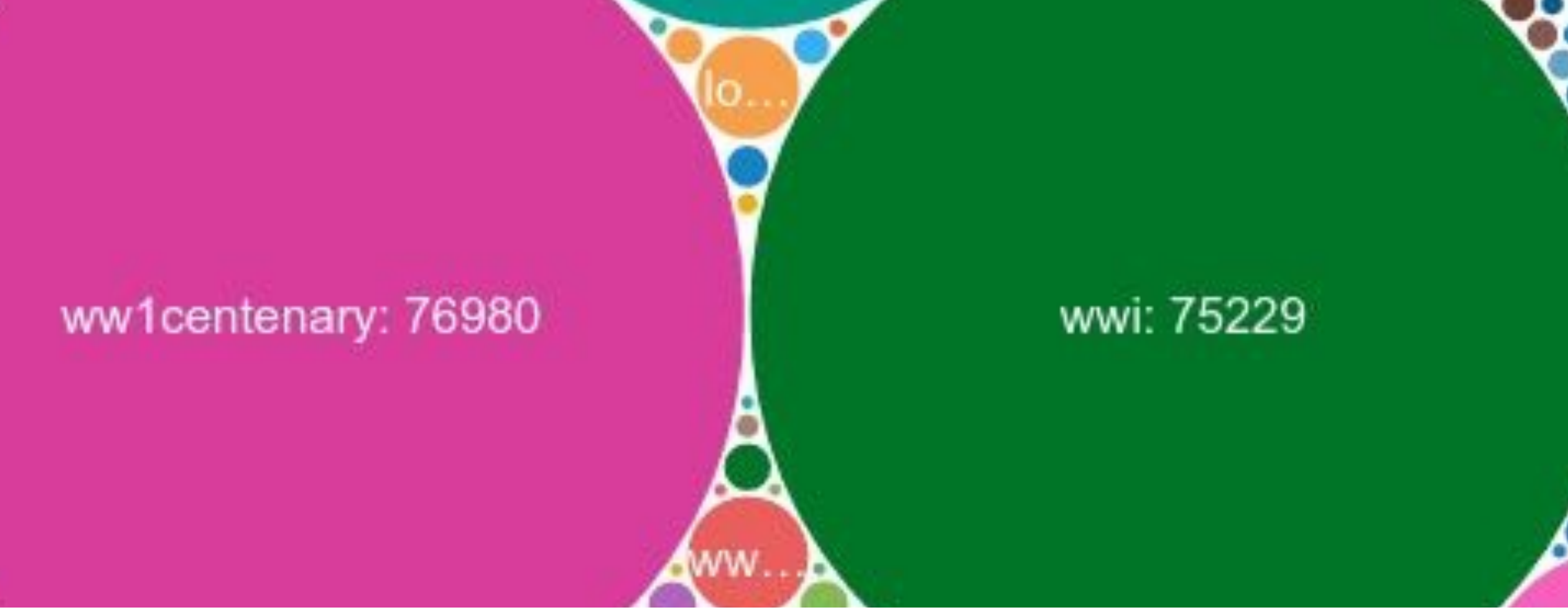


#ww1. Feedback from a data-driven research led by a non-data-scientist historian

Frédéric Clavert – Université de Lausanne
frederic.clavert@unil.ch - 26 mai 2016 - MASHS

INTRODUCTION

A bit of self-history

- Trained as a narrative historian
  coming from (mostly narrative) political sciences
- International/monetary history
- Long-running interest for computing
  - Database
- Turned to Digital Humanities starting from 2008
  - European integration history digital library

Genesis of a project: learning to collect Tweets

- Use of Twitter since 2008
- REAL use of Twitter since 2009
  - Academic interest: following several sessions of the same conference
  - Talking with other researchers
  - Interdisciplinary

Genesis of a project: collecting #ww1

- 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)
- Strong media exposure from time to time
  - Example: 1990s and Vichy
    cf. 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 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

ww1, wwi, wwiafrica, 1gm, 1GM, 1wk, wk1, 1Weltkrieg, centenaire, centenaire14, centenaire1914, GrandeGuerre, centenaire2014, centenary, fww, WW1centenary, 1418Centenary, 1ereGuerreMondiale, WWIcentenary, 1j1p, 11NOV, 11novembre, WWI, poppies, WomenHeroesofWWI, womenofworldwarone, womenofww1, womenofwwi, womenww1, ww1athome, greatwar, 100years, firstworldwar, Verdun, Verdun2016, Somme, PoilusVerdun

The current state of the corpus

- 1 April 2014 - 13 April 2016
- 2,096,968 tweets
  - Around 2/3 of retweets
- 542,570 Twitter accounts
  - private individuals, institutions, project-based account, bots and many others

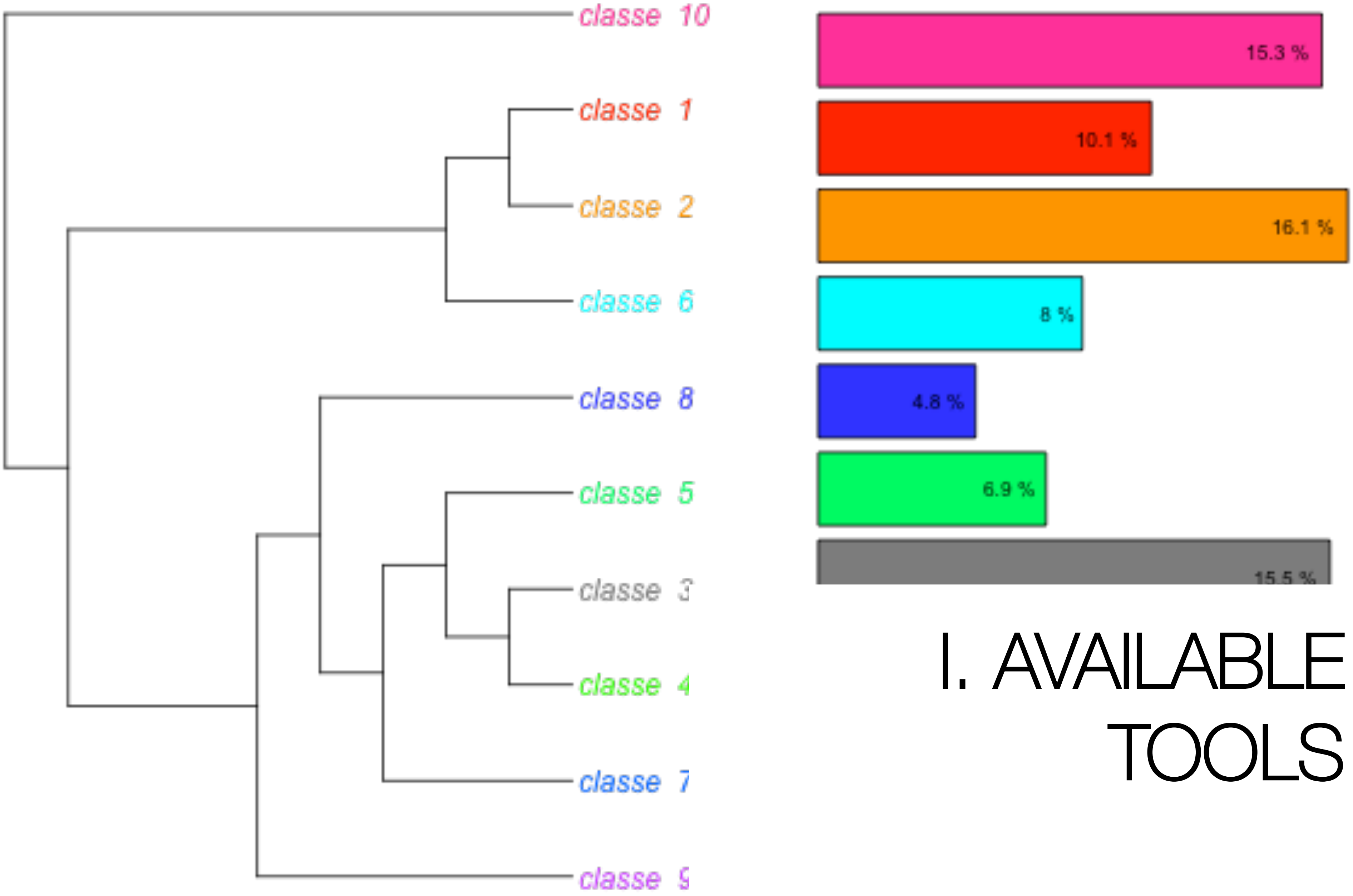
- 124,424 hashtags
  - 54,566 used only once
  - 84,936 three times or less
  - 107,047 tenth times or less
- Not a lot of noise, except for three hashtags: #11Nov/Verdun/Somme

A historian facing a sea of data

How to read two 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 litterature
  - *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 tools
    - number of tweets per day for instance
  - Data / Text mining
  - Network analysis
- Imply to deal with not-that-structured data

