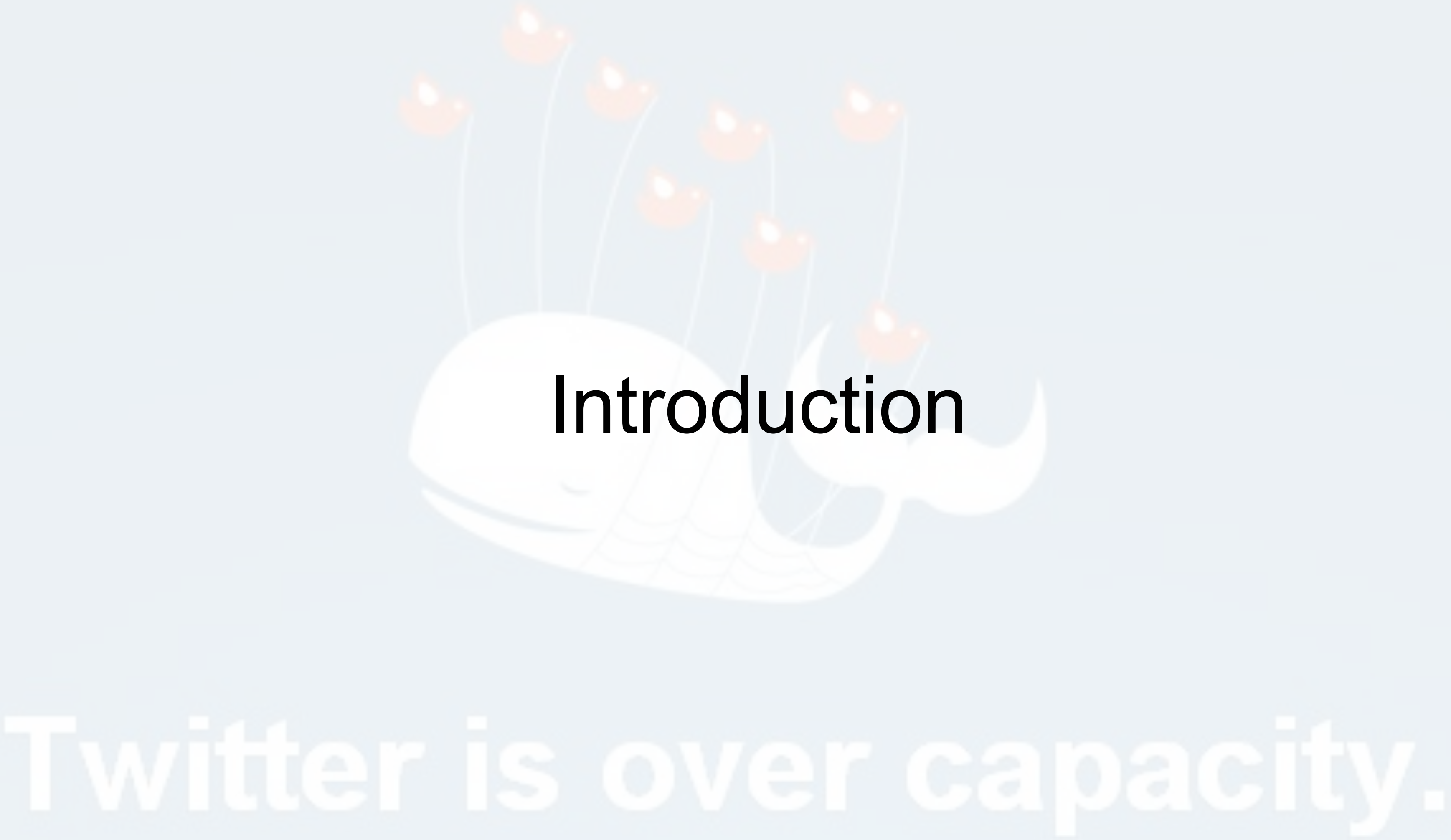
Twitter data as primary sources for historians: a critical approach

Lessons from two projects: the 2015 Greek referendum and the Centenary of the Great War on Twitter

Sofia Papastamkou (European Center for the Humanities and Social Sciences, Lille)
Frédéric Clavert (C2DH, University of Luxembourg)

Introduction



Introduction: social networks online, a definition

» We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site. «

(Boyd Danah M et Ellison Nicole B., « Social Network Sites: Definition, History, and Scholarship », *Journal of Computer-Mediated Communication* 13 (1), 01.10.2007, p. 210-230.)

Introduction: Twitter studies

- Since the beginning of Twitter, its data has been used in Humanities and Social Sciences for different purposes

  See: Williams Shirley A., Terras Melissa M. et Warwick Claire, « What do people study when they study Twitter? Classifying Twitter related academic papers », Journal of Documentation 69 (3), 10.05.2013, pp. 384‑410. En ligne: <https://doi.org/10.1108/JD-03-2012-0027>, consulté le 24.10.2018.

- Two kinds of historian’s work
  - Collective memory

    Turgeon Alexandre, « Comment Travailler La Mémoire Sur Twitter. Quelques réflexions d’ordreméthodologique à partir de la Grande Noirceur et Révolution Tranquille 2.0 », Études canadiennes / Canadian Studies. Revue interdisciplinaire des études canadiennes en France (76), 01.07.2014, pp. 11‑26.
  - Current time events

    Ruest Nick et Milligan Ian, « An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter », The Code4Lib Journal (32), 25.04.2016. En ligne: <http://journal.code4lib.org/articles/11358>, consulté le 24.10.2018.

    Documenting the Now: https://www.docnow.io/

Introduction: why Twitter?

