DELIVERABLE SUBMISSION SHEET

To: Susan Fraser *(Project Officer)*

EUROPEAN COMMISSION
Directorate-General Information Society and Media
EUFO 1165A
L-2920 Luxembourg

From:

Project acronym: PHEME Project number: 611233

Project manager: Kalina Bontcheva

Project coordinator The University of Sheffield (USFD)

The following deliverable:

Deliverable title: LOD-based Reasoning about Rumours: Final Prototype

Deliverable number: D4.1.2

Deliverable date: 31 August 2016

Partners responsible: Ontotext AD

Status: ☒ Public ☐ Restricted ☐ Confidential

is now complete. ☒ It is available for your inspection.
☒ Relevant descriptive documents are attached.

The deliverable is:

☒ a document
☐ a Website (URL: .....)
☐ software (.....)
☐ an event
☒ other (...Prototype.....)

Sent to Project Officer:	Sent to functional mail box:	On date:
Susan.Fraser@ec.europa.eu	*CNECT-ICT*-611233*@ec.europa.eu*	30 August 2016

D4.1.2 LOD-based reasoning

Laura Tolo̧i (Ontotext), Georgi Georgiev (Ontotext), Atanas Popov (Ontotext)

Abstract.
FP7-ICT Collaborative Project ICT-2013-611233 PHEME
Deliverable D4.1.2 (WP4.1)

In this deliverable we show that LOD annotations are a powerful tool for semantic enrichment of Social Media microposts, allowing for reasoning with information that is not transmitted directly through the Social Media channels, but available in rich knowledge bases. Given the very short text of tweets for example, such enrichment provides the necessary context, which is crucial for understanding opinions, trends, veracity in Social Media. We choose a showcase that is highly relevant for the international political scene at the moment of the deliverable, namely a post-Brexit analysis. We show that one can automatically mine the general opinion of the main UK administrative regions and cities. We also identified the main actors of the political scene, with side comments on their age. Controversiality of the main political figures also becomes easily available, by aggregation of the support/deny annotations of the tweets mentioning them.

Keyword list: LOD, semantic annotation, semantic search, Brexit, Twitter, Geonames, DBpedia.

Project	PHEME No. 611233
Delivery Date	August 29, 2016
Contractual Date	August 30, 2016
Nature	Report
Reviewed By	Leon Derczynski
Web links	http://pheme.ontotext.com/graphdb/sparql
Dissemination	PU

PHEME **Consortium**

This document is part of the PHEME research project (No. 611233), partially funded by the FP7-ICT Programme.

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

MODUL University Vienna GMBH
Am Kahlenberg 1
1190 Wien
Austria
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

ATOS Spain SA
Calle de Albarracin 25
28037 Madrid
Spain
Contact person: Tomás Pariente Lobo
E-mail: tomas.parientelobo@atos.net

iHub Ltd.
NGONG, Road Bishop Magua Building
4th floor
00200 Nairobi
Kenya
Contact person: Rob Baker
E-mail: robbaker@ushahidi.com

The University of Warwick
Kirby Corner Road
University House
CV4 8UW Coventry
United Kingdom
Contact person: Rob Procter
E-mail: Rob.Procter@warwick.ac.uk

Universitaet des Saarlandes
Campus
D-66041 Saarbrücken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Georgi Georgiev
E-mail: georgiev@ontotext.com

King's College London
Strand
WC2R 2LS London
United Kingdom
Contact person: Robert Stewart
E-mail: robert.stewart@kcl.ac.uk

SwissInfo.ch
Giacomettistrasse 3
3000 Bern
Switzerland
Contact person: Peter Schibli
E-mail: Peter.Schibli@swissinfo.ch

Changes

Version	Date	Author	Changes
1.0	26.08.2016	Laura Toloşi	creation
	28.08.2016	Leon Derczynski	review
1.1	28.08.2016	Laura Toloşi	correction

Executive Summary

In this deliverable we show that LOD annotations are a powerful tool for semantic enrichment of Social Media microposts, allowing for reasoning with information that is not transmitted directly through the Social Media channels, but available in rich, hopefully unbiased knowledge bases. Given the very short text of tweets for example, such enrichment provides with the necessary context, which is crucial for understanding opinions, trends, veracity in Social Media. LOD enrichment allows for the computer algorithms to 'understand' tweets in a way that a human would, by referring to common knowledge external to the micropost.