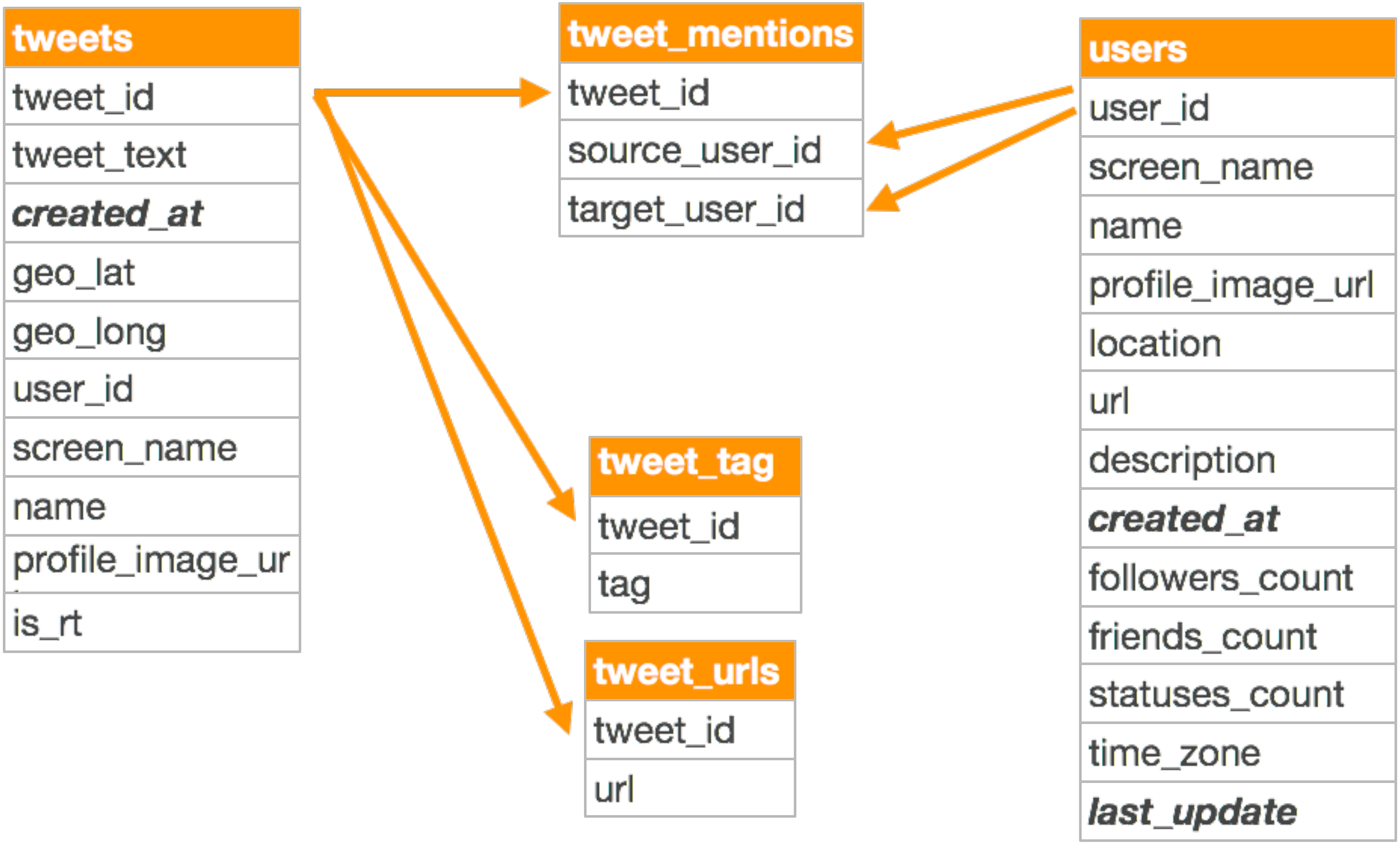
Numerous tools are at our disposal…

- …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 ground 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?

Tools for #ww1: harvesting data

- LAMP server
  - From a home-based server to a more professionalised one
- PHP script: 140dev.com
  - Other candidates: DMI-TCAT
  - 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

Storing data: data-model



tweets
tweet_id
tweet_text
created_at
geo_lat
geo_long
user_id
screen_name
name
profile_image_ur
is_rt

tweet_mentions
tweet_id
source_user_id
target_user_id

tweet_tag
tweet_id
tag

tweet_urls
tweet_id
url

users
user_id
screen_name
name
profile_image_url
location
url
description
created_at
followers_count
friends_count
statuses_count
time_zone
last_update

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

Text-mining

- 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
- Grounded on the French researcher Max Reinert's 'mondes lexicaux'

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

Un énoncé traduit donc davantage un point de vue particulier plutôt qu'une représentation, le point de vue : impliquant en son centre l'existence d'un "sujet' dans une certaine modalité du faire ou de l'être. Cette notion n'a donc rien d'absolu. Elle est relative à l'activité ou à l'état d'un sujet et ne renvoie pas obligatoirement à un système de références préétabli, celui-ci pouvant l'être ou non, reconstruit, imaginé. . . Notre hypothèse principale consiste justement à considérer le vocabulaire d'un énoncé particulier comme une trace pertinente de ce point de vue il est à la fois la trace d'un lieu référentiel et d'une activité cohérente du sujet-énonciateur. **Nous appelons mondes lexicaux, les traces les plus prégnantes de ces activités dans le lexique.**

Text-mining: similitude analysis

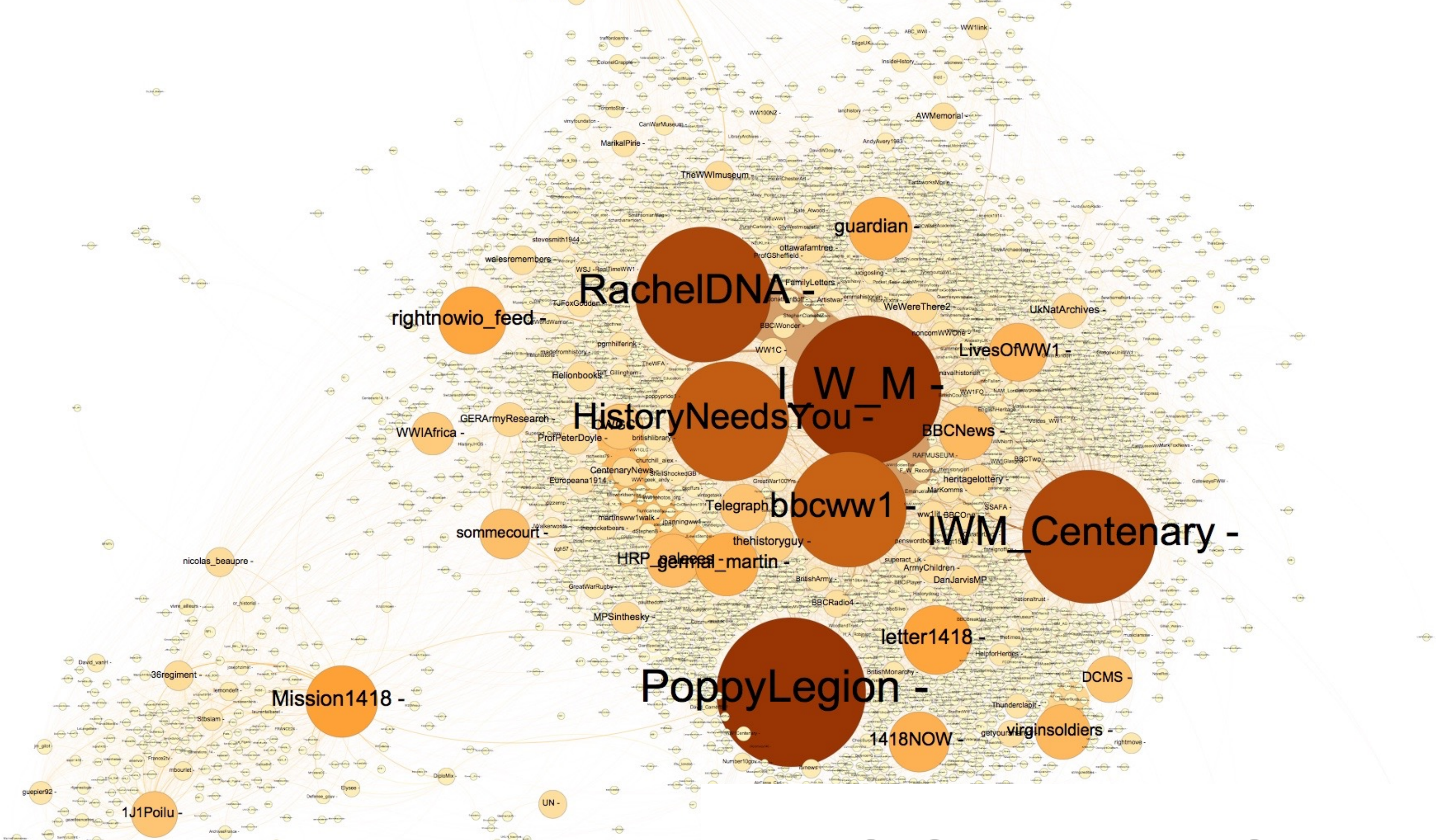
- How the words relate to each other? How are they connected?
  - Clustering is a way to see differences between different lexical worlds
  - Similitude analysis is a way to see how words are linked to each other

Network analysis

- Gephi
- In the case of network analysis, the difficulties are the following:
  - It requires a sense of aesthetics
  - It requires to study sociological studies that are grounding it