Because we can:

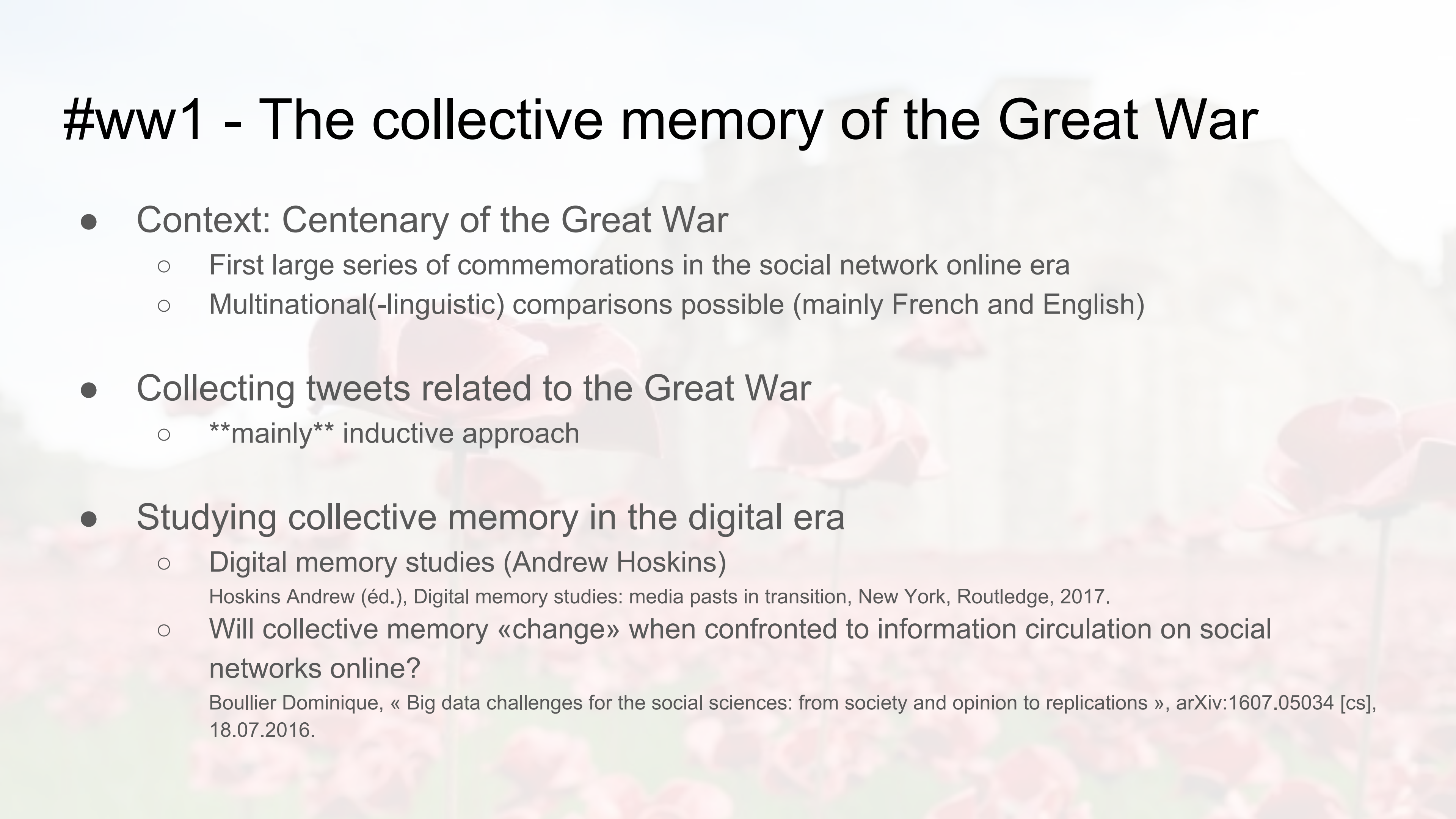
- Relatively (though less and less) open APIs
- Several free and one paying APIs
    - Search API (history)
    - Streaming API (what’s going on, < 1% of)
    - «If you pay you can get whatever you want» API (but we don’t pay, do we?)
- Lots of tools to collect tweets
    - twarc,
    - DMI-TCAT,
    - TAGS,
    - etc.

Introduction

Plan

I. Two projects, two theoretical backgrounds
II. Two projects, many methods and tools for the creation and the analysis of the corpus
III. Twitter hermeneutics

I. Two projects, two theoretical backgrounds



#ww1 - The collective memory of the Great War

- Context: Centenary of the Great War
  - First large series of commemorations in the social network online era
  - Multinational(-linguistic) comparisons possible (mainly French and English)

- Collecting tweets related to the Great War
  - **mainly** inductive approach

- Studying collective memory in the digital era
  - Digital memory studies (Andrew Hoskins)

    Hoskins Andrew (éd.), Digital memory studies: media pasts in transition, New York, Routledge, 2017.
  - Will collective memory «change» when confronted to information circulation on social networks online?

    Boullier Dominique, « Big data challenges for the social sciences: from society and opinion to replications », arXiv:1607.05034 [cs], 18.07.2016.

#greferendum - Studying the 2015 greek referendum

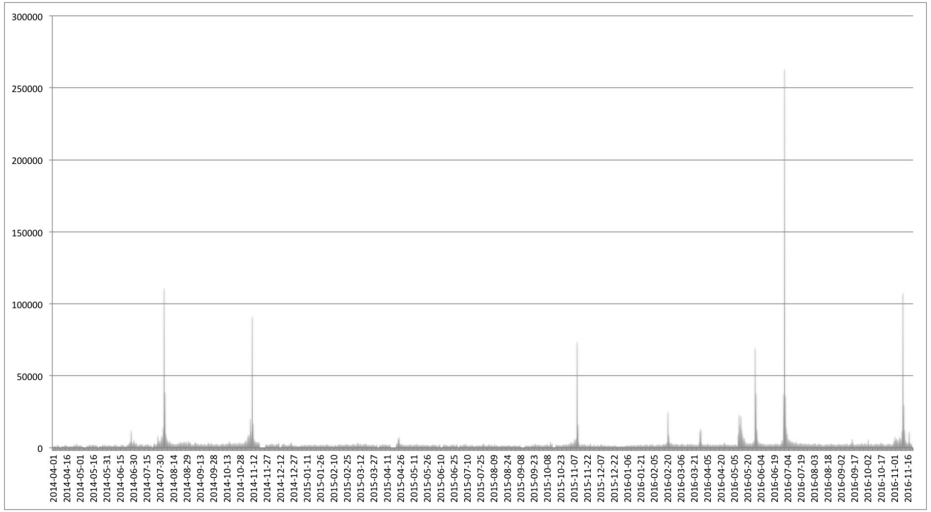
- Context: Greek debt crisis, Eurozone crisis,
  - A rich, born-digital (SNS), transnational documentation
  - A personal experimentation: archiving and analysing an event
- Collecting the #greferendum tweets
  - An *ad hoc* collect
  - Holistic approach (by hashtag)
- Studying the event
  - An important concept for historians (Seignobos 1898; Nora 1972; Le Goff 1999; Sirinelli 2002)
  - Twitter: the medium of the event
- Studying Twitter as a source for historians
  - non-institutional; decentralised; wild; born-digital

