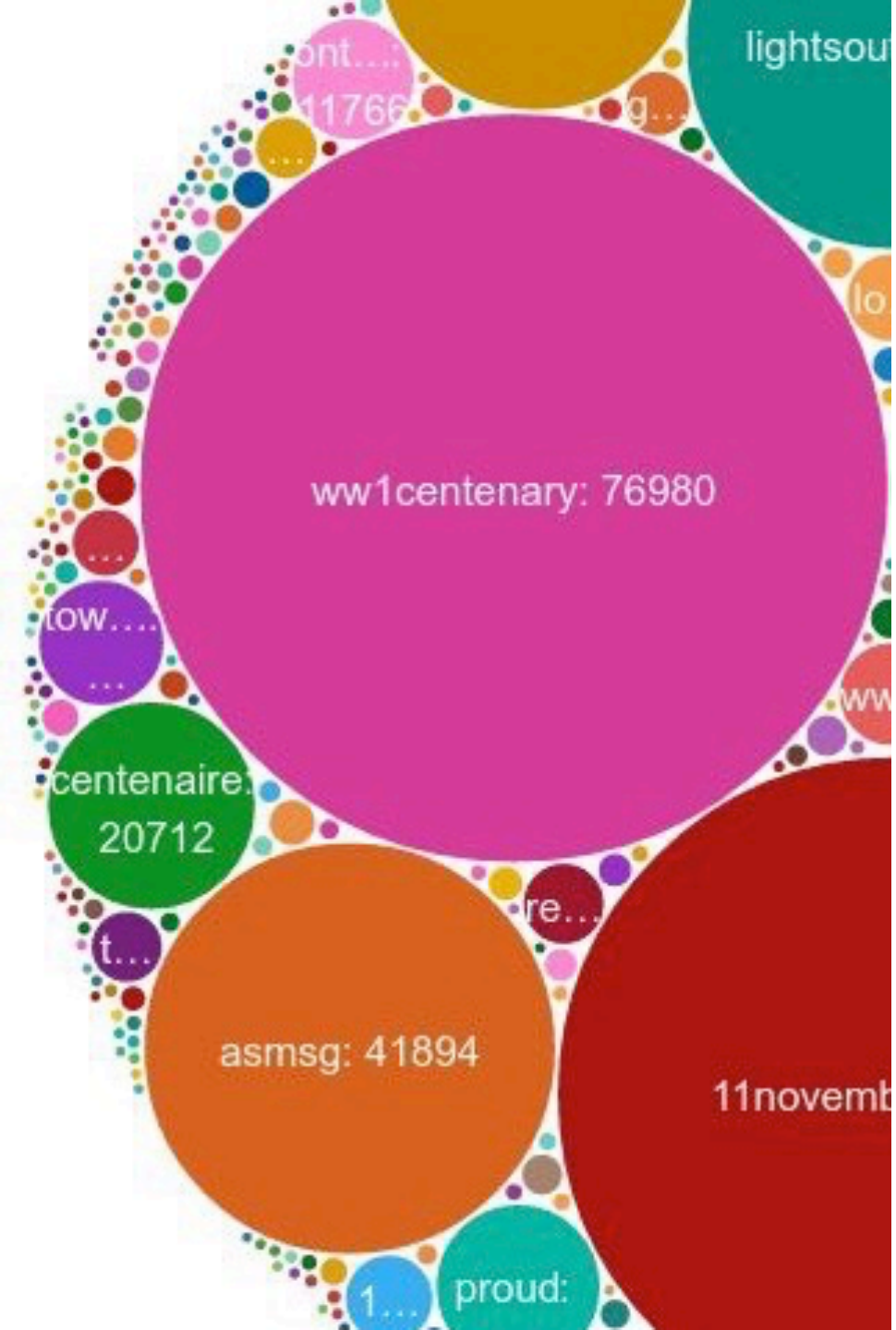
Sources en flux

Collecter, analyser, archiver, pérenniser

Enseignements d'une collecte de tweets sur le Centenaire de la Grande Guerre.

Frédéric Clavert (frederic.clavert@unil.ch) - 9 mai 2016



#ww1

AUX SOURCES D'UNE RECHERCHE

Apprentissage d'un savoir-faire

- Découverte de Twitter lors de conférences *Digital Humanities*
  - *2009*
- Collecte des tweets liés aux conférences
  - 2012
- L'échec d'une première collecte massive
  - #ledebat - 2012
- L'apprentissage de l'API publique de streaming
  - #manifpourtous