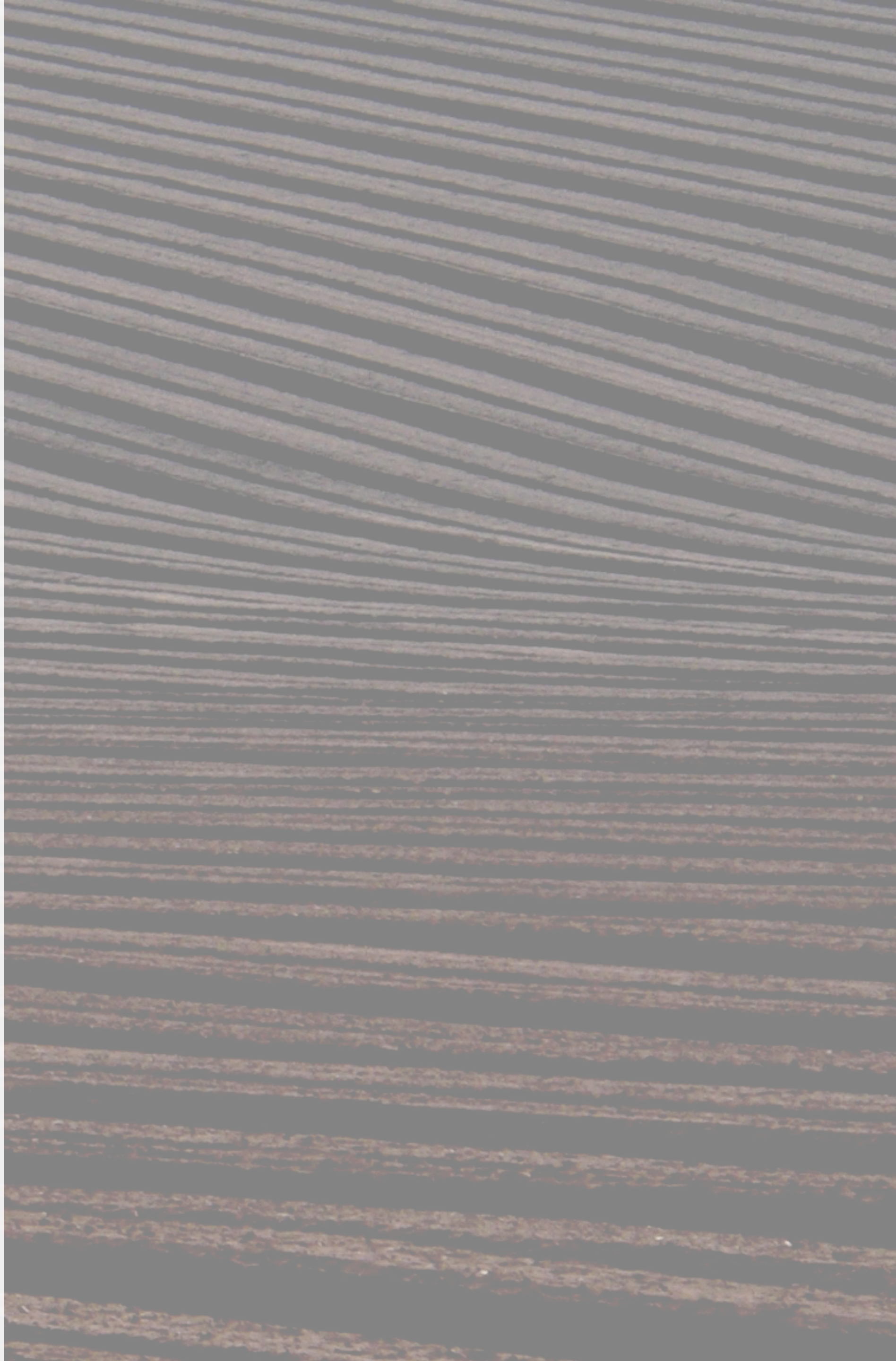
Other and less used software

- 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



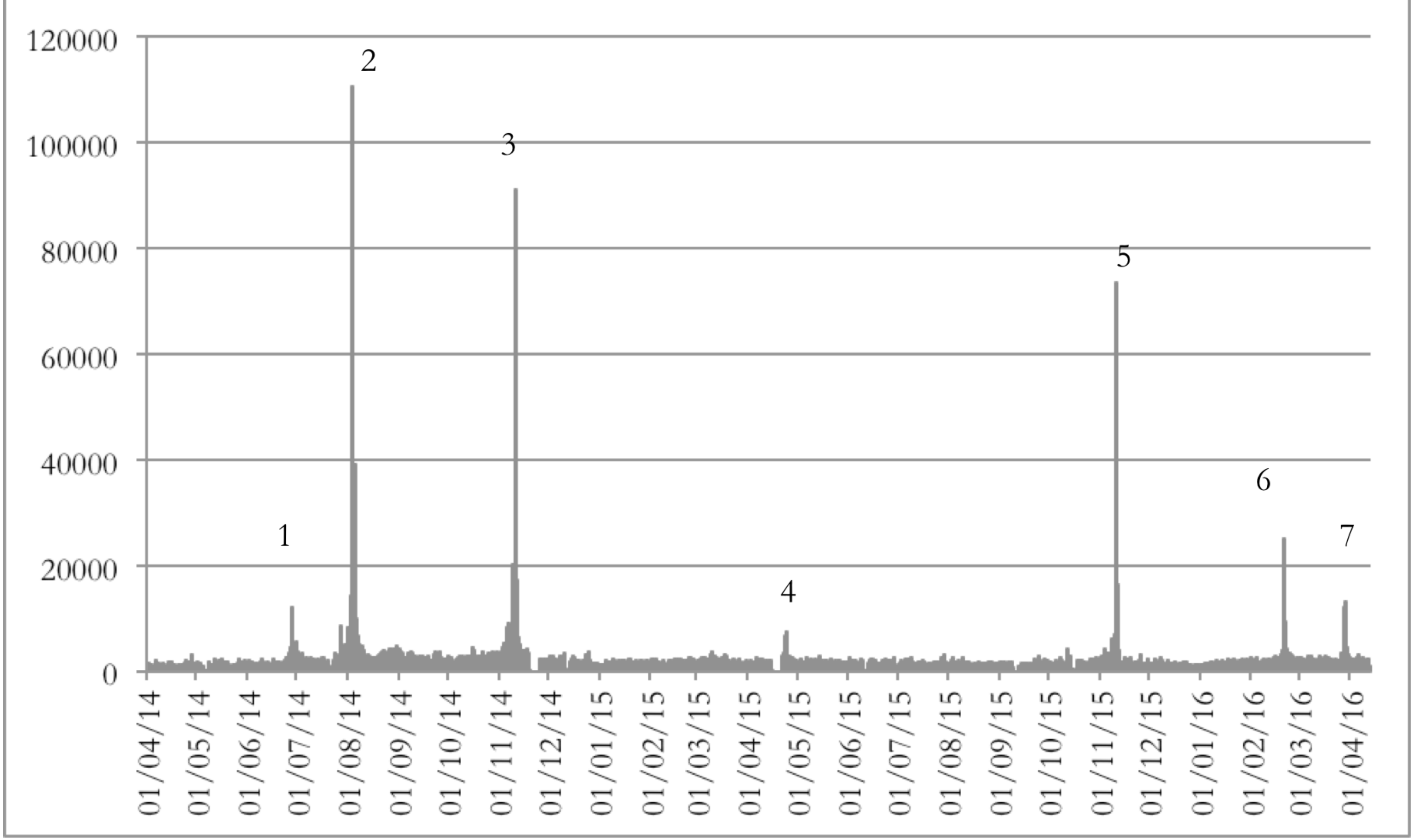
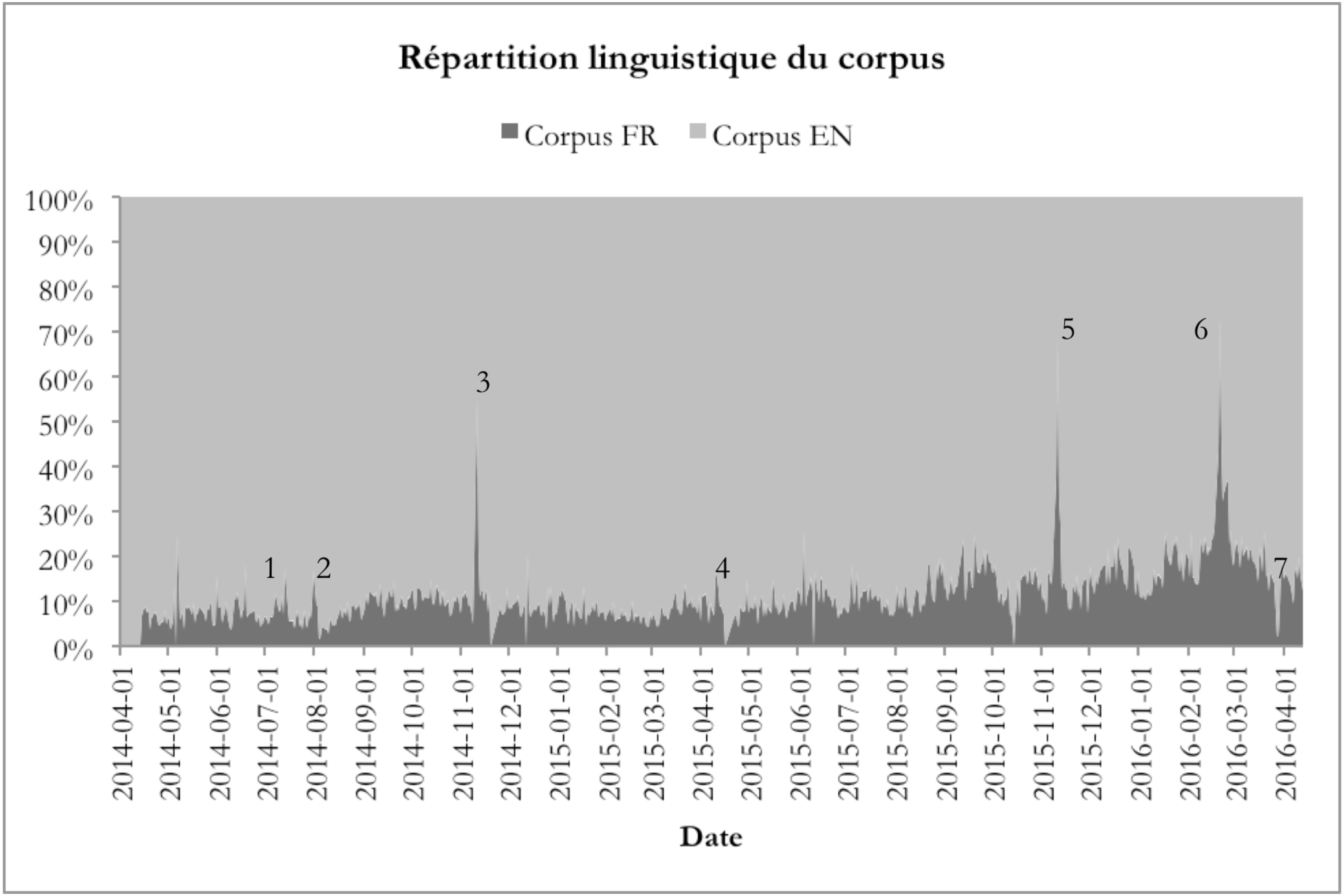
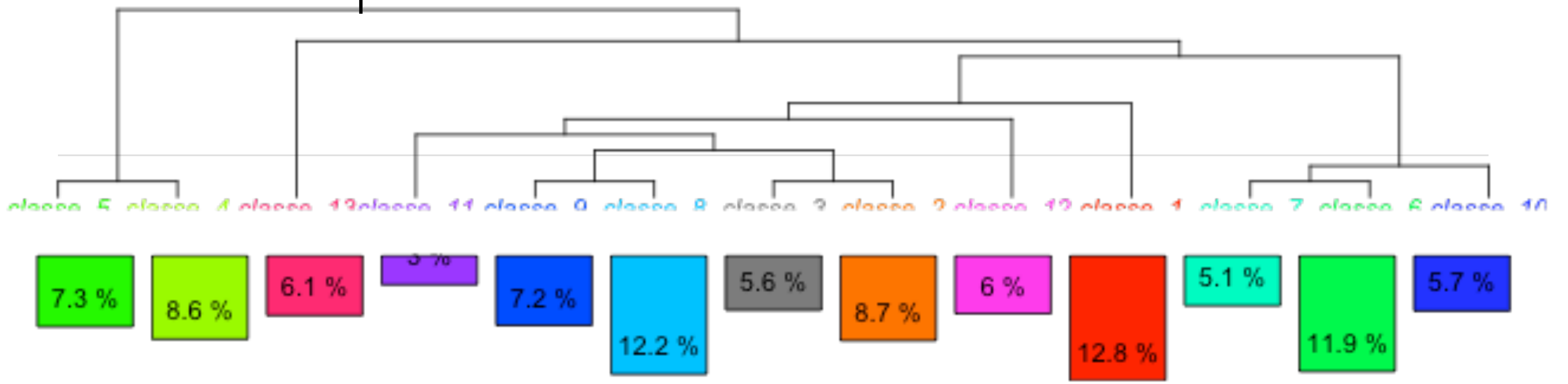