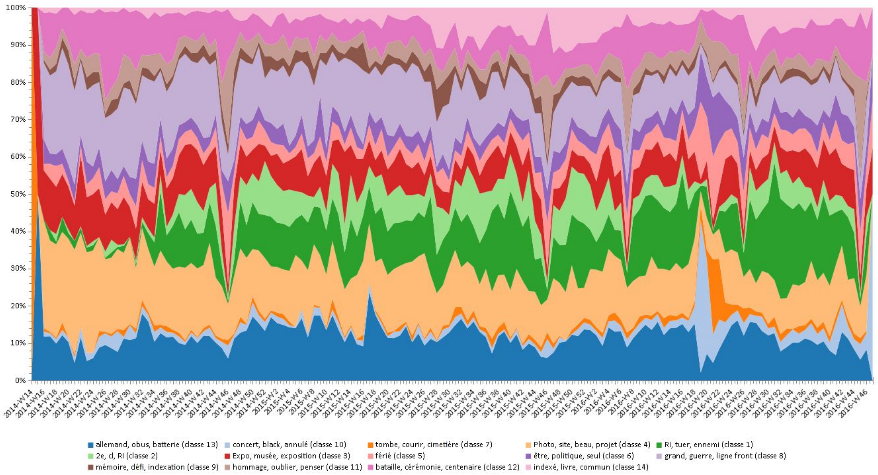
II. Two projects
Many methods and tools

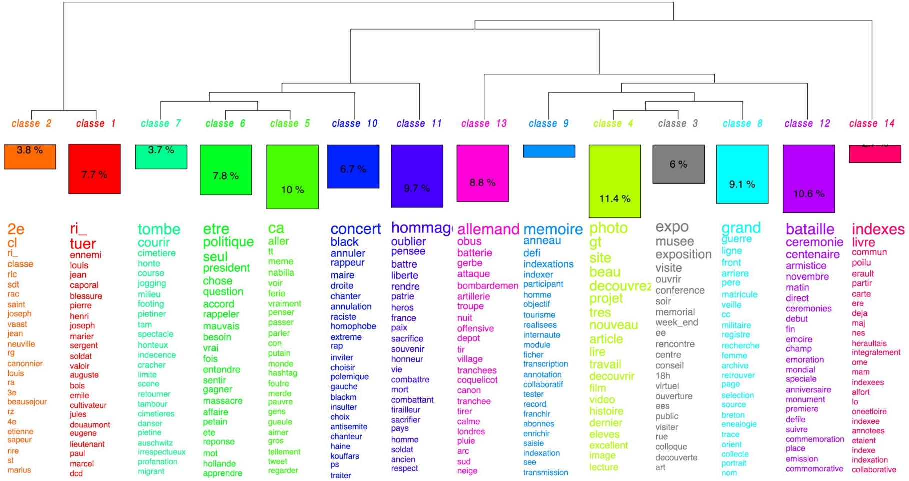
#ww1

- Collecting tweets
    - 140dev [abandonned] and [the incredible] DMI-TCAT

      Rieder Bernhard et Borra Erik, « Programmed method: developing a toolset for capturing and analyzing tweets », Aslib Journal of Information Management 66 (3), 19.05.2014, p. 262‑278.
    - a regularly updated tool, that can manage the many and regular changes in Twitter API
    - 5 Millions+ tweets, 1 million users (and GPDR headache) stored in a mariaDB database
- Preparing Twitter data for analysis
    - spreadsheets, OpenRefine, Dataïku DSS, SQL query, etc.
    - «in-between» tools that we don’t always talk about (but we should)
- Analysing tweets
    - IRaMuTeQ (iramuteq.org) = data mining
    - Gephi = social networks analysis
    - Dataïku DSS / spreadsheets for simple stats


300000
250000
200000
150000
100000
50000
0
2014-04-01
2014-04-16
2014-05-01
2014-05-16
2014-05-31
2014-06-15
2014-06-30
2014-07-15
2014-07-30
2014-08-14
2014-08-29
2014-09-13
2014-09-28
2014-10-13
2014-10-28
2014-11-12
2014-11-27
2014-12-12
2014-12-27
2015-01-11
2015-01-26
2015-02-10
2015-02-25
2015-03-12
2015-03-27
2015-04-11
2015-04-26
2015-05-11
2015-05-26
2015-06-10
2015-06-25
2015-07-10
2015-07-25
2015-08-09
2015-08-24
2015-09-08
2015-09-23
2015-10-08
2015-10-23
2015-11-07
2015-11-22
2015-12-07
2015-12-22
2016-01-06
2016-01-21
2016-02-05
2016-02-20
2016-03-06
2016-03-21
2016-04-05
2016-04-20
2016-05-05
2016-05-20
2016-06-04
2016-06-19
2016-07-04
2016-07-19
2016-08-03
2016-08-18
2016-09-02
2016-09-17
2016-10-02
2016-10-17
2016-11-01
2016-11-16