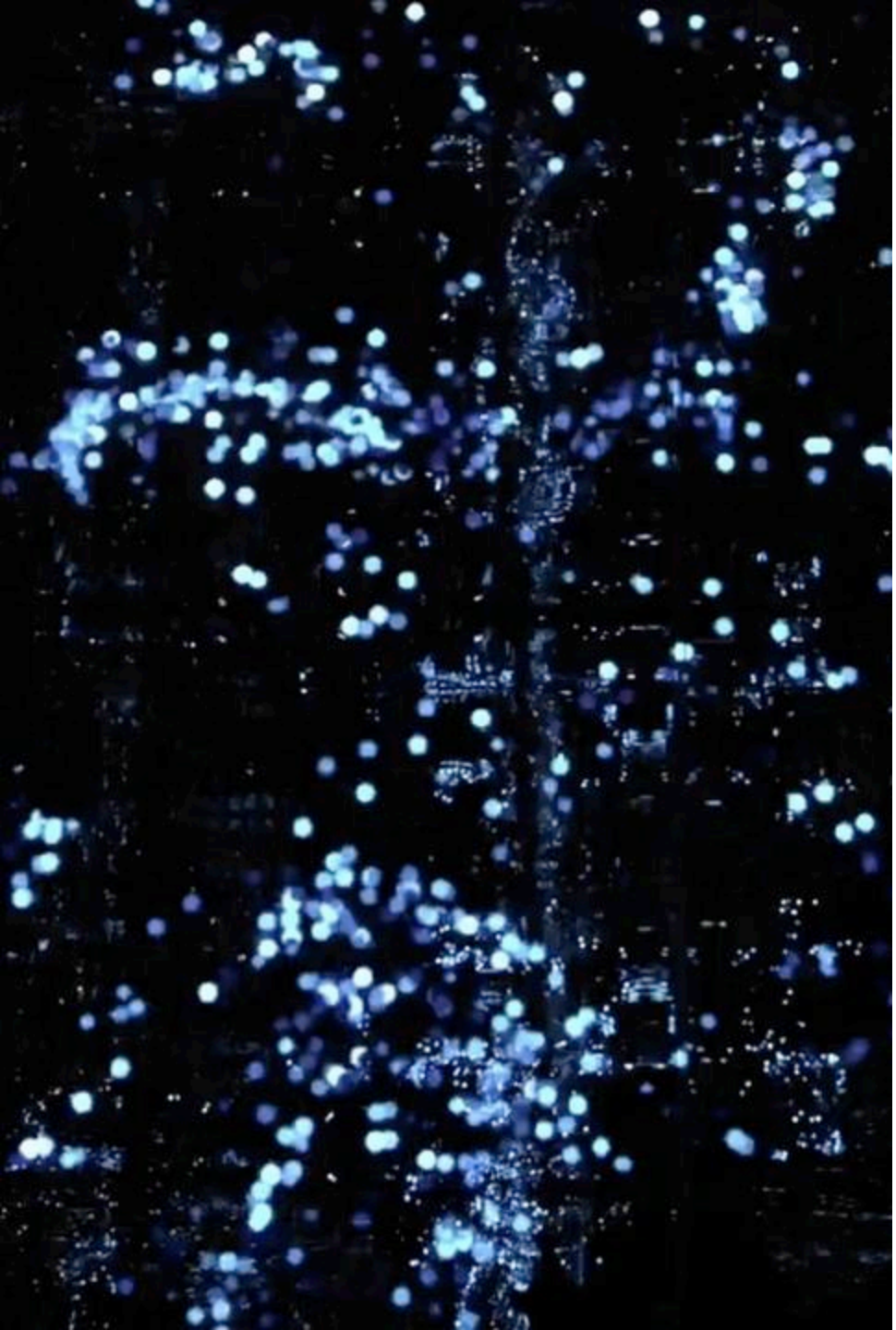
#ww1

- Application de ces nouveaux savoir-faire au Centenaire de la Grande Guerre
  - Lancé en France le 11 novembre 2013
  - Collecte à partir d'avril seulement
- Intérêts scientifiques
  - Analyser la relation histoire-mémoire sur Twitter
  - La perception du passé et sa mémoire change-t-elle avec les réseaux-sociaux?

L'historien.ne et les données massives

- Questions méthodologiques fondamentales
  - Le corpus est le fruit d'une recherche
  - L'ordre illusoire
  - Pourquoi Twitter?
  - Relations chercheur / bibliothécaires et archivistes / informaticiens / …
  - La question du « bricolage »
- Questions pratiques
  - Pourquoi Twitter?
  - Choix des outils de collecte et d'analyse
  - Archivage et pérennisation des données

Comment fonder une recherche sur des sources primaires en flux?