Figure 1 - Nombre de tweets par jour

1. Centenaire de l'assassinat de l'Archiduc François-Ferdinand, 28 juin 2014, 12 029 tweets ;
2. Centenaire de l'entrée en guerre du Royaume Uni, 4 août 2014, 110684 tweets ;
3. Commémoration de l'armistice, 11 novembre 2014, 91108 tweets ;
4. ANZAC Day, 25 avril, 7605 tweets ;
5. Commémoration de l'armistice, 11 novembre 2015, 73 520 ;
6. Centenaire du déclenchement de la guerre de Verdun, 21 février 2016, 25 011 tweets ;
7. Easter rising, 28 et 29 mars 2016, 12021 et 13299.


Répartition linguistique du corpus
Corpus FR
Corpus EN
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
2014-04-01
2014-05-01
2014-06-01
2014-07-01
2014-08-01
2014-09-01
2014-10-01
2014-11-01
2014-12-01
2015-01-01
2015-02-01
2015-03-01
2015-04-01
2015-05-01
2015-06-01
2015-07-01
2015-08-01
2015-09-01
2015-10-01
2015-11-01
2015-12-01
2016-01-01
2016-02-01
2016-03-01
2016-04-01
Date

French topics



2e
cl
rire
joseph
sdt
ric
classe
rac
jean
marius
hôpital
louis
caporal
beauséjour
emile
françois
4e
rg
béziers
bc
ra
auguste
16e
mpf
paul
3e
hurlus
dec
canonnier
bcp

tuer
rire
ennemi
naître
louis
dcd
blessure
pierre
marier
décéder
meuse
jean
henri
suite
362e
belgique
coray
douaumont
haumont
neuville
bonneau
alain
vaast
valoir
avril
septembre
caporal
eugène
disparaître
jules

livre
indexés
ficher
commur
indexatic
matricule
poilu
registre
partir
indexées
indexer
collaborativ
défi
mam
indexations
carte
homme
transcrire
maj
mdh
compléter
memoire
participant
inscrire
désormais
entièrement
intégralemer
indexons
icaunais

mondi
premie
guerre
1ère
lecteur
dessiner
inventer
au_delà
londres
jardin
tropical
peugeot
infographie
mooc
bande
fois
allier
lorrain
fin
conflit
invitation
évacuation
quentin
1ere
accepter
webdocumer
conséquence
be
chaîne
canada
apocalypse

lire
lettre
dernier
article
raconter
dossier
nouveau
blog
revue
récit
lecture
numéro
recherche
destin
page
témoignage
courrier
épisode
breton
écouter
historien
roman
corresponda
relire
nouvel
sujet
publier
coeur

grand
guerre
découvrir
exposition
archive
gt
site
photo
expo
journal
histoire
sélection
image
musée
colloque
retrouver
retour
front
document
source
veille
collecte
collection
dessin
travers
fonds
art
ressource
quotidien
civil

soir
novem
dimanch
conféren
prochain
samedi
mercredi
ouvrir
visite
18h
mardi
rdv
rendez_voi
porte
guider
programme
table
soirée
ouvert
manquer
11h
lundi
salle
19h
ronde
cinéma
h
18h30

beau
élève
matin
école
travail
militaire
ville
inviter
maire
cérémoni
projet
année
participer
reportage
cimetière
chanter
lycée
rue
marseillais
voir
public
pari
parisien
mission
remise
musique
hôtel

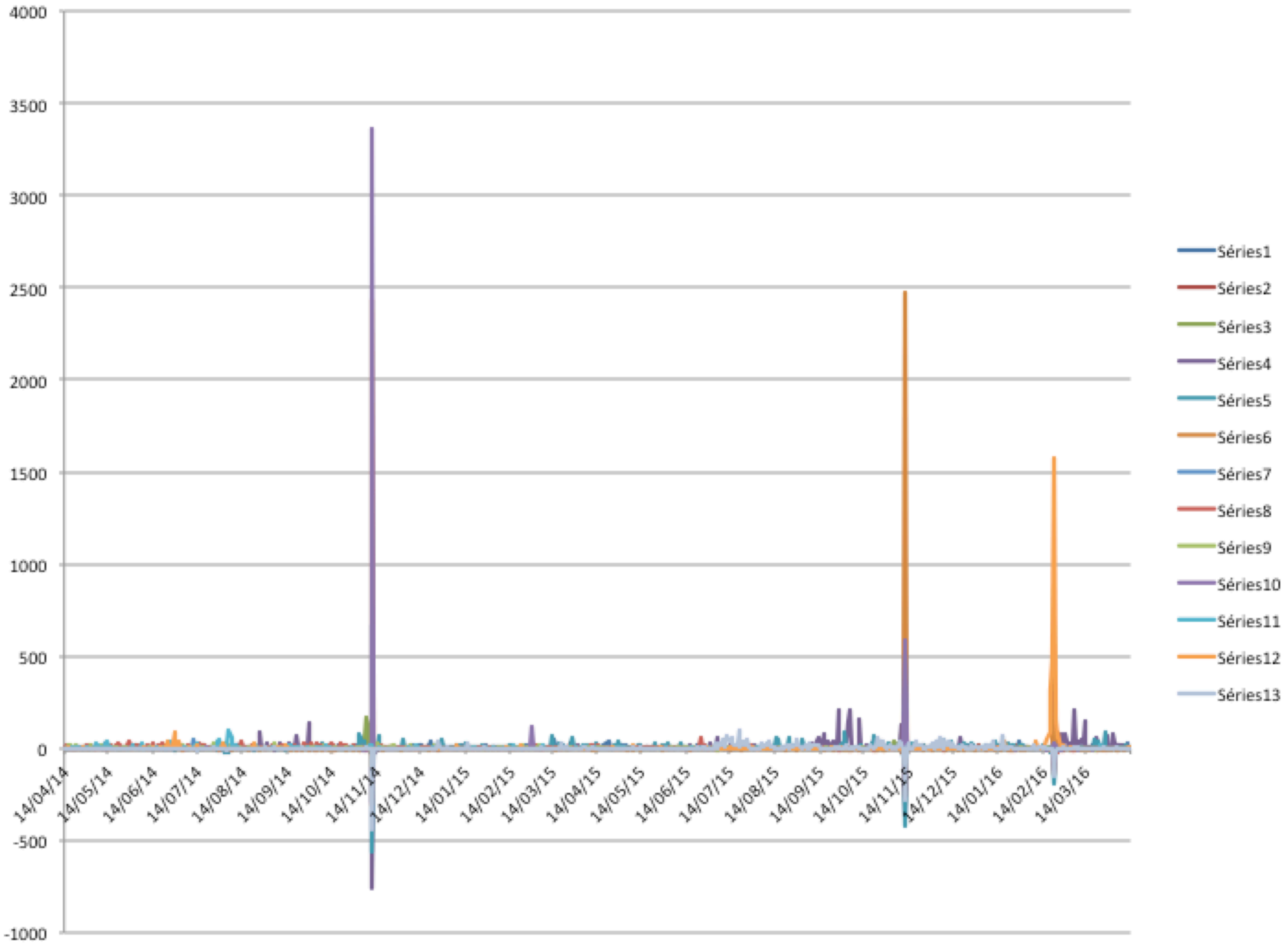
bataille
verdun
champ
enfer
début
direct
centenaire
commémora
débuter
elysées
siècle
symbole
suivre
mémorial
célébration
émission
australien
chiffre
réouverture
mythe
krumeich
ossuaire
rouvrir
élysées
gerd
spécial
live
xxe
jaurès
fresque
joe

allema
obus
tranchée
batterie
nuit
tirer
canon
attaque
coup
bombarden
calme
artillerie
tir
heure
offensive
pluie
7e
troupe
arriver
blessé
neige
bombarder
infanterie
officier
côté
feu
village
repos
occuper
général