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2014-W14
2014-W16
2014-W18
2014-W20
2014-W22
2014-W24
2014-W26
2014-W28
2014-W30
2014-W32
2014-W34
2014-W36
2014-W38
2014-W40
2014-W42
2014-W44
2014-W46
2014-W48
2014-W50
2014-W52
2015-W2
2015-W4
2015-W6
2015-W8
2015-W10
2015-W12
2015-W14
2015-W16
2015-W18
2015-W20
2015-W22
2015-W24
2015-W26
2015-W28
2015-W30
2015-W32
2015-W34
2015-W36
2015-W38
2015-W40
2015-W42
2015-W44
2015-W46
2015-W48
2015-W50
2015-W52
2016-W2
2016-W4
2016-W6
2016-W8
2016-W10
2016-W12
2016-W14
2016-W16
2016-W18
2016-W20
2016-W22
2016-W24
2016-W26
2016-W28
2016-W30
2016-W32
2016-W34
2016-W36
2016-W38
2016-W40
2016-W42
2016-W44
2016-W46
allemand, obus, batterie (classe 13)
concert, black, annulé (classe 10)
tombe, courir, cimetière (classe 7)
Photo, site, beau, projet (classe 4)
RI, tuer, ennemi (classe 1)
2e, cl, RI (classe 2)
Expo, musée, exposition (classe 3)
férié (classe 5)
être, politique, seul (classe 6)
grand, guerre, ligne front (classe 8)
mémoire, défi, indexation (classe 9)
hommage, oublier, penser (classe 11)
bataille, cérémonie, centenaire (classe 12)
indexé, livre, commun (classe 14)


classe 2
classe 1
classe 7
classe 6
classe 5
classe 10
classe 11
classe 13
classe 9
classe 4
classe 3
classe 8
classe 12
classe 14
3.8 %
7.7 %
3.7 %
7.8 %
10 %
6.7 %
9.7 %
8.8 %
11.4 %
6 %
9.1 %
10.6 %
2e cl ri_ classe ric sdt rac saint joseph vaast jean neuville rg canonnier louis ra 3e beausejour rz 4e etienne sapeur rire st marius
ri_ tuer ennemi louis jean caporal blessure pierre henri joseph marier sergent soldat valoir auguste bois emile cultivateur jules douaumont eugene lieutenant paul marcel dcd
tombe courir cimetiere honte course jogging milieu footing pietiner tam spectacle honteux indecence cracher limite scene retourner tambour cimetieres danser pietine auschwitz irrespectueux profanation migrant
etre politique seul president chose question accord rappeler mauvais besoin vrai fois entendre sentir gagner massacre affaire petain ete reponse mot hollande apprendre
ca aller tt meme nabilla voir ferie vraiment penser passer parler con putain monde hashtag foutre merde pauvre gens gueule aimer gros tellement tweet regarder
concert black annuler rappeur maire droite chanter annulation raciste homophobe extreme rap inviter choisir polemique gauche blackm insulter choix antisemite chanteur haine kouffars ps traiter
hommage oublier pensee battre liberte rendre patrie heros france paix sacrifice souvenir honneur vie combattre mort combattant tirailleur sacrifier pays homme soldat ancien respect
allemand obus batterie gerbe attaque bombardemen artillerie troupe nuit offensive depot tir village tranchees coquelicot canon tranchee tirer calme londres pluie arc sud neige
memoire anneau defi indexations indexer participant homme objectif tourisme realisees internaute module ficher transcription annotation collaboratif tester record franchir abonnes enrichir saisie indexation see transmission
photo gt site beau decouvrez projet tres nouveau article lire travail decouvrir film video histoire dernier eleves excellent image lecture
expo musee exposition visite ouvrir conference soir memorial week_end ee rencontre centre conseil 18h virtuel ouverture ees public visiter rue colloque decouverte art
grand guerre ligne front arriere pere matricule veille cc militaire registre recherche femme archive retrouver page selection source breton enealogie trace orient collecte portrait nom
bataille ceremonie centenaire armistice novembre matin direct ceremonies debut fin emoire champ emoration mondial speciale anniversaire monument premiere defile suivre commemoration place emission commemorative
indexes livre commun poilu erault partir carte ere deja maj nes heraultais integralement ome mam indexees alfort lo oneetloire indexee annotees etaient indexe indexation collaborative

#greferendum: collecting tweets

- Dates: 6-16 July 2015 (the "international" phase)
- Holistic collect - main hashtag: #greferendum
- NodeXL: ≈ 20,000 tweets per day

204 714 tweets:

- 139 945 retweets (68,36 %)
- 8 686 replies (4,24 %)
- 56 086 unique tweets (27,39 %)

#greferendum: preparing data for analysis

OpenRefine => data (:hashtag) cleaning (:clustering)

TEI P5 XML => (a very basic) text encoding of data (:tweet text) - subcorpus par date

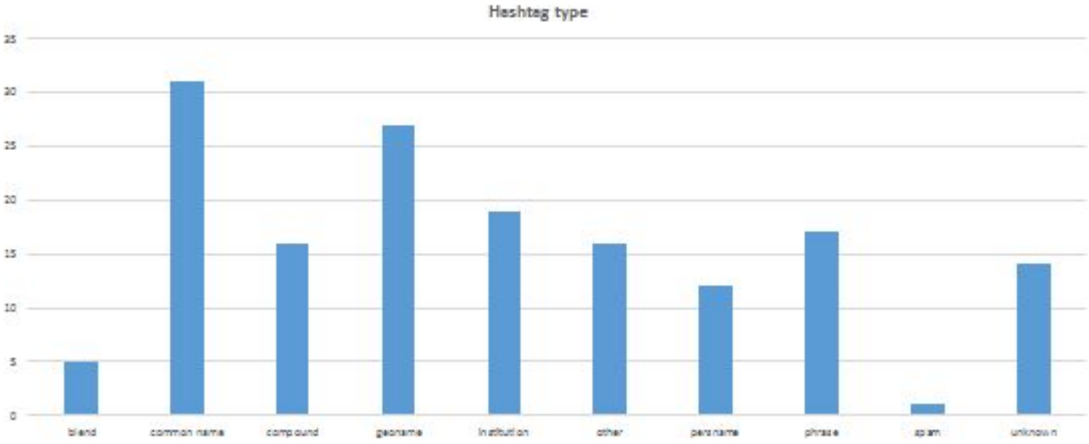
#greferendum: analysing tweet data

- Hashtags
  Qualitative work : a typology of the most frequent hashtags (frq>99, 158 words)
  R (wordcloud package) => textual data visualisation (:hashtags)
- Tweets

  text statistical analysis (:cooccurrences) => TXM textométrie

- Users
  network metrics and visualisation => Gephi
  Qualitative work on most central accounts
- Domains: simple statistics with Voyant tools