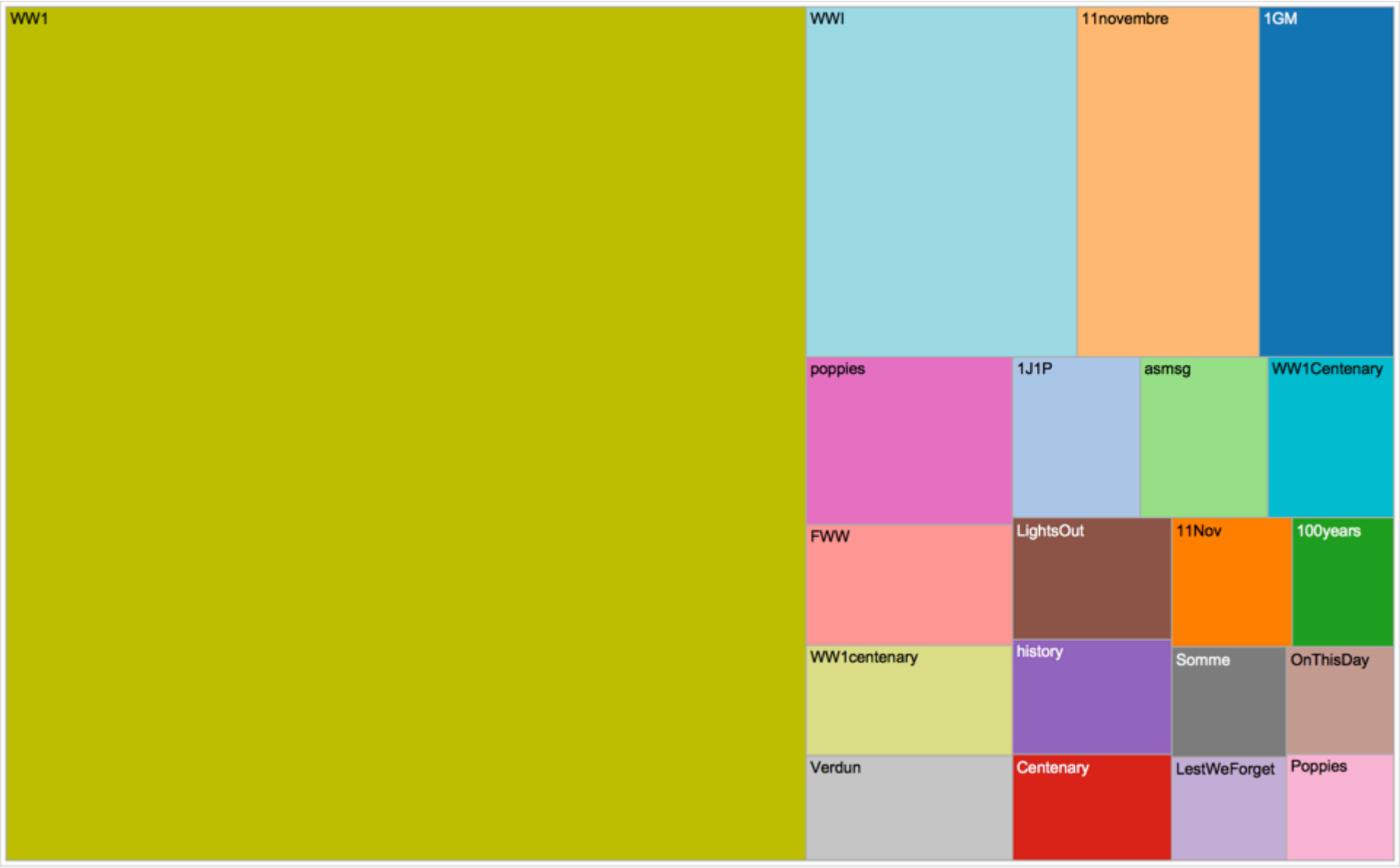


COLLECTER

Présentation du corpus #ww1

- Collecte
  - Commencée le 1er avril 2014, toujours en cours
  - Multilingue
  - Artisanal… puis plus professionalisée
- Script serveur: 140dev
  - Collecte en XML / Stocke en SQL
  - Collecte par mots clés

Les mots-clés collectés



Descriptif du corpus (13 avril 2016)

- 2 096 968 de tweets
  - Peu de bruit, sauf pour #11Nov
  - Bruit en augmentation (Somme, Verdun)
  - 730 111 sans les retweets
- Émis par 542 570 comptes
  - De toutes sortes
  - individus, institutions mémorielles, médias, projets de recherche… voire bots
- 124 424 hashtags
  - 54 566 utilisés une seule fois
  - 107 047 dix fois ou moins

Problèmes méthodologiques liés à la collecte

Une notion fondamentale: l'API

- *Application Programming Interface*
  - Échanger fonctionnalités ou données
- Fondamental?
  - En fonction de l'API, les données récoltées ne sont pas les mêmes
- Twitter
  - API search / API streaming
  - Public API / Full API
  - Le 1%
  - Le flou de Twitter sur ses API: intégralité ou non?