We choose a showcase that is highly relevant for the international political scene at the moment of the deliverable, namely a post-Brexit analysis. Post-Brexit discussions on Twitter provide with insights on the mixed feelings, attitude, propaganda, interests that follow the referendum and precede the political actions that need to be taken.

We show that one can automatically mine the general opinion of the main UK administrative regions and cities. We also identified the main actors of the political scene, with side comments on their age - an aspect that has been so many times brought to the public attention and even used for manipulating the opinion of the voters. Reasoning about age or birth year is only possible via LOD annotations. Controversiality of the main political figures also becomes easily available, by aggregation of the support/deny annotations of the tweets mentioning them.

Contents

Chapter 1

Semantic enrichment of post-Brexit tweets

The goal is to demonstrate the benefit of semantic annotations of Social Media microposts. We took advantage of the already functional Journalism Dashboard (WP8), which allows to make a request for a certain topic to be streamed via the Twitter API, analyzed by the PHEME journalism pipeline and become available for querying via GraphDB SPARQL endpoint[1]. A topic on Brexit was started on 06.07.2016, which to the date of this deliverable had gathered more than 800,000 annotated tweets in GraphDB.

Note that insightful analyses of Twitter datasets on Brexit has been conducted before the referendum by Ontotext (Toloşi, 2016).

1.1 Querying Pheme GraphDB for LOD annotations

We show several relevant queries about UK regions and cities, as well as public figures mentioned in Brexit tweets.

1.1.1 UK Regions

With the SPARQL query in Figure 1.1 to the Pheme SPARQL repository, one can extract all tweets that mention at least one of the large administrative regions of UK (England, Scotland, Wales and Northern Ireland). The administrative regions are available in Geonames [2] (line 18 in Figure 1.1). The parent country (in our case UK) also comes from Geonames (line 17 in Figure 1.1).

We call the resulting dataset the *regions* dataset. It consists of $19,674$ records with the

[1] `http://pheme.ontotext.com/graphdb/sparql`

[2] `www.geonames.org`

```
1  PREFIX sioc: <http://rdfs.org/sioc/ns#>
2  PREFIX dlpo: <http://www.semanticdesktop.org/ontologies/2011/10/05/dlpo#>
3  PREFIX pub: <http://ontology.ontotext.com/taxonomy/>
4  PREFIX geo-ont: <http://www.geonames.org/ontology#>
5
6  select ?a ?text ?date ?tgsinst ?name where {
7      ?a a pheme:Tweet .
8      ?a pheme:createdAt ?date .
9      ?a dlpo:textualContent ?text .
10     ?a pheme:containsMention ?mention .
11     ?mention pheme:name ?name .
12     ?mention pheme:mentionType ?type.
13     FILTER (?type = "Location"^^xsd:string) .
14     ?mention pheme:inst ?tgsinst .
15     ?tgsinst pub:exactMatch ?o.
16     ?o a geo-ont:Feature.
17     ?o geo-ont:parentCountry <http://sws.geonames.org/2635167/>.
18     ?o geo-ont:featureCode geo-ont:A.ADM1.
19     ?a pheme:dataChannel "6c9fcb94" .
20     ?a pheme:version "v7" .
21 }
```

Figure 1.1: SPARQL query for large administrative regions in UK mentioned in Twitter.

following information: *tweet id*, *tweet text*, *tweet date*, *URI* of the region mentioned, and its *name*, as spelled in the tweet.

1.1.2 UK Cities

The SPARQL query in Figure 1.2 retrieves tweets that mention cities in UK that have a population larger than $5,000$ people and the region to which they belong. The information is available from Geonames.