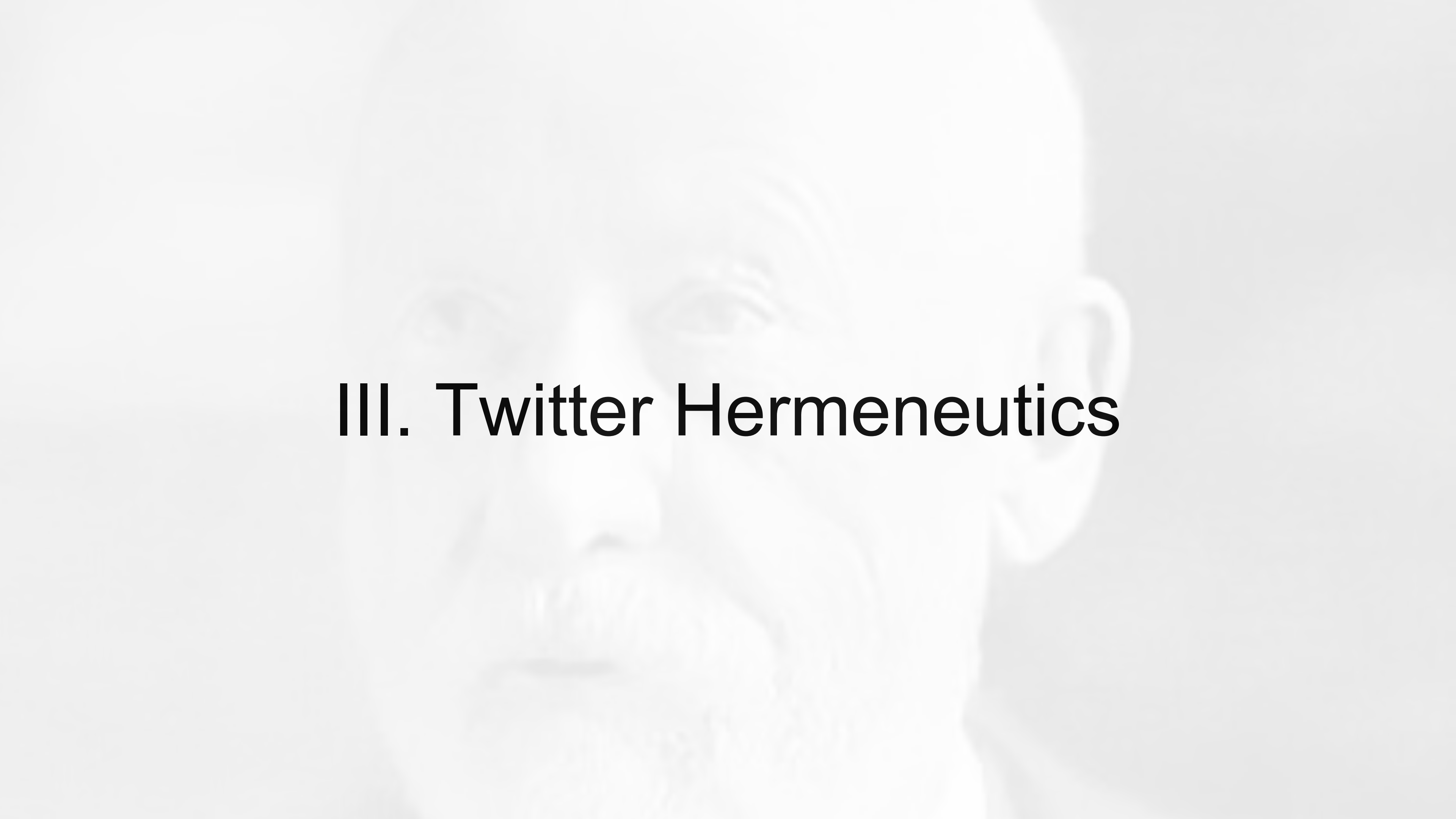
Typology 1


Hashtag type
35
30
25
20
15
10
5
0
blend
common name
compound
geoname
institution
other
persname
phrase
spam
unknown

Typology 2: hashtag function

Commentary: 14/158 (8,861 %)

Tag: 144/158 (91,139 %)



III. Twitter Hermeneutics

Hermeneutics of APIs

Twitter APIs constraints: choosing an API as the first step to interpretation

- Search API: 7 days in the past, around 3000 tweets per hour
  (some workarounds: https://github.com/taspinar/twitterscraper)
  Either sampling / or small corpus
- Streaming API: anticipation of what will be the past
  Limitation: 1% of the tweets that are being published
  Progressive construction of massive corpus

Hermeneutics of keywords: hashtags

Most research on twitter are based on keywords/hashtags which means that:

- The studied object must be quite well-known by the researcher to find the best keywords
- Hashtags / keywords are not conversation

  D’heer Evelien, Vandersmissen Baptist, Neve Wesley De et al., « What are we missing? An empirical exploration in the structural biases of hashtag-based sampling on Twitter », First Monday 22 (2), 16.01.2017.
- Therefore
  - collecting massive data != collecting exhaustive data
  - sampling data can be better than massively collecting data
- Numerous ways to understand what a hashtag is

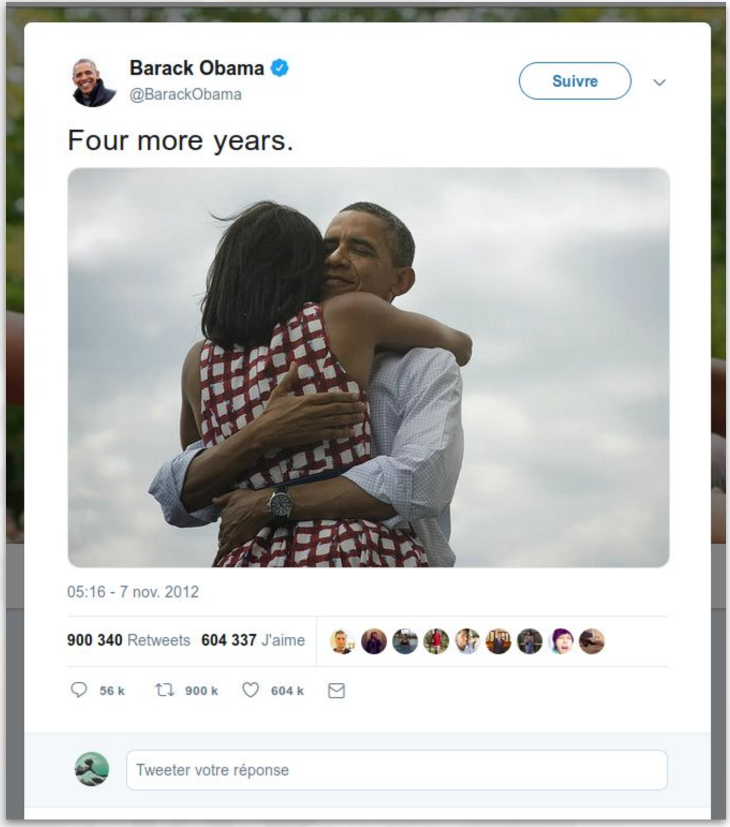
… some thoughts from the #greferendum corpus