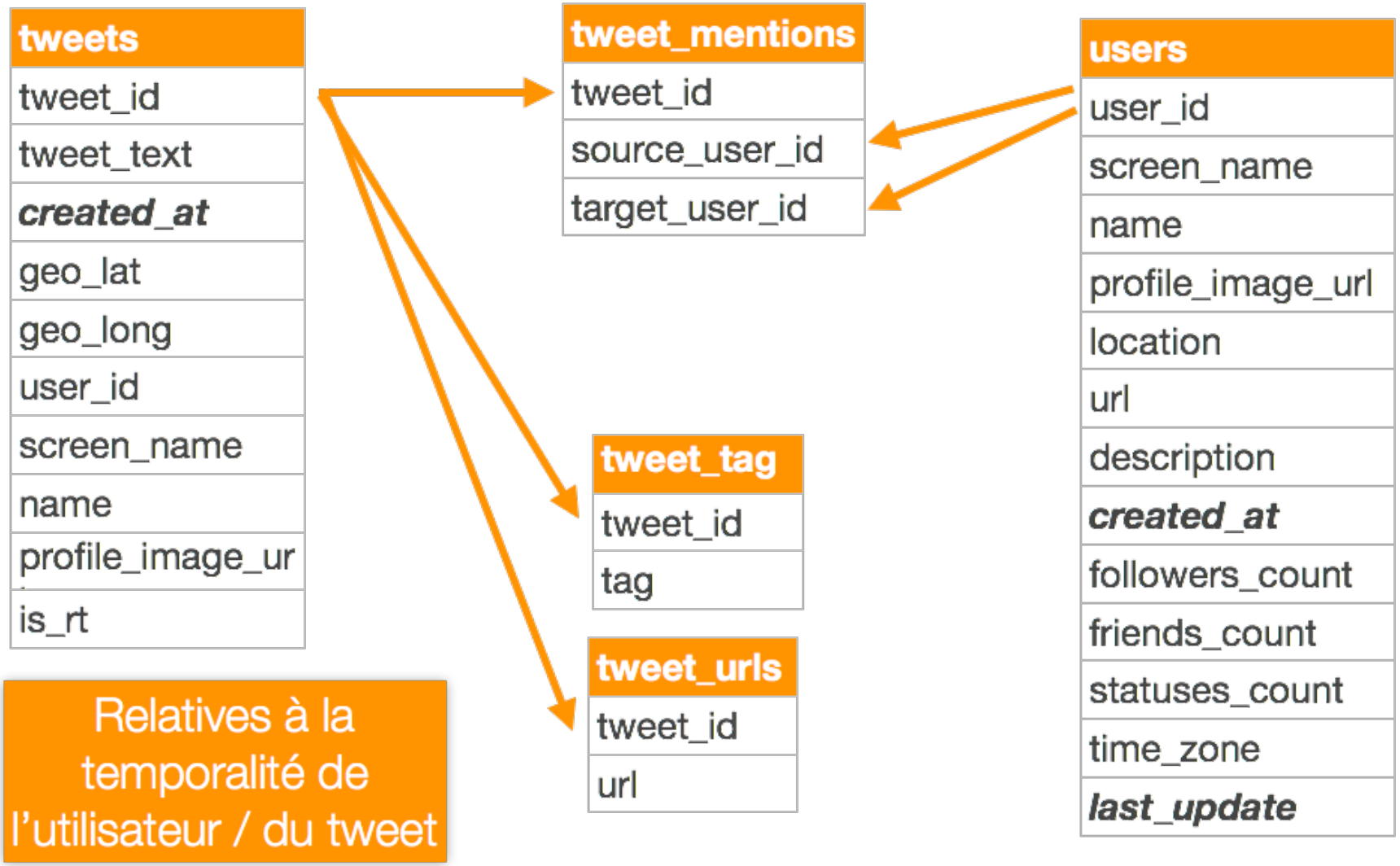
Les affres du Big Data

- Notion relative
  - Point de vue de l'historien.ne ≠ point de vue de Gnip
- L'ordre illusoire
  - «intégralité»?
- Le tropisme anglo-saxon

Les affres du flux

- L'API *search* de Twitter ne peut être utilisé dans ce cas
  - Permet de chercher dans l'historique des tweets quelques jours en arrière, 3000 tweets / heure
- L'API *streaming* doit l'être
  - Ne permet pas de chercher dans l'historique
  - Un oubli de hashtag / mot-clé ne peut être rattrapé (sauf par la voie commerciale)
- L'anticipation est reine
  - Elle n'est pas toujours possible

Des données peu structurées



Historien face
à une mer de données

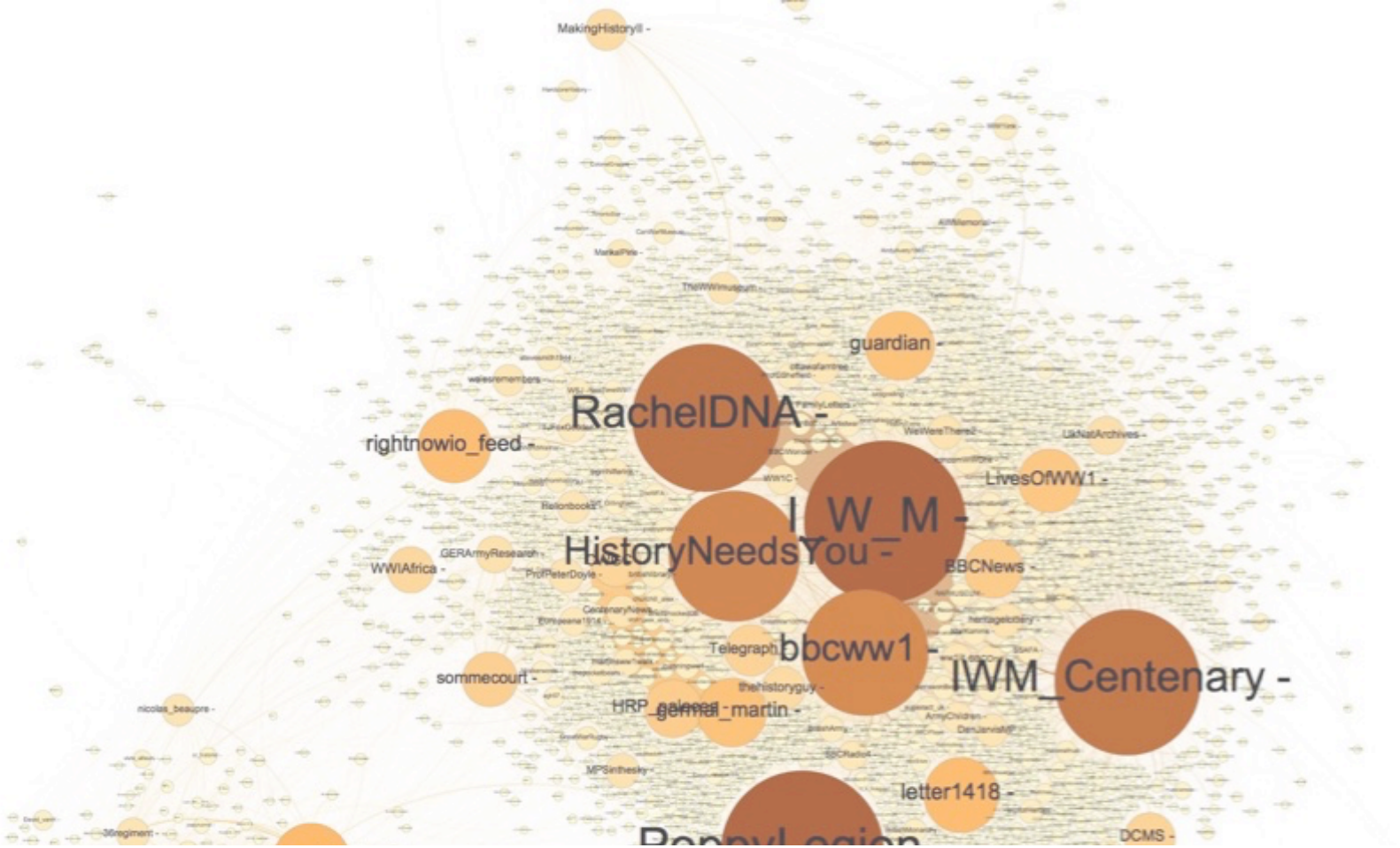
ANALYSER

Comment lire deux millions de tweets?

Fischer, Frank; Göbel, Mathias; Kampkaspar, Dario; Kittel, Christopher; Trilcke, Peer (2015): Drama Networks Superposter. https://dx.doi.org/10.6084/m9.figshare.1461761.v1

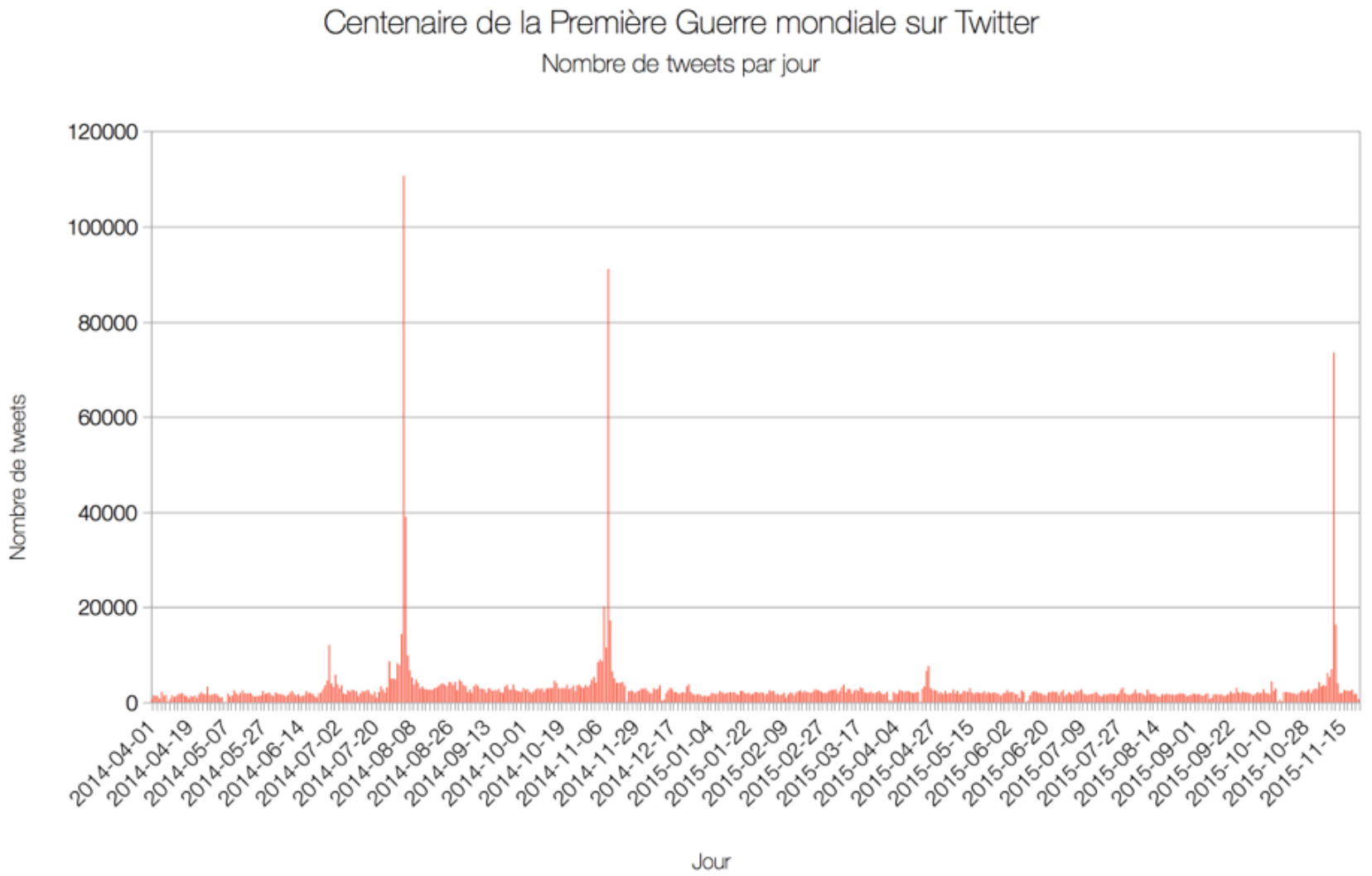
Distant Reading

- Notion de lecture distante
- Introduite par Franco Moretti dans *Graphs, Maps and Trees* (Verso, 2007)
  - Histoire des grands romans ou histoire de la littérature?
- Articulation entre lecture distante et lecture proche (lecture critique classique de l'historien)
  - À chaque étape, garder un lien avec le tweet singulier
- Fouille de texte (text mining), analyse réseaux…