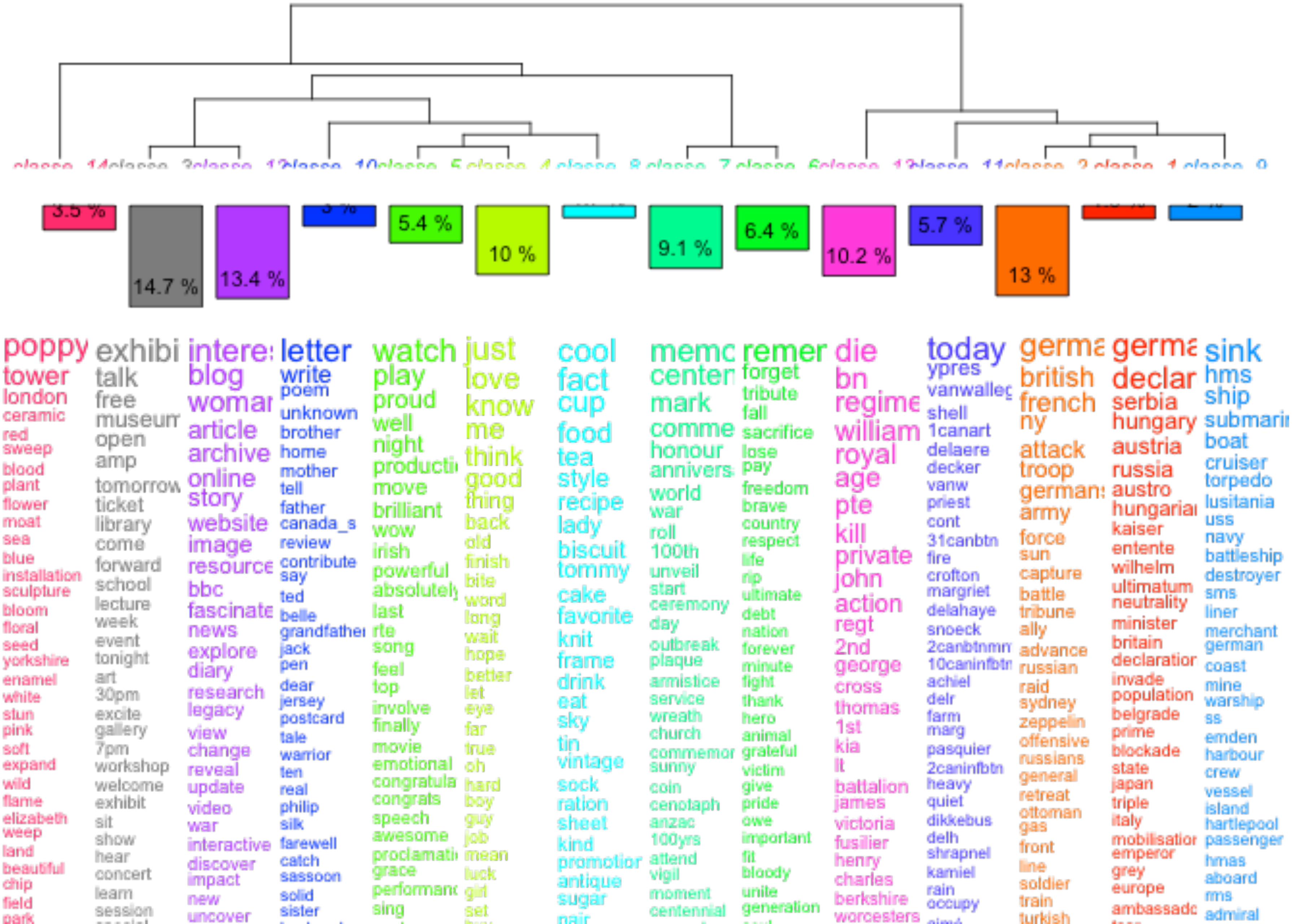
homm
rendre
france
mort
victime
bel
tomber
industriel
combattant
soldat
unir
anatole
poilu
déclarer
royaume
héros
faber
joueur
allemagne
héraultais
patrie
champion
tour
vainqueur
stade
ardenne
driant
mourir
loi
rugbyman
mention

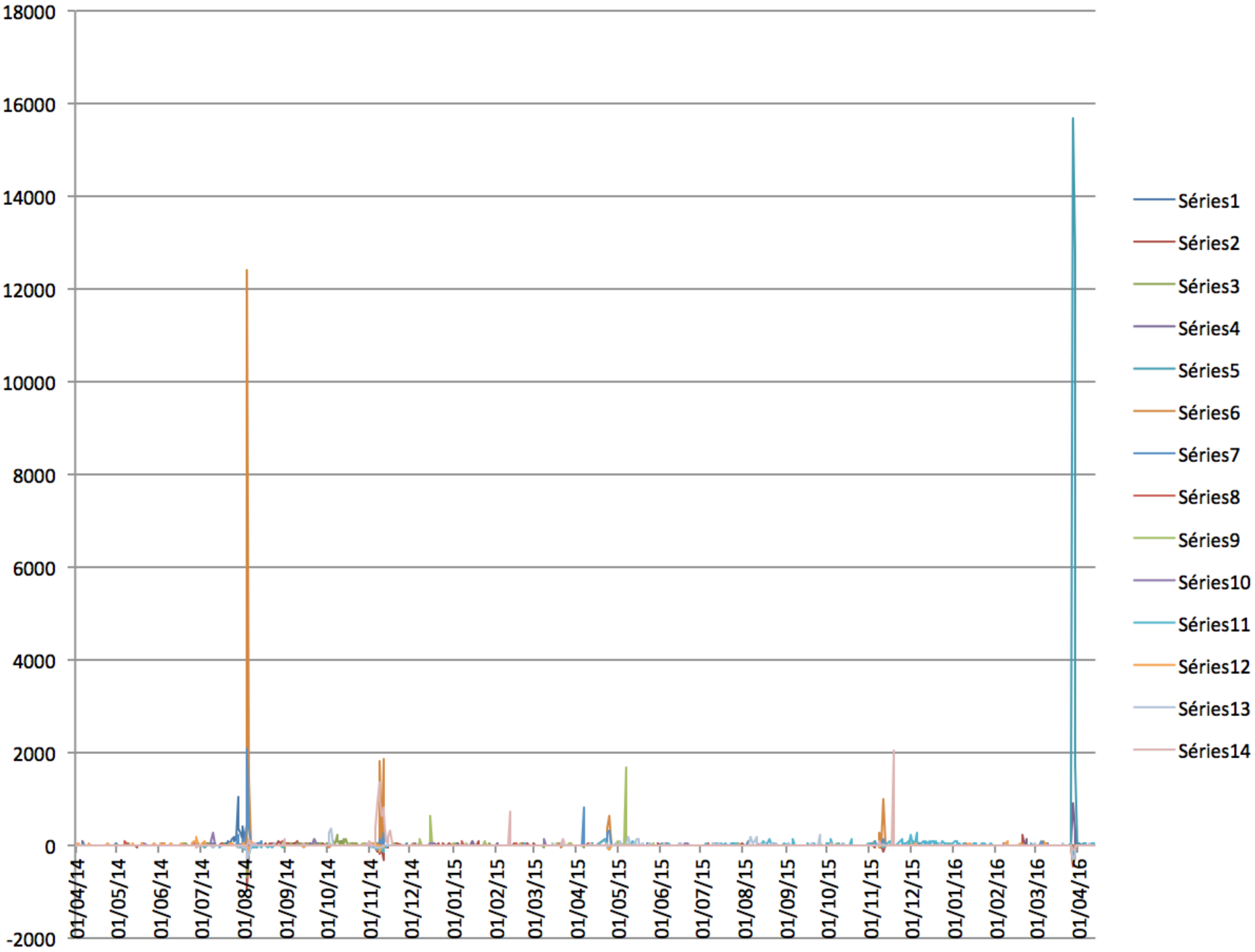
pensé
oublier
battre
souvenir
liberté
paix
pays
france
honneur
respect
sacrifice
patrie
défendre
vie
combattre
héros
père
donner
sacrifier
honorer
sauver
mort
peuple
merci
tomber
français
rester
vivre
génération

férié
nabilla
tt
gens
chose
con
penser
parler
important
vraiment
hollande
monde
foutre
hashtag
putain
aller
préférer
jour
fête
honte
main
pauvre
poignée
passer
gars
respecter
ya
mec
cool


4000
3500
3000
2500
2000
1500
1000
500
0
-500
-1000
14/04/14
14/05/14
14/06/14
14/07/14
14/08/14
14/09/14
14/10/14
14/11/14
14/12/14
14/01/15
14/02/15
14/03/15
14/04/15
14/05/15
14/06/15
14/07/15
14/08/15
14/09/15
14/10/15
14/11/15
14/12/15
14/01/16
14/02/16
14/03/16
Séries1
Séries2
Séries3
Séries4
Séries5
Séries6
Séries7
Séries8
Séries9
Séries10
Séries11
Séries12
Séries13

English topics




18000
16000
14000
12000
10000
8000
6000
4000
2000
0
-2000
01/04/14
01/05/14
01/06/14
01/07/14
01/08/14
01/09/14
01/10/14
01/11/14
01/12/14
01/01/15
01/02/15
01/03/15
01/04/15
01/05/15
01/06/15
01/07/15
01/08/15
01/09/15
01/10/15
01/11/15
01/12/15
01/01/16
01/02/16
01/03/16
01/04/16
Séries1
Séries2
Séries3
Séries4
Séries5
Séries6
Séries7
Séries8
Séries9
Séries10
Séries11
Séries12
Séries13
Séries14