Hashtags

- tell the big story (quantitative + relational analysis)
- reveal different temporalities related to connected histories of the Eurozone crisis
- a common conversation? a European space? (cf. works of Camille Roth)

Hermeneutics of networks

See: http://theconversation.com/four-more-years-that-obama-tweet-and-the-politics-of-intimacy-10606



Barack Obama
@BarackObama

Four more years.

05:16 - 7 nov. 2012

900 340 Retweets 604 337 J'aime

… some thoughts from the #greferendum corpus

- main actors in a graphe - main actors irl?
- what is a Twitter network?

Hermeneutics of tools

Hypothesis: a tool = a method = a theory = a specific way to interpret data

- **Gephi**

  Visualizing social networks => sociology of social networks != sociology of field and *habitus*
- **IRaMuTeQ**

  Théorie des mondes lexicaux

  Reinert Max, « Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte », Les cahiers de l'analyse des données 8 (2), 1983, pp. 187-198.
  - French School of Data Analysis (yes, there is one)
  - Mondes lexicaux: one point of view = one coherent set of words = social representations (= Émile Durkheim)

… some thoughts from the #greferendum corpus

- Dataviz is useful… metrics are important
- How to be comfortable with the algorithm? (transparency, stability issues)
- Need for tools that behave well with multilingual corpora (TXM is fine)
- Preservation and sharing issues

Hermeneutics of Twitter: Twitter as a primary source

- A primary source in the historian's point of view
  Traditionally something that is fixed within a set framework (= the Archive)

- Twitter is always moving, is a "source" in the original meaning (source of water): something that is endlessly flowing, that cannot by definition be fixed
  Ex: The "four more years" Obama tweet

- What we do
  - transforming something that is not supposed to stay still into an archive, something that is fixed
  - What do we lose in this process?

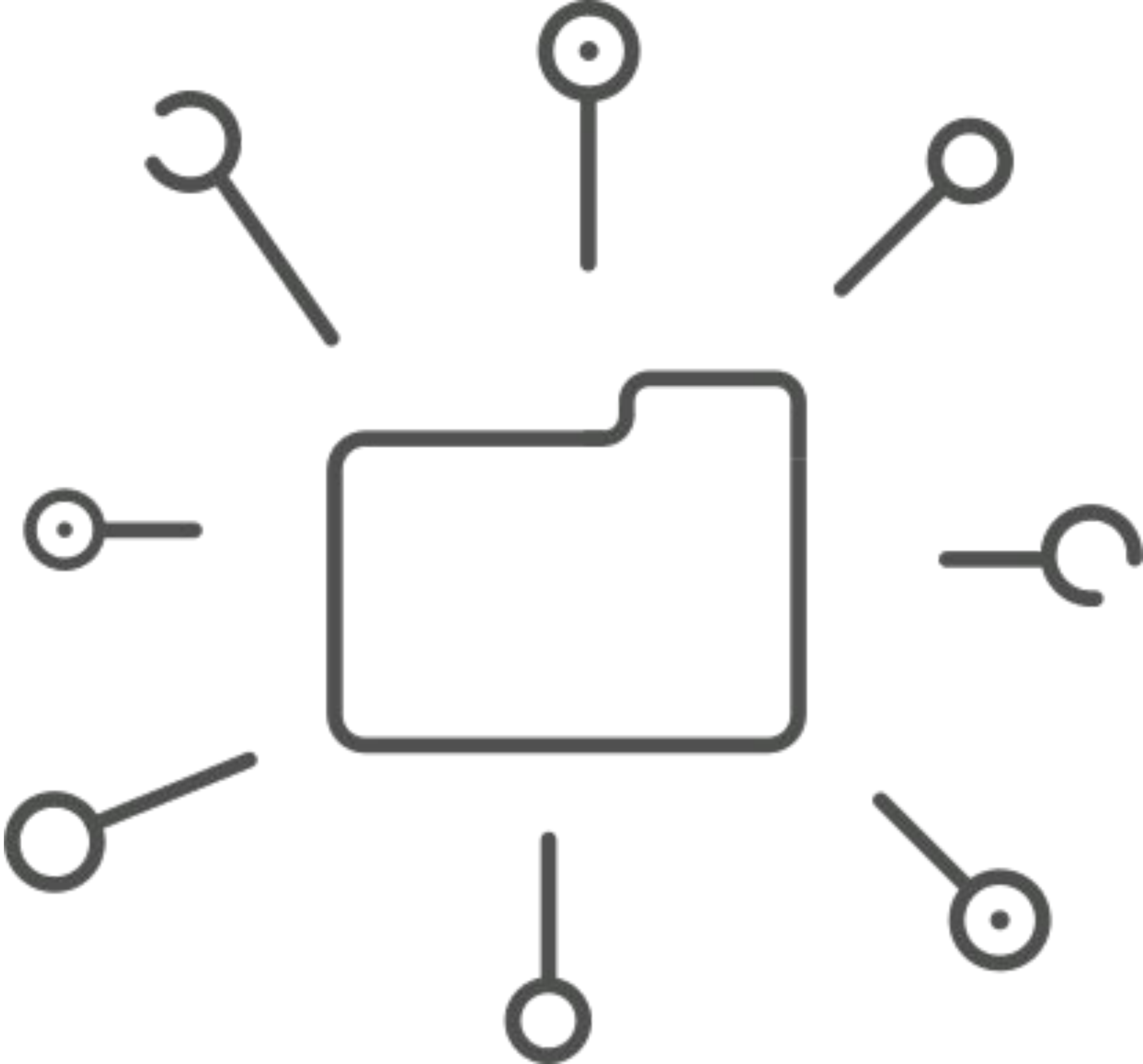
Hermeneutics of Metadata

Information embedded in the metadata are crucial for the interpretation of a / numerous tweet(s)