We call the resulting dataset the *cities* dataset. It consists of $21,130$ records with the following fields: *tweet id*, *tweet text*, *tweet date*, *URI* of the city mentioned, its *name* as spelled in the tweet, the *geographic coordinates* of the city, the *larger administrative region*

1.1.3 People

The SPARQL query in Figure 1.3 is used to retrieve the mentions of public personalities with their birth date. The information is available from DBpedia[3] (Bizer et al., 2009) (lines 17-18 in Figure 1.3). Lines 11 and 12 of the query filter out so called 'generated' entities, that are predicted by Ontotext's concept tagger as of type Person, but no DBpedia article is found about her. Such a filter focuses the analysis on public figures.

[3]wiki.dbpedia.org

```
3  PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
4  PREFIX pheme: <http://www.pheme.eu/ontology/pheme#>
5  PREFIX sioc: <http://rdfs.org/sioc/ns#>
6  PREFIX dlpo: <http://www.semanticdesktop.org/ontologies/2011/10/05/dlpo#>
7  PREFIX pub: <http://ontology.ontotext.com/taxonomy/>
8  PREFIX geo-ont: <http://www.geonames.org/ontology#>
9
10 select ?a ?text ?date ?tgsinst ?name ?coordinates ?region ?regionname where {
11     ?a a pheme:Tweet .
12     ?a pheme:createdAt ?date .
13     ?a dlpo:textualContent ?text .
14     ?a pheme:containsMention ?mention .
15     ?mention pheme:name ?name .
16     ?mention pheme:mentionType ?type.
17     FILTER (?type = "Location"^^xsd:string) .
18     ?mention pheme:inst ?tgsinst .
19     ?tgsinst pub:coordinateLocation ?coord .
20     ?coord pub:hasValue ?coordinates .
21     ?tgsinst pub:exactMatch ?o.
22     ?o a geo-ont:Feature.
23     ?o geo-ont:population ?population.
24     ?o geo-ont:parentCountry <http://sws.geonames.org/2635167/>.
25     ?o geo-ont:parentADM1 ?region .
26     ?region geo-ont:name ?regionname .
27     ?a pheme:dataChannel "6c9fcb94" .
28     ?a pheme:version "v7" .
29 }
```

Figure 1.2: SPARQL query for cities in UK mentioned in Twitter.

We call the resulting dataset the *people* dataset. It consists of $53,355$ mentions with the following information: *tweet id*, *tweet text*, *tweet date*, *URI* of the person mentioned, its *name* as spelled in the tweet and the *birth date*.

Annotations of whether a tweet is supporting, denying or interrogative are available in GraphDB. The algorithm that automatically assigns such labels is presented in Lukasik et al. (2015). Figure 1.4 is essentially counting how many support/deny/ question tweets are mentioning a particular person (Theresa May in the example). Line 12 specifies the URI of the person of interest. In the following Chapter, we applied this query repeatedly to the top ten most mentioned personalities in the Brexit context.

```
PREFIX pheme: <http://www.pheme.eu/ontology/pheme#>
PREFIX dlpo: <http://www.semanticdesktop.org/ontologies/2011/10/05/dlpo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX pub: <http://ontology.ontotext.com/taxonomy/>
PREFIX geo-ont: <http://www.geonames.org/ontology#>
select ?a ?text ?date ?tgsinst ?mention ?name  ?born where {
    ?a a pheme:Tweet .
    ?a dlpo:textualContent ?text .
    ?a pheme:createdAt ?date .
    ?a pheme:containsMention ?mention .
    ?mention pheme:generated ?gen .
    FILTER (?gen = "false").
    ?mention pheme:name ?name .
    ?mention pheme:mentionType ?type.
    FILTER (?type = "Person"^^xsd:string) .
    ?mention pheme:inst ?tgsinst .
    ?tgsinst pub:dateOfBirth ?dob .
    ?dob pub:hasValue ?born .
    #FILTER(?tgsinst != <http://ontology.ontotext.com/resource/tsk4vb736328>).
    ?a pheme:dataChannel "6c9fcb94" .
    ?a pheme:version "v7" .
}order by ?name
```

Figure 1.3: SPARQL query for people mentioned in the Brexit topic on Twitter.

1.2 Relevance to PHEME

1.2.1 Relevance to project objectives

The Pheme GraphDB [4] repository [5] stores at the moment tweets and annotations that are being produced by the PHEME journalism pipeline. A large variety of annotations like named entities with LOD references, as well as veracity-specific metadata such as rumor score and support/deny indications are openly accessible via SPARQL queries. This chapter demonstrates how to query the Brexit topic with a few relevant examples.