The missing temporality

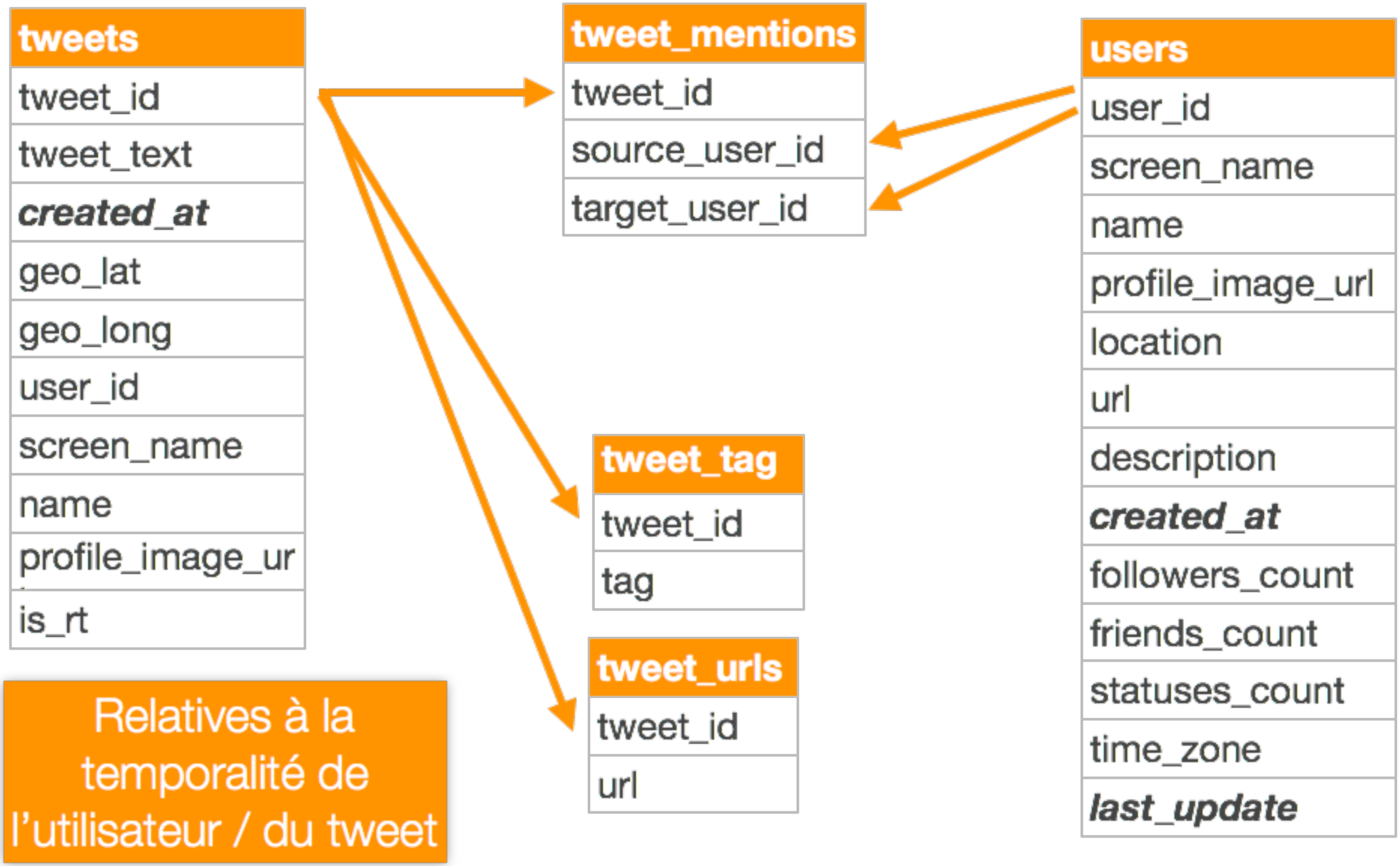
- Artifacts of times within tweets

‘jadis’/‘yore’
≠ ‘Il y a cent ans’/‘on this day in history’
≠ 11/11/1918

- Are those different ways to express ‘time’ reflecting different kinds of perceptions of time/different kinds of memories?

- Need to use NLP/Named Entities Recognition

The poverty of time-based metadata



tweets
tweet_id
tweet_text
created_at
geo_lat
geo_long
user_id
screen_name
name
profile_image_ur
is_rt

tweet_mentions
tweet_id
source_user_id
target_user_id

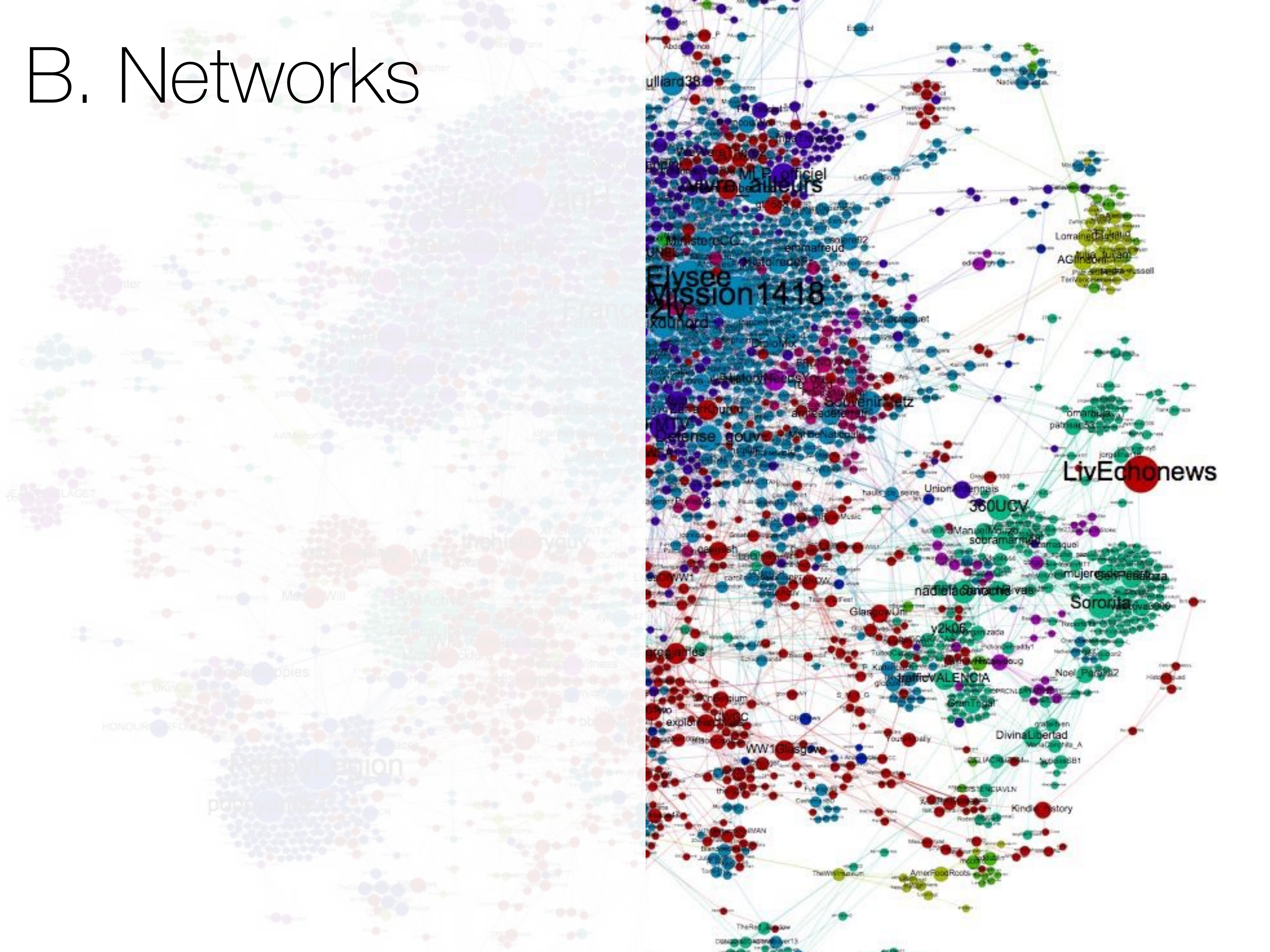
tweet_tag
tweet_id
tag

tweet_urls
tweet_id
url

users
user_id
screen_name
name
profile_image_url
location
url
description
created_at
followers_count
friends_count
statuses_count
time_zone
last_update