Ex: timestamps and the interpretation of temporalities

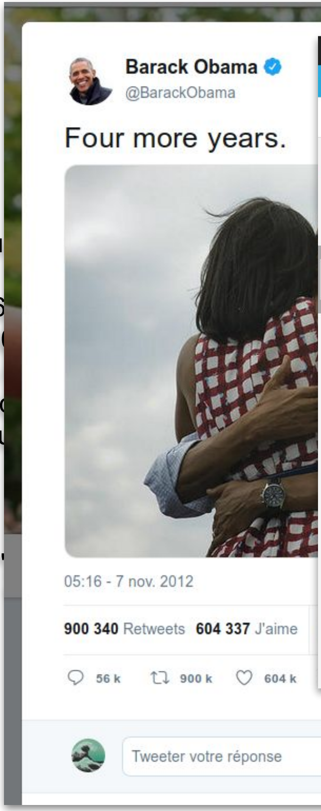
- Timestamps in tweet metadata correspond to the unending (well…) and continuous feed of tweets that is the essence of Twitter
    - Western vision of time
- Many more artifacts of temporalities are embedded in the text of a tweet
    - How to deal with other kind of temporalities whereas collective memory, for instance, is the result of an interlacing of temporalities

Conclusion: what is the allure of born digital archive?



What is a tweet?

```
[{
  "created_at": "Thu
2017",
  "id": 87799460456
  "id_str": "8779946
  "text": "Creating a
Angular, Part 1: Adc
https://t.co/xFox78ju
  "truncated": false,
  "entities": {
    "hashtags": [{
      "text": "Angular"
      "indices": [103,
    }],
```

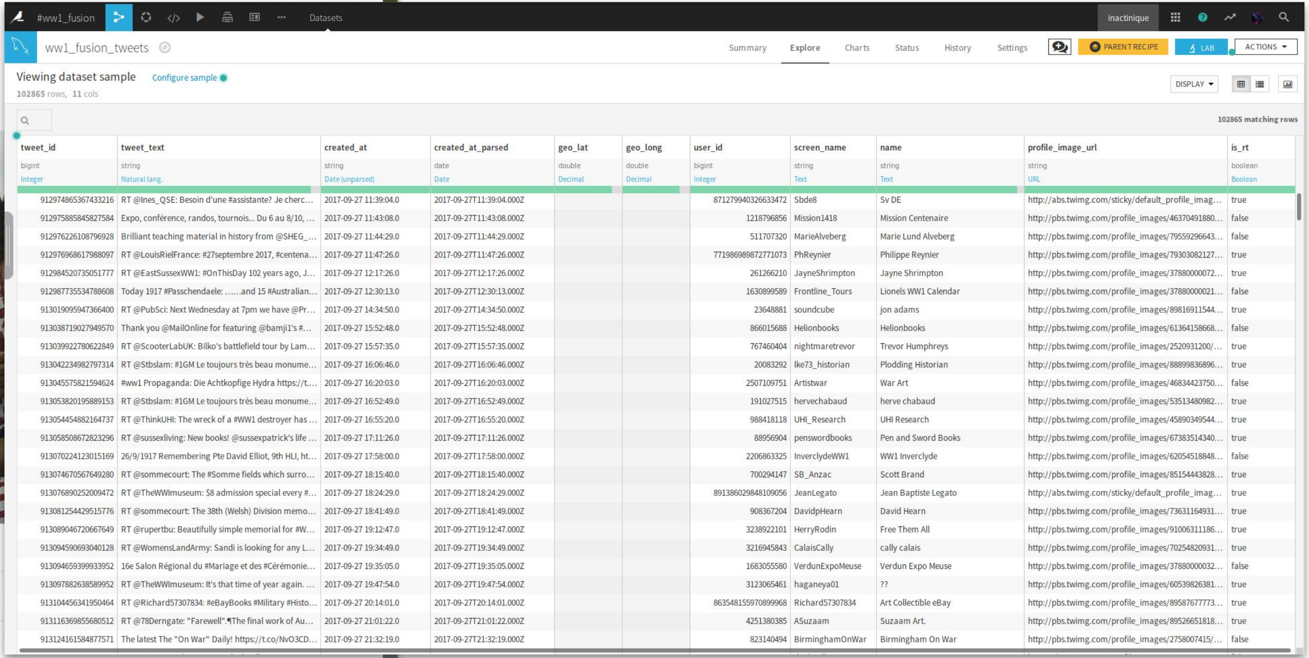


Barack Obama
@BarackObama

Four more years.