1.2.2 Relation to other workpackages

All SPARQL queries are best understood by taking a look at the schema of the Pheme ontology (Figure 1.5), which has been developed in WP2 and was subject to continuous improvements during the integration stages (WP6) and the development of the Journalism Dashboard (WP8).

[4] http://ontotext.com/products/graphdb/
[5] http://pheme.ontotext.com/graphdb/sparql

```
PREFIX pheme: <http://www.pheme.eu/ontology/pheme#>
PREFIX dlpo: <http://www.semanticdesktop.org/ontologies/2011/10/05/dlpo#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX pub: <http://ontology.ontotext.com/taxonomy/>
select ?contr (count(?contr) as ?c) where {
    ?a a pheme:Tweet .
    ?a pheme:containsMention ?mention .
    ?a pheme:sdq ?contr .
    ?mention pheme:generated ?gen .
    FILTER (?gen = "false").
    ?mention pheme:inst ?tgsinst .
    FILTER(?tgsinst = <http://ontology.ontotext.com/resource/tsnkdntepk3k>).
    ?a pheme:dataChannel "6c9fcb94" .
    ?a pheme:version "v7" .
} group by ?contr

```

Figure 1.4: SPARQL query for counting the support / deny / question tweets mentioning Theresa May.

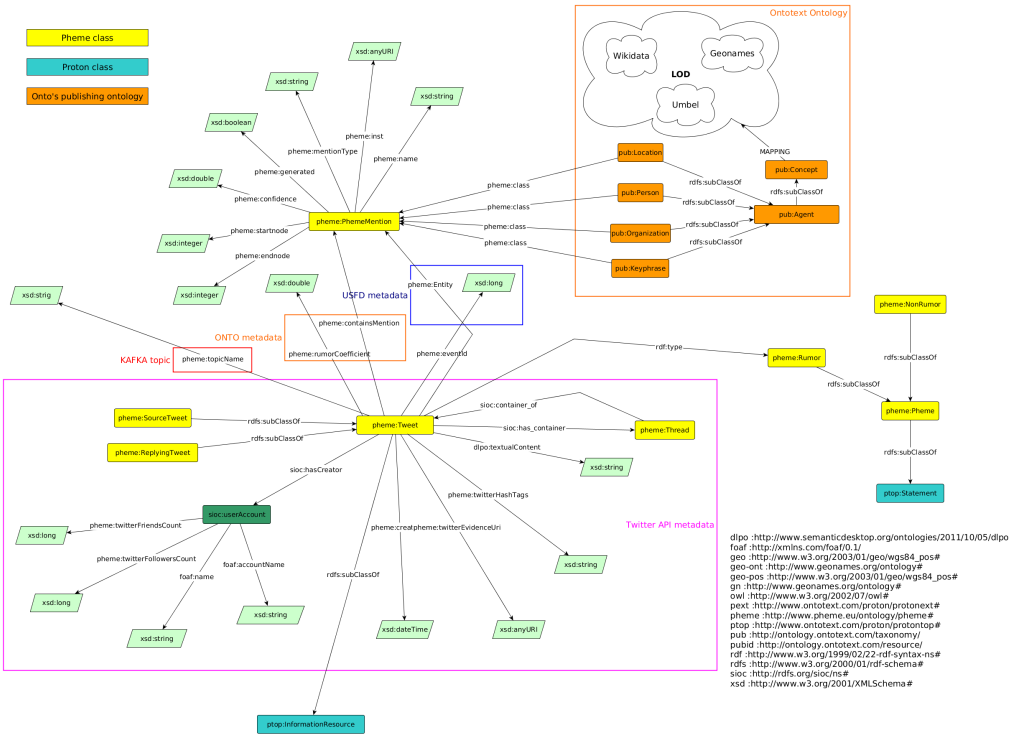

Figure 1.5: Schematic representation of Pheme Ontology and links to LOD.

Chapter 2

Brexit followup analysis

In this section, we show how LOD annotations and reasoning help gain insights about popular topics in Social Media after Brexit, with focