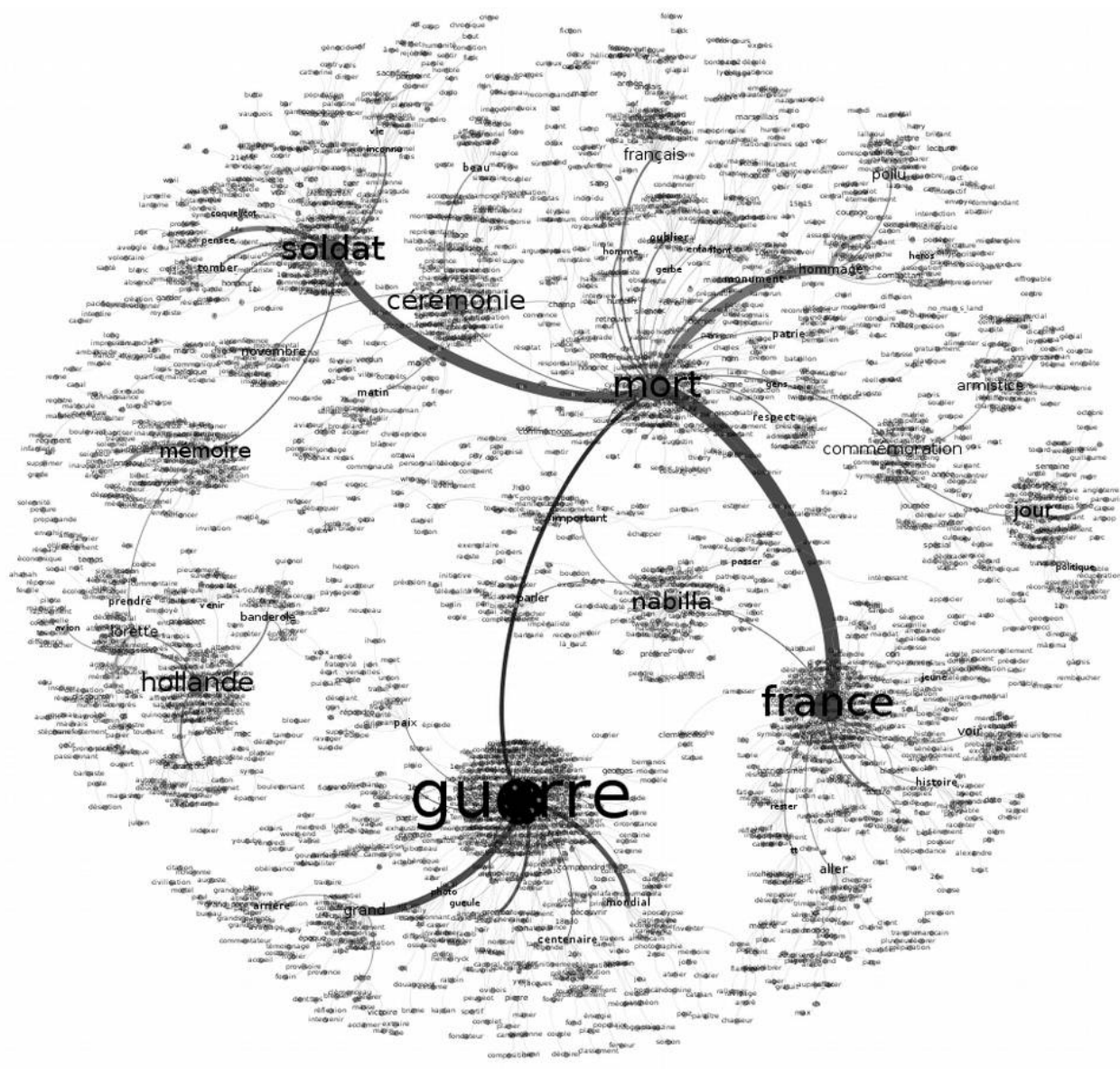
# B. Networks Viz



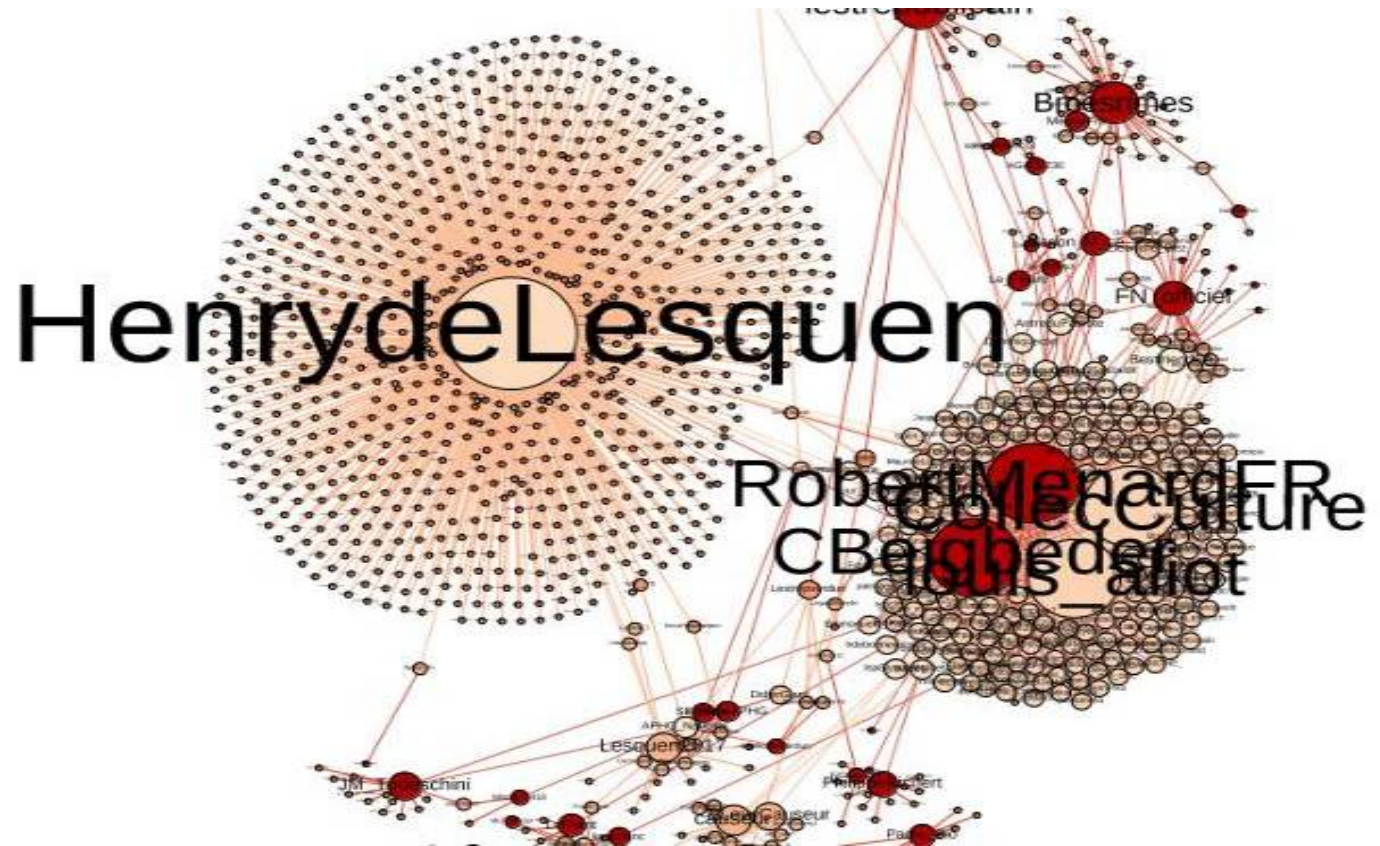




# Networks of words

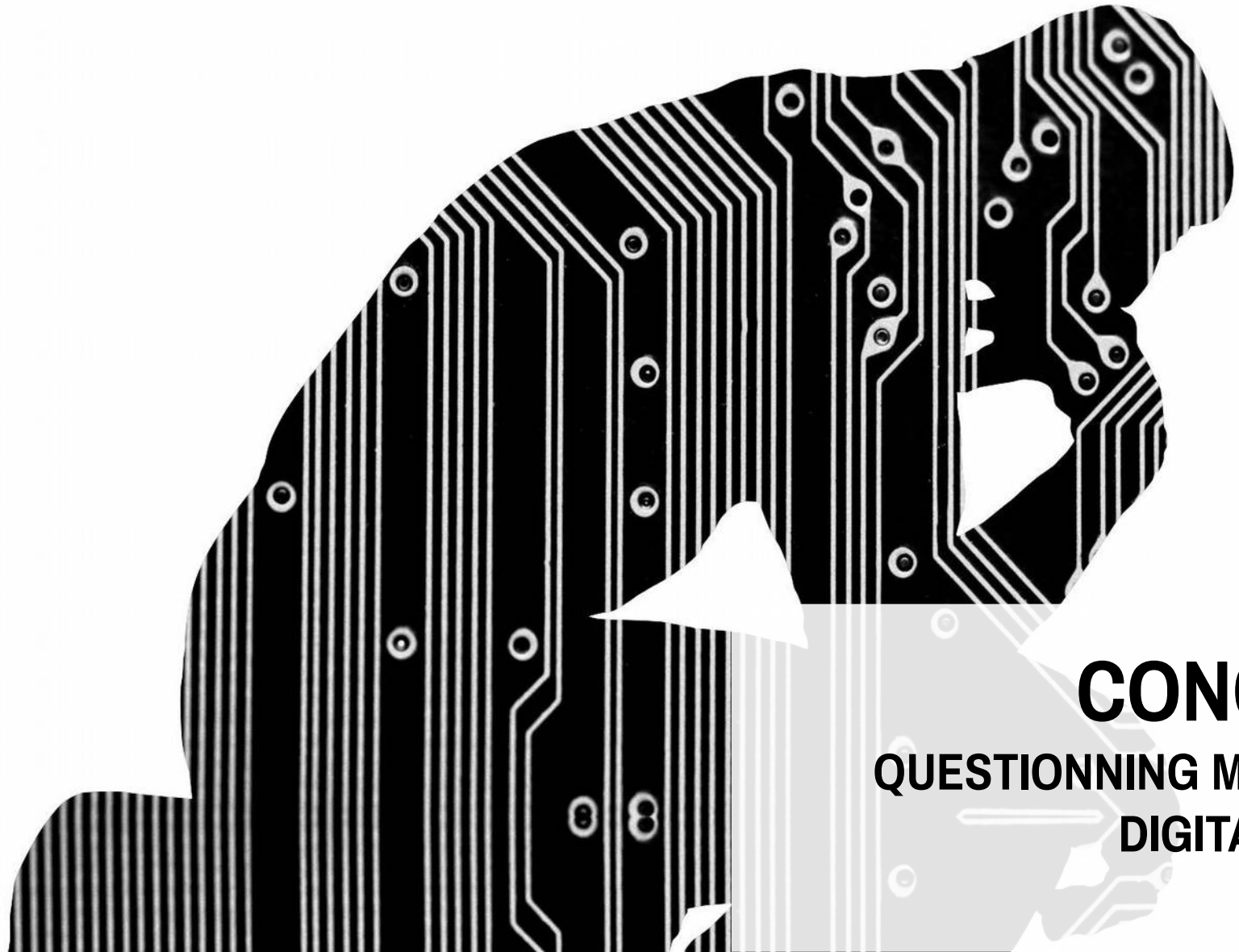


Black M's cancelled concert at Verdun



Intense “Wavelets” (D. Boullier)





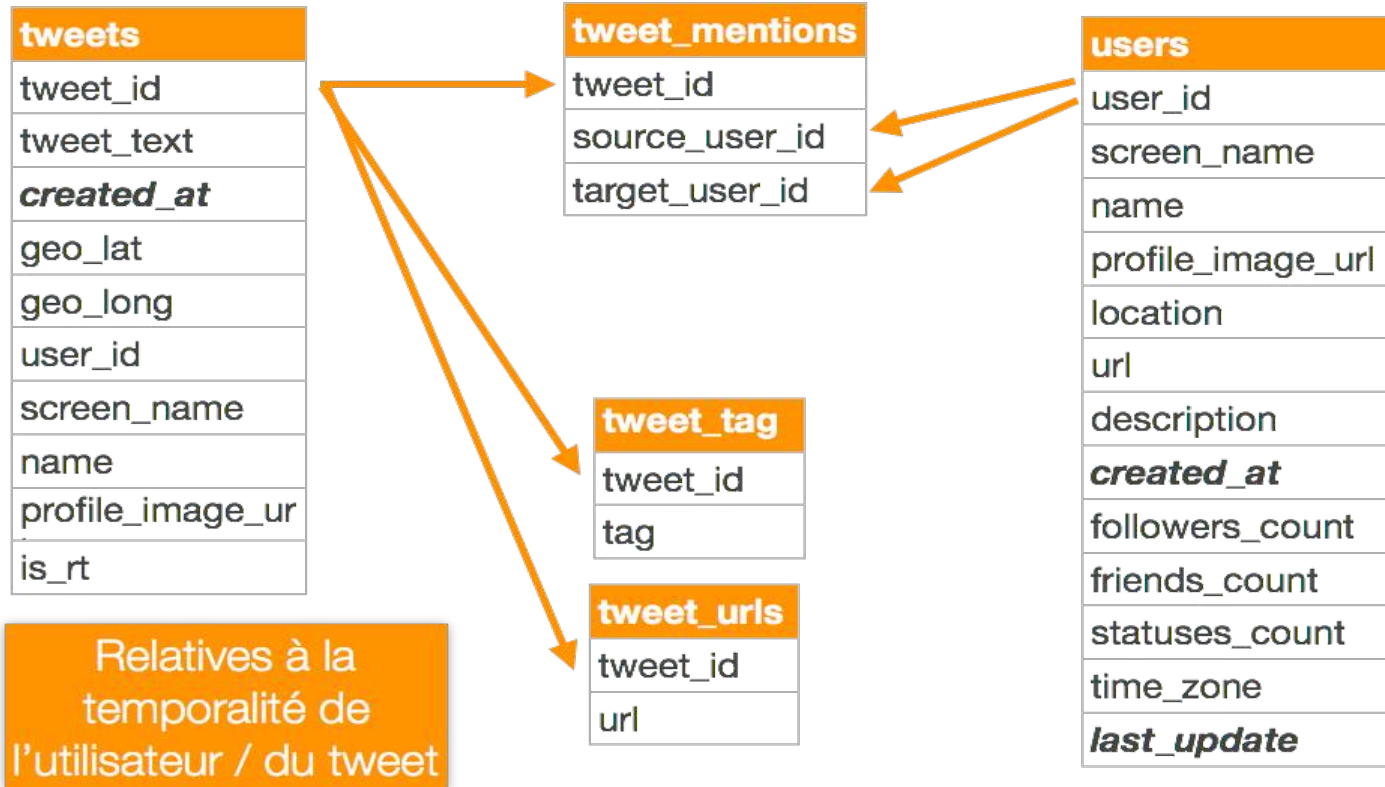
**CONCLUSION**  
**QUESTIONNING MY RESEARCH :**  
**DIGITAL BRICOLAGE**

# Digital Bricolage

---

- *La pensée sauvage* (1962), Claude Levi- Strauss (anthropology)
  - Criticism : see Ricoeur (1963) or Derrida
  - Can be used to understand social innovation
- Intellectual bricolage : concrete thinking allowing social organisation and collective rebalancing, when scientific thinking can lead to destabilization of a social order
- Digital bricolage is hence here understood as an (academic) answer to technological disruption
  - How to carry on your research, while tools, methods, and even primary sources (its form and its volume) are fastly changing whereas you are not able to read / understand all the literature you should read and understand
  - In concrete terms : how to choose a tool, how to use it, how to know its limitations, how to be aware of your own technical, methodological and epistemological limitations while still doing research

# The poverty of time-based metadata



# Neither computing nor statistics but...

---

- Thinking about born digital primary sources
- Why Twitter?
  - Because we can: A rather open API system
- Would have needed a developer for other kinds of source
- Risks
  - Algopol and Facebook: arbitrary politics of APIs
  - My aim is to collect tweets up to 2019 (Centenary of the Versailles Treaty)
  - Twitter might change or shut down its API, might disappear...

# Limits of « home-based » Big Data analysis

---

- Big Data from a historian's point of view
  - When Gnip Inc plays with 'small data', they handle 5 to 6 million tweets...
- Those pieces of software have a limited ability to analyse massive data corpora
  - Are their way to do statistics outdated with regards to today's massiveness of data?
- Questions the historian's training
- Questions her status in the historical narrative / social memory production chain

# How to go through the data analysis jungle?

---

- Too many tools
  - Too many unflexible tools
- Too many tools that do not answer researcher's needs
  - An example: based on words and not on expressions/groups of words
- Too many tools that are standardizing research

# How to understand weak signals?

---

- All my analyses are about *Poilus* (France) or battlefields (UK)
- What about all the other ones?  
Women, prisoners, inhabitants of occupied lands, soldiers from colonies, sentenced to death, dissenters...
- Are they subjects of memories for smaller communities that my tools (my methods) are not able to see?
- What about weak signals? How to see snippets within the feed of information?