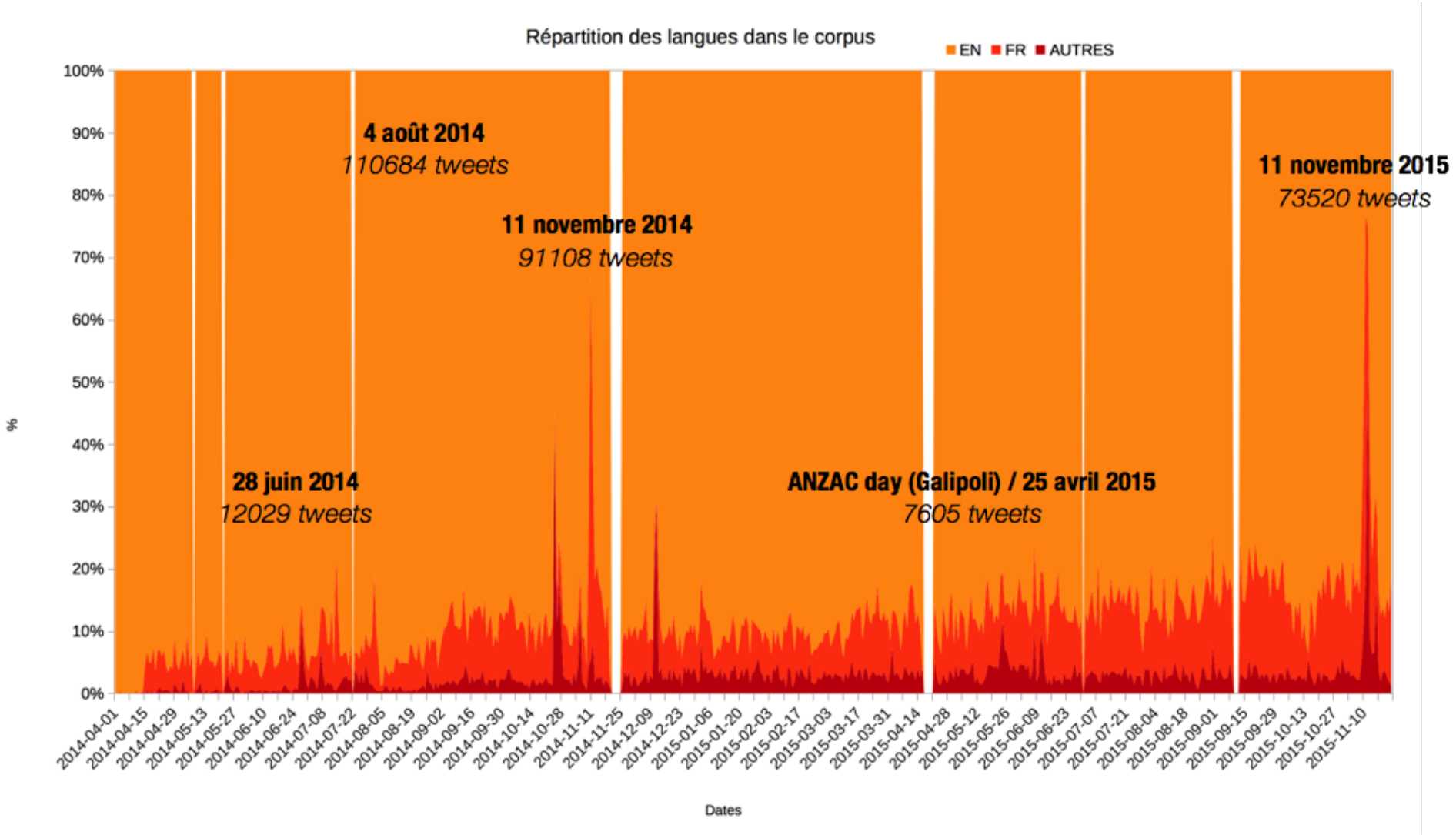


Quelques résultats

Le flux

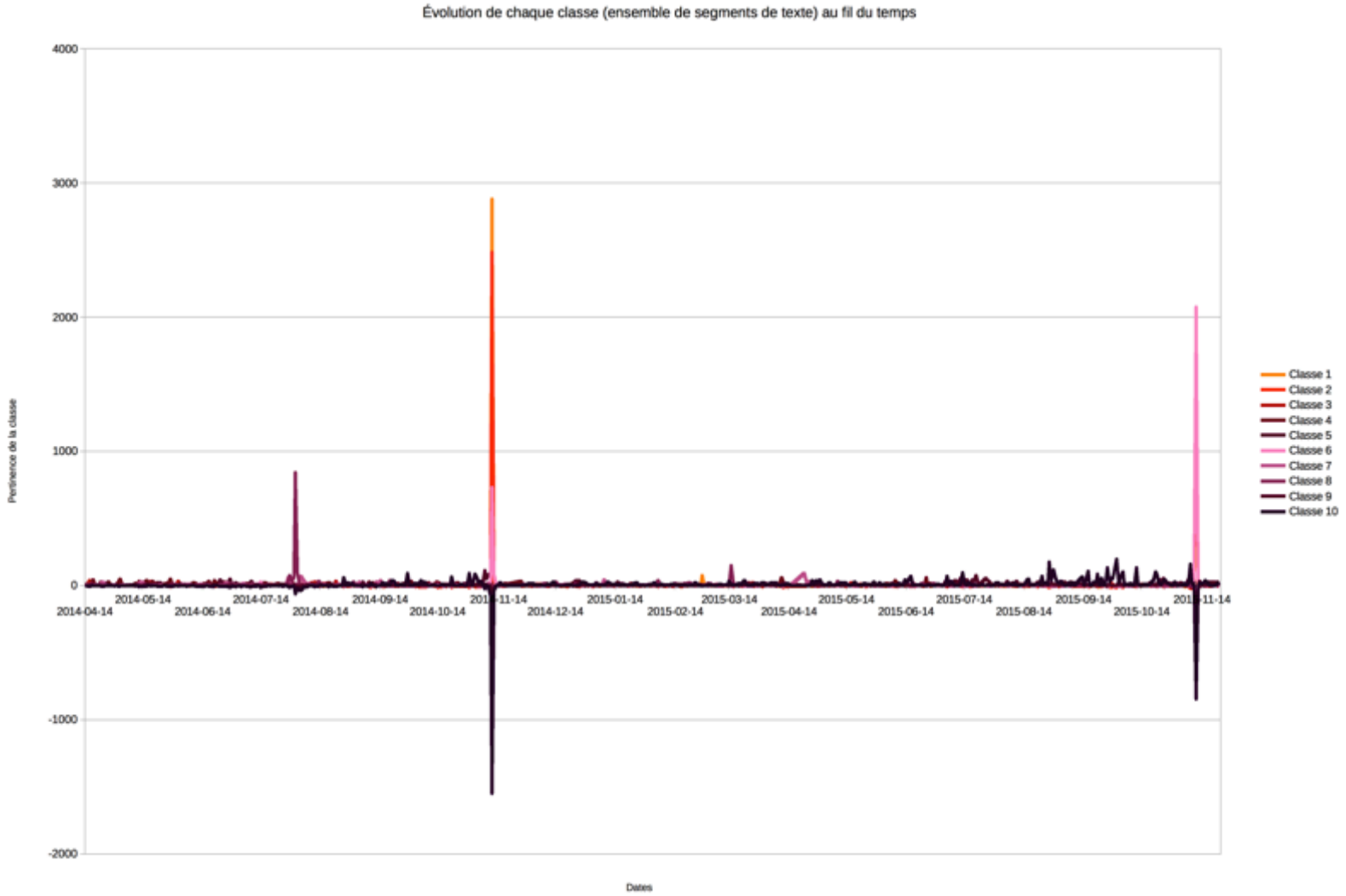

Centenaire de la Première Guerre mondiale sur Twitter
Nombre de tweets par jour
Nombre de tweets
120000
100000
80000
60000
40000
20000
0
2014-04-01
2014-04-19
2014-05-07
2014-05-27
2014-06-14
2014-07-02
2014-07-20
2014-08-08
2014-08-26
2014-09-13
2014-10-01
2014-10-19
2014-11-06
2014-11-29
2014-12-17
2015-01-04
2015-01-22
2015-02-09
2015-02-27
2015-03-17
2015-04-04
2015-04-27
2015-05-15
2015-06-02
2015-06-20
2015-07-09
2015-07-27