Relatives à la temporalité de l'utilisateur / du tweet

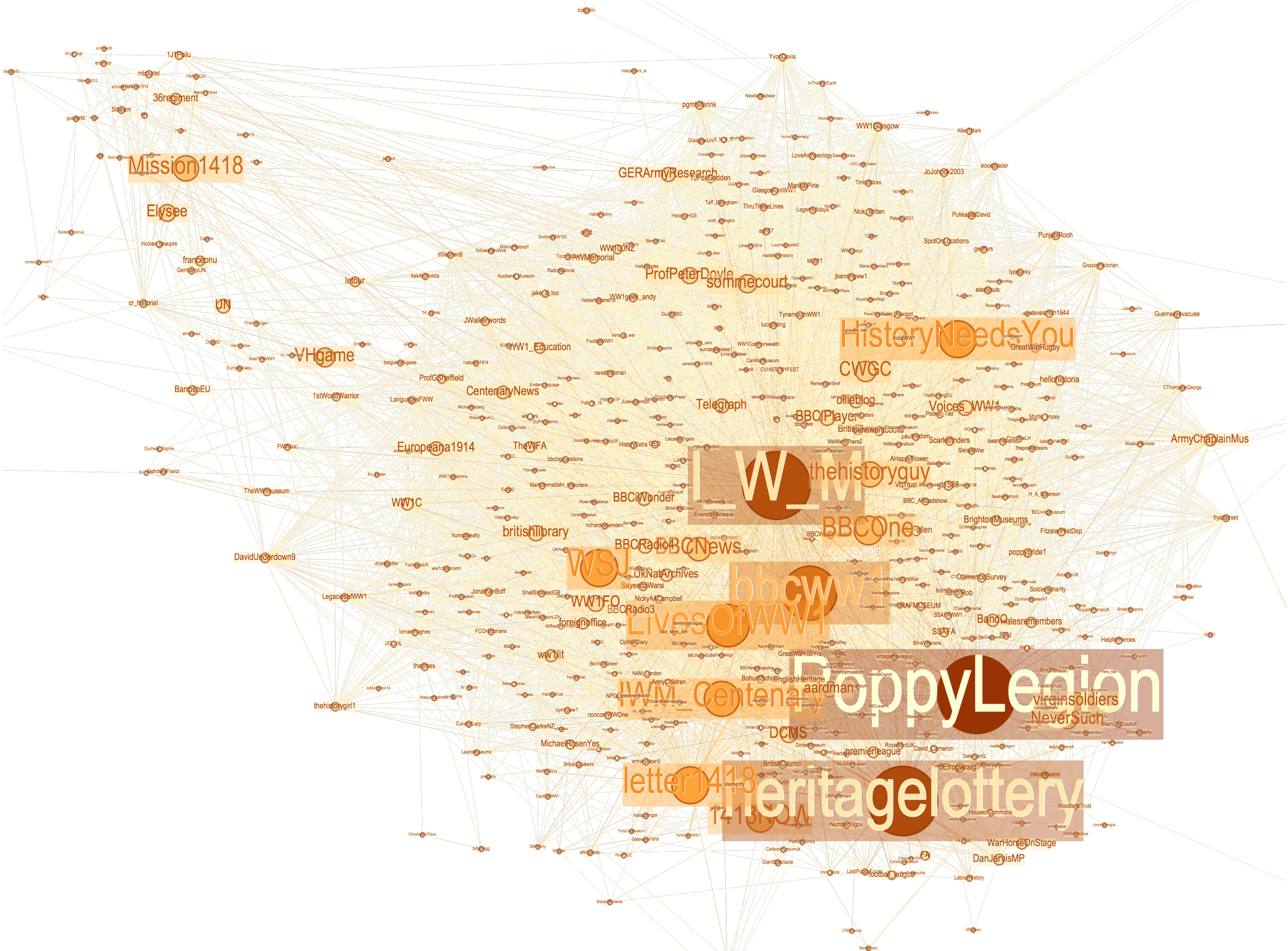
B. Networks


MLP_officiel
vivre_ailleurs
emmafreud
Elysee
Mission1418
Defense_gouv
Souvenirmetz
LivEchonews
360UCV
Sororita
mujeres
GlasgowUni
y2k06
TrafficVALENCIA
DivinaLibertad
Noel_Perez82
WW1Glasgow
Kindle history
AmerFoodRoots
thehistoryguy

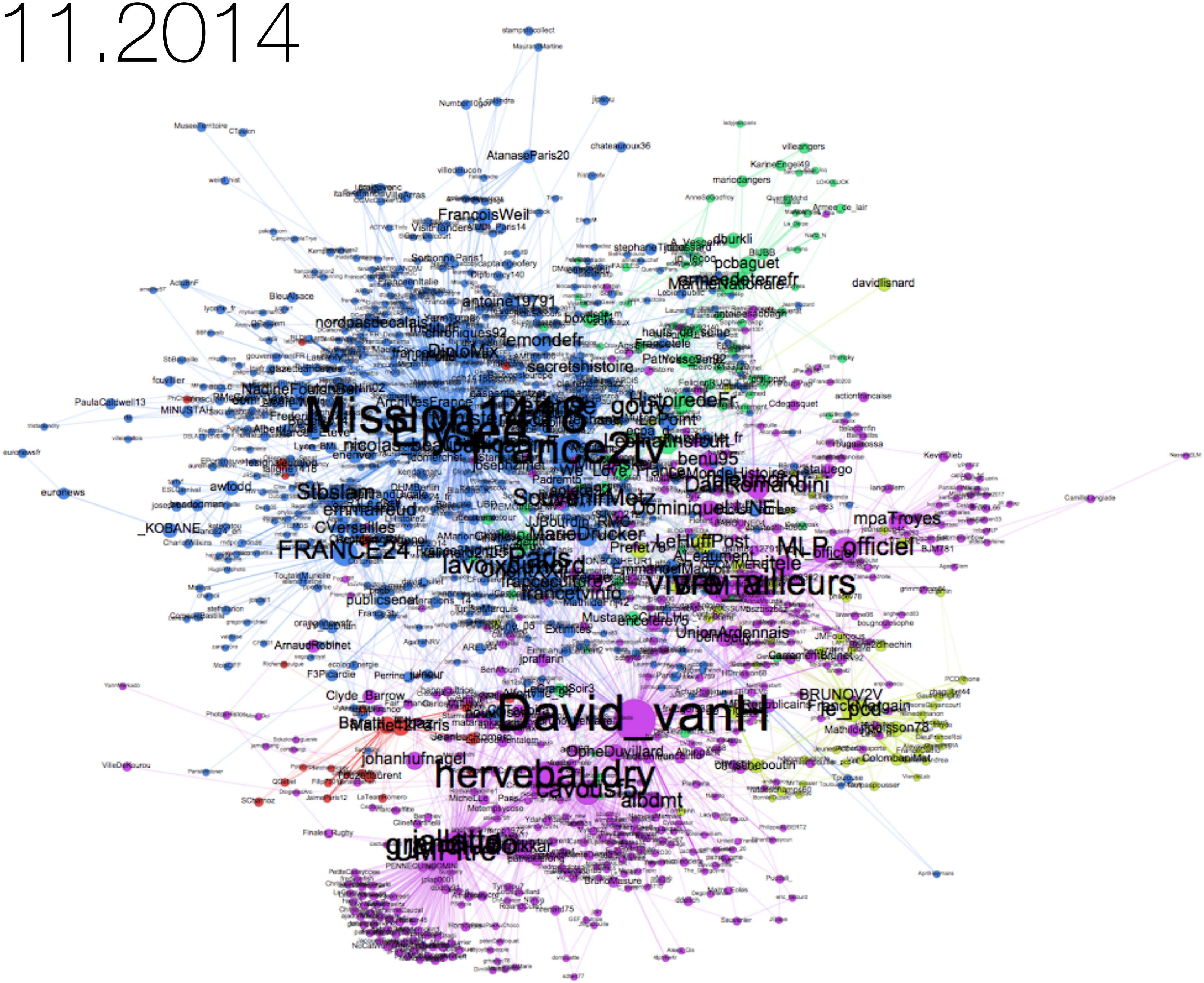
The (already seen a million time) network dataviz

- Networks dataviz are based on
  - aesthetics
  - Strong sociological theory
- Difficult not to be YAND
  - ‘yet another network dataviz’