05:16 - 7 nov. 2012

900 340 Retweets 604 337 J'aime



Viewing dataset sample

102865 rows, 11 cols

102865 matching rows

tweet_id	tweet_text	created_at	created_at_parsed	geo_lat	geo_long	user_id	screen_name	name	profile_image_url	is_rt
bigint	string	string	date	double	double	bigint	string	string	string	boolean
Integer	Natural lang.	Date (unparsed)	Date	Decimal	Decimal	Integer	Text	Text	URL	Boolean
912974865367433216	RT @Ines_QSE: Besoin d'une #assistante? Je cherc…	2017-09-27 11:39:04.0	2017-09-27T11:39:04.000Z			871279940326633472	Sbde8	Sv DE	http://abs.twimg.com/sticky/default_profile_imag…	true
912975885845827584	Expo, conférence, randos, tournois... Du 6 au 8/10, …	2017-09-27 11:43:08.0	2017-09-27T11:43:08.000Z			1218796856	Mission1418	Mission Centenaire	http://pbs.twimg.com/profile_images/46370491880…	false
912976226108796928	Brilliant teaching material in history from @SHEG_…	2017-09-27 11:44:29.0	2017-09-27T11:44:29.000Z			511707320	MarieAlveberg	Marie Lund Alveberg	http://pbs.twimg.com/profile_images/79559296643…	false
912976968617988097	RT @LouisRielFrance: #27septembre 2017, #centena…	2017-09-27 11:47:26.0	2017-09-27T11:47:26.000Z			771986989872771073	PhReynier	Philippe Reynier	http://pbs.twimg.com/profile_images/79303082127…	true
912984520735051777	RT @EastSussexWW1: #OnThisDay 102 years ago, J…	2017-09-27 12:17:26.0	2017-09-27T12:17:26.000Z			261266210	JayneShrimpton	Jayne Shrimpton	http://pbs.twimg.com/profile_images/37880000072…	true
912987735534788608	Today 1917 #Passchendaele: ……and 15 #Australian…	2017-09-27 12:30:13.0	2017-09-27T12:30:13.000Z			1630899589	Frontline_Tours	Lionels WW1 Calendar	http://pbs.twimg.com/profile_images/37880000021…	false
913019095947366400	RT @PubSci: Next Wednesday at 7pm we have @Pr…	2017-09-27 14:34:50.0	2017-09-27T14:34:50.000Z			23648881	soundcube	jon adams	http://pbs.twimg.com/profile_images/89816911544…	true
913038719027949570	Thank you @MailOnline for featuring @bamji1's #…	2017-09-27 15:52:48.0	2017-09-27T15:52:48.000Z			866015688	Helionbooks	Helionbooks	http://pbs.twimg.com/profile_images/61364158668…	false
913039922780622849	RT @ScooterLabUK: Bilko's battlefield tour by Lam…	2017-09-27 15:57:35.0	2017-09-27T15:57:35.000Z			767460404	nightmaretrevor	Trevor Humphreys	http://pbs.twimg.com/profile_images/2520931200/…	true
913042234982797314	RT @Stbslam: #1GM Le toujours très beau monume…	2017-09-27 16:06:46.0	2017-09-27T16:06:46.000Z			20083292	lke73_historian	Plodding Historian	http://pbs.twimg.com/profile_images/88899836896…	true
913045575821594624	#ww1 Propaganda: Die Achtkopfige Hydra https://t.…	2017-09-27 16:20:03.0	2017-09-27T16:20:03.000Z			2507109751	Artistwar	War Art	http://pbs.twimg.com/profile_images/46834423750…	false
913053820195889153	RT @Stbslam: #1GM Le toujours très beau monume…	2017-09-27 16:52:49.0	2017-09-27T16:52:49.000Z			191027515	hervechabaud	herve chabaud	http://pbs.twimg.com/profile_images/53513480982…	true
913054454882164737	RT @ThinkUHI: The wreck of a #WW1 destroyer has …	2017-09-27 16:55:20.0	2017-09-27T16:55:20.000Z			988418118	UHI_Research	UHI Research	http://pbs.twimg.com/profile_images/45890349544…	true
913058508672823296	RT @sussexliving: New books! @sussexpatrick's life …	2017-09-27 17:11:26.0	2017-09-27T17:11:26.000Z			88956904	penswordbooks	Pen and Sword Books	http://pbs.twimg.com/profile_images/67383514340…	true
913070224123015169	26/9/1917 Remembering Pte David Elliot, 9th HLI, ht…	2017-09-27 17:58:00.0	2017-09-27T17:58:00.000Z			2206863325	InverclydeWW1	WW1 Inverclyde	http://pbs.twimg.com/profile_images/62054518848…	false
913074670567649280	RT @sommecourt: The #Somme fields which surro…	2017-09-27 18:15:40.0	2017-09-27T18:15:40.000Z			700294147	SB_Anzac	Scott Brand	http://pbs.twimg.com/profile_images/85154443828…	true
913076890252009472	RT @TheWWImuseum: $8 admission special every #…	2017-09-27 18:24:29.0	2017-09-27T18:24:29.000Z			891386029848109056	JeanLegato	Jean Baptiste Legato	http://abs.twimg.com/sticky/default_profile_imag…	true
913081254429515776	RT @sommecourt: The 38th (Welsh) Division memo…	2017-09-27 18:41:49.0	2017-09-27T18:41:49.000Z			908367204	DavidpHearn	David Hearn	http://pbs.twimg.com/profile_images/73631164931…	true
913089046720667649	RT @rupertbu: Beautifully simple memorial for #W…	2017-09-27 19:12:47.0	2017-09-27T19:12:47.000Z			3238922101	HerryRodin	Free Them All	http://pbs.twimg.com/profile_images/91006311186…	true
913094590693040128	RT @WomensLandArmy: Sandi is looking for any L…	2017-09-27 19:34:49.0	2017-09-27T19:34:49.000Z			3216945843	CalaisCally	cally calais	http://pbs.twimg.com/profile_images/70254820931…	true
913094659399933952	16e Salon Régional du #Mariage et des #Cérémonie…	2017-09-27 19:35:05.0	2017-09-27T19:35:05.000Z			1683055580	VerdunExpoMeuse	Verdun Expo Meuse	http://pbs.twimg.com/profile_images/37880000032…	false
913097882638589952	RT @TheWWImuseum: It's that time of year again. …	2017-09-27 19:47:54.0	2017-09-27T19:47:54.000Z			3123065461	haganeya01	??	http://pbs.twimg.com/profile_images/60539826381…	true
913104456341950464	RT @Richard57307834: #eBayBooks #Military #Histo…	2017-09-27 20:14:01.0	2017-09-27T20:14:01.000Z			863548155970899968	Richard57307834	Art Collectible eBay	http://pbs.twimg.com/profile_images/89587677773…	true
913116369855680512	RT @78Derngate: "Farewell".¶The final work of Au…	2017-09-27 21:01:22.0	2017-09-27T21:01:22.000Z			4251380385	ASuzaam	Suzaam Art.	http://pbs.twimg.com/profile_images/89526651818…	true
913124161584877571	The latest The "On War" Daily! https://t.co/NvO3CD…	2017-09-27 21:32:19.0	2017-09-27T21:32:19.000Z			823140494	BirminghamOnWar	Birmingham On War	http://pbs.twimg.com/profile_images/2758007415/…	false