2015-08-14
2015-09-01
2015-09-22
2015-10-10
2015-10-28
2015-11-15
Jour

Les temporalités linguistiques

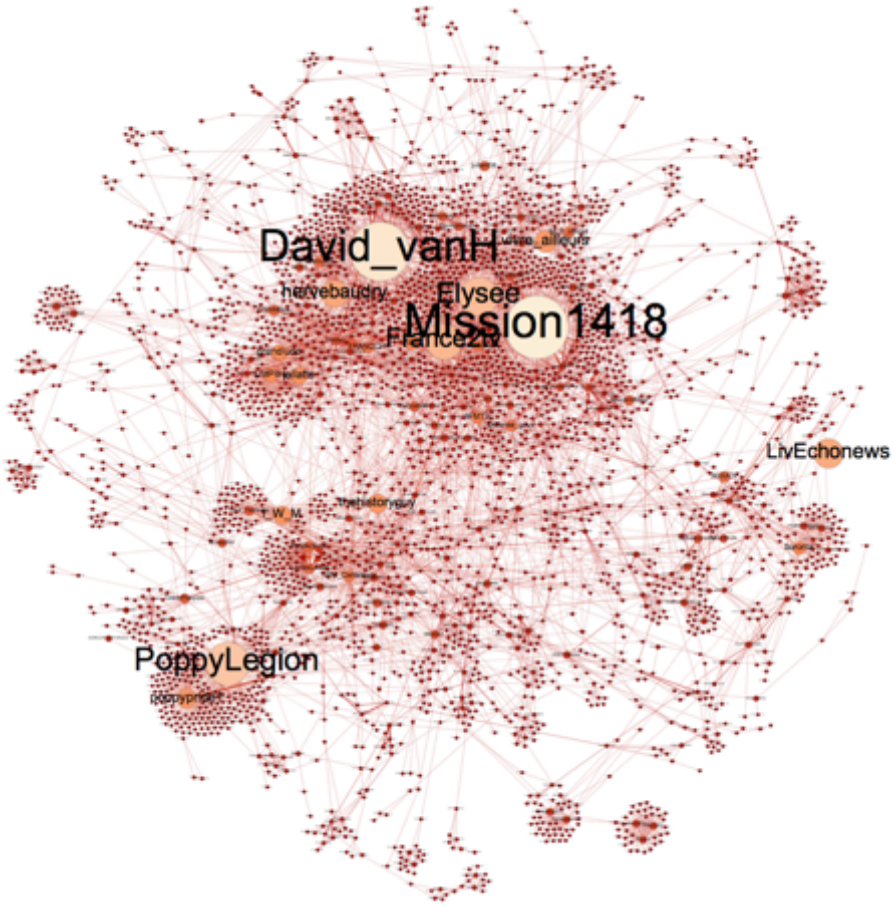


Fouille de texte (corpus francophone)…

…projetée dans le temps



La visualisation réseau (11/11/ 14)



Limites de cette approche

- Les limites techniques
  - Données peu structurées
  - La capacité matérielle (processeur, mémoire vive…)
  - Donc l'infrastructure technique
- Les limites de Twitter
- Les limites de l'historien.ne
  - Savoir technique à acquérir
  - Savoir méthodologique