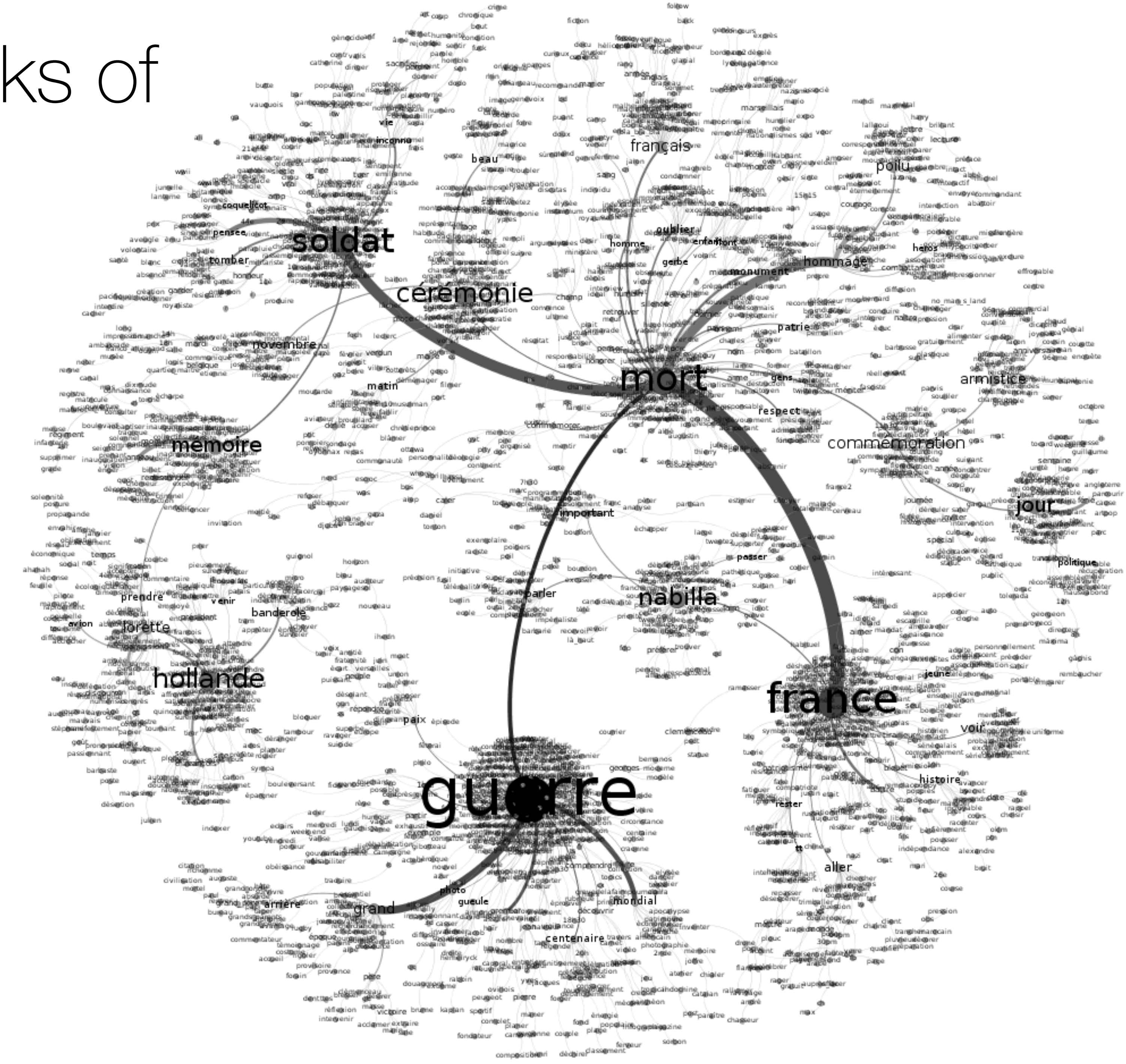
Frenchs, Englishs and others



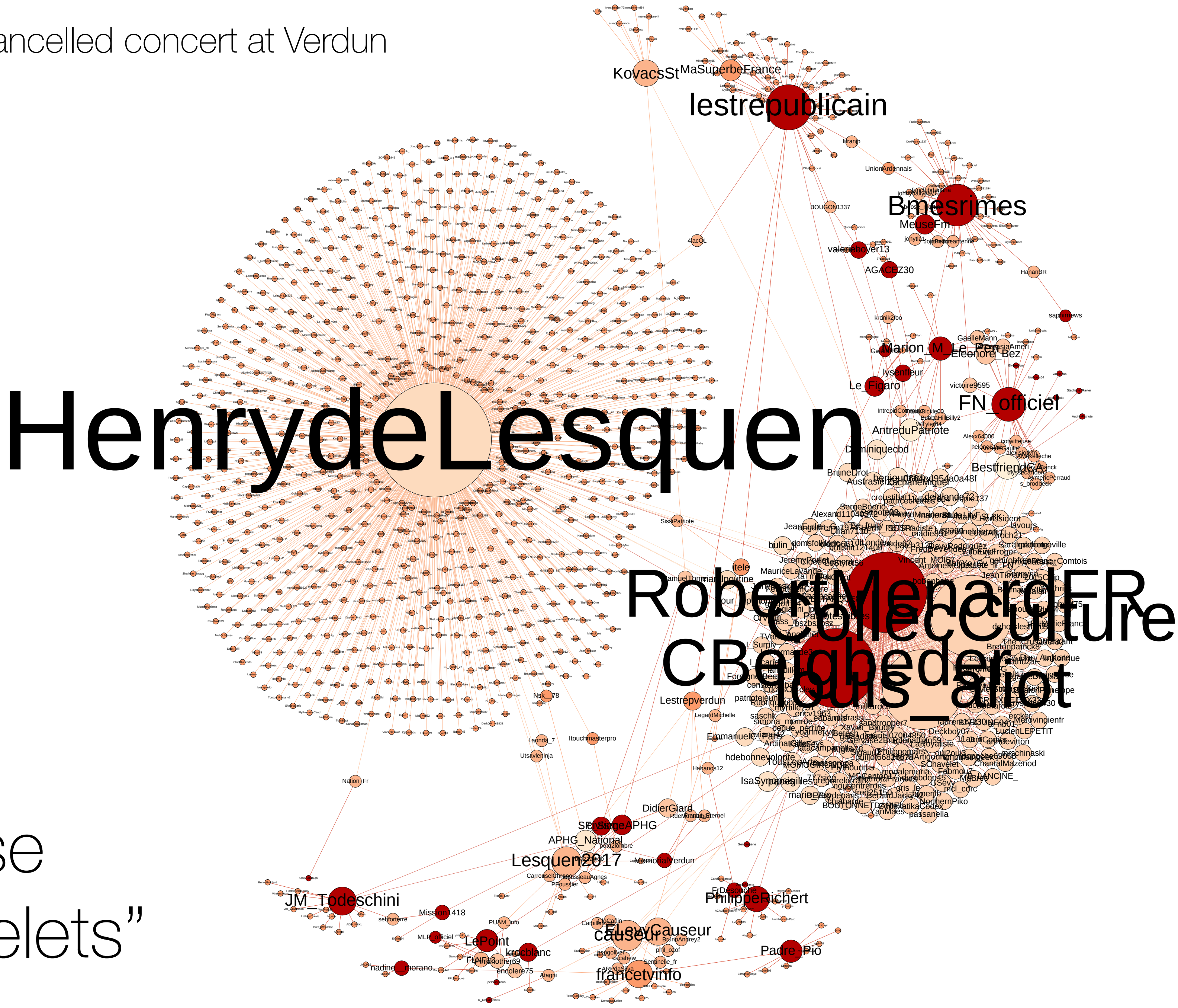
11.11.2014


stampsfocollect
MauraisMartine
jipsou
AtanaseParis20
chateauroux36
villedoucon
MuseeTerritoire
FrancoisWeil
villeangers
KarineEngel49
mariodangers
Armee_de_lair
dburkli
stephaneTjdossard
pcbaguet
MarineNationale
davidlisnard
antoine19791
nordpasdecalais
BleuAlsace
lemondefr
DiploMix
secretshistoire
Francetele
PaulaCaldwell13
MINUSTAH
fcuvilier
NadineFouchoniari
ArchivesFrance
Mission1418
France2tv
HistoiredeFr
gouy
LePoint
actionfrancaise
Cdegasquet
benu95
euronews
awtodd
Stbslam
emmarroud
DanRomandini
MondeHistoire
KevinDieb
mpaTroyes
DominiqueLUNEL
SouvenirMetz
_KOBANE
CVersailles
FRANCE24
MarieDrucker
JJBourdin_RMC
LeHuffPost
MLP_officiel
Prefet78
lavoixdunord
francetvinfo
EmmanuelMacron
publicsenat
ArnaudRobinet
UnionArdennais
F3Picardie
jpraffarin
BRUNOV2V
Clyde_Barrow
David_vanH
johanhufnagel
OpheDuvillard
hervebaudry
cavous15
albdmt
christineboutin
VilleDeKourou
SChamoz
Finales_Rugby
BrunoMasure
AprilHumans
euronewsfr

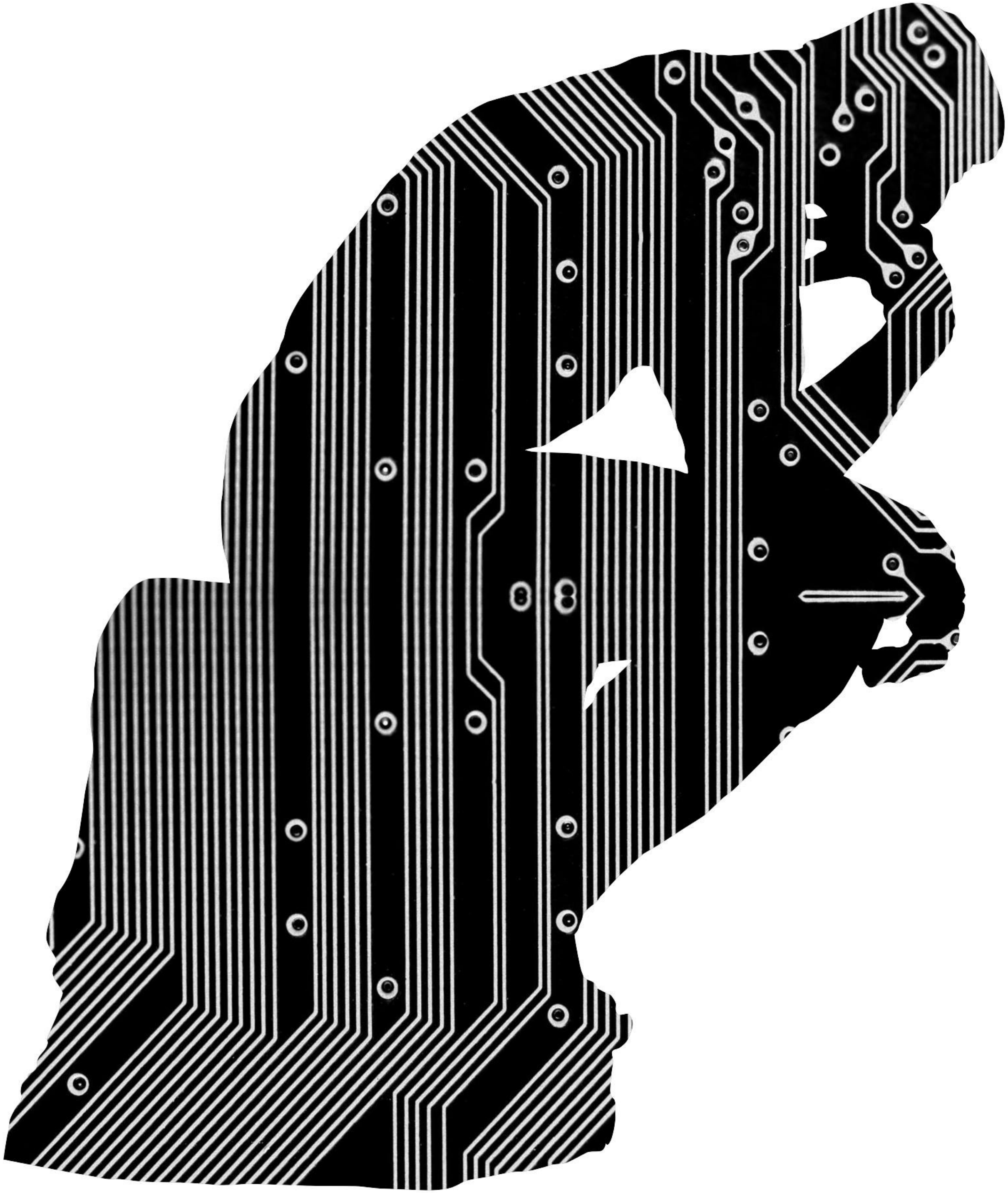
Networks of words



Black M’s cancelled concert at Verdun



Intense “Wavelets”



CONCLUSION

QUESTIONNING MY RESEARCH

Neither computing nor statistics but…

- 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?