Les limites techniques (1)

- Données peu structurées: exemple des temporalités
- Une seule date réellement intéressante
  - la date de publication des tweets
- Ne peut rendre compte de la richesse des temporalités entremêlées de ce corpus
  - Le flux: Twitter
  - Le Centenaire
  - La Grande Guerre
  - Chaque utilisateur
- Absence qui doit être compensée
  - Ex: reconnaissance d'entités nommées

Les limites techniques (2)

- Capacité matérielle et infrastructure
- Projet type «garage band» / artisanat
  - Serveur de récupération
  - Connexion ADSL privée
  - Interruptions dans la collecte
- Migration vers une infrastructure universitaire
  - Professionnalisation du projet
  - Inégalité entre les pays

Les limites techniques (3)

- Les logiciels utilisés
  - Gephi
  - Iramuteq
  - LibreOffice
- Usages intenses de la mémoire vive
  - Pas le cas des outils Big Data (ruby, python)
  - Limite leur usage sur un corpus conséquent
- Il faut aussi une infrastructure pour l'analyse
  - Renchérit le coût des SHS

Les limites de Twitter

- Twitter Inc. peu disert sur ce qui est effectivement collecté (API publique de streaming)
- Interruptions dans la collecte
  - Pannes de l'API
  - Changements techniques de l'API
- Épée de Damoclès
  - Société commerciale déficitaire en grande difficulté
  - Si rachat, l'API restera-t-elle aussi disponible?
  - Risque: Algopol et Facebook
- Ce qui est sur Twitter reste-t-il sur Twitter?

Les limites de l'historien.ne

- Les savoir-faire techniques à apprendre sont conséquents
  - Ne peuvent tous être acquis
  - Nécessité d'un travail d'équipe
  - Nécessité d'une culture numérique
- Les savoir-faire méthodologiques le sont encore plus
  - Choisir un logiciel = choisir une méthodologie
  - Fouille de texte: savoir statistique très précis
  - Sociologie de l'analyse des réseaux «sociaux» / sociologie des controverses
- Pose la question de la formation des chercheurs / du travail en équipe / du travail par projets
  - Remis en cause comme correspondant aux évolutions «néolibérales» de l'université
  - Cf. LARB et *Variations*

Que faire des données?

ARCHIVER ET PÉRENNISER

La difficile question de l'archivage

- Plusieurs niveaux d'archivage
  - L'archivage du web
  - Les réseaux sociaux numériques dans leur ensemble
  - Les données liées à une recherche précise
- L'archivage du web ne comprends pas (par défaut) l'archivage des RSN
  - Y compris Internet Archive
  - En France: BNF et INA (dépôt légal), .fr plus sélection
    - Dans certains cas, archivages de tweets.

Le cas Suisse

- Assuré par la bibliothèque suisse, en coopération avec bibliothèques cantonales notamment
- Politique sélective
  - « sites web patrimoniaux qui ont un fort lien avec la Suisse et qui sont accessibles librement »
  - sites web sur les cantons et les communes, domaines spécifiques tels que sciences sociales ou littérature suisse
  - Collections spéciales
  - Pas de collecte systématique du .ch
- Pas de collecte des RSN

Les réseaux sociaux numériques

- Twitter semble être le seul RSN archivé par une institution publique
  - Library of Congress
  - Depuis 2006
  - Archive non consultable: difficultés techniques trop importantes pour le moment
- Les autres RSN, dont Facebook
  - Politique d'archivage dépend de chaque RSN
  - Spectre de la disparition…
    - GeoCities. MySpace?
  - Inégalité des chercheurs face à l'accès aux données

L’archivage des données des chercheurs

- Question essentielle
  - Plusieurs dimensions: archivage des données / des logiciels pour les exploiter
  - Années 1990: nombreuses bases de données perdues
- Infrastructures nationales
  - TGIR Huma-Num en France
  - Colloque récent sur les données de la recherche à Lausanne
- MAIS…
  - Conditions d’utilisation restrictives de Twitter
  - Consultation ne peut être que limitée

Quelques questionnements

- Problème de l'archivage des utilisateurs
  - Cf. Louise Merzeau sur les questions d'autorité sur Twitter
  - Un tweet n'a pas la même autorité selon son émetteur (nombre de followers / following comme indice d'autorité)
  - Twitter fournit des statistiques pour chaque tweets. Quel archivage?
    - Important, car ces statistiques intègrent aussi les «impressions» donc le comportement des comptes qui ne publient pas mais lisent
- Problème de l'archivage de l'interface
  - #ww1: par d'archivage de l'interface, problème méthodologique
  - En outre, interfaces multiples (Twitter, Tweetdeck, echofon, interfaces mobiles…)

Archiver un éco-système?

- Avec ses API plutôt ouvertes, Twitter a créé tout un éco-système
  - Ex: mesure de l'autorité / popularité / audience avec klout.com
  - Services d'images, les raccourcisseurs d'URLs…
- Twitter encourage le partage des liens
  - Archivage de ces liens?
- Archivage des tweets / comptes / trends sponsorisés?
  - Des trends non sponsorisés
- Comment archiver des pratiques et usages qui évoluent constamment?

Comment archiver tout un éco-système?

Conclusion

Les sources en flux

- Twitter est un flux
  - Caractéristiques du flux: continu et en temps réel
- L'utiliser pour ses recherches puis l'archiver, c'est vouloir capter et figer ce flux
- Opposition flux / archives
  - nombreuses tensions méthodologiques
  - difficultés (techniques et méthodologiques) soulevées par l'archivage des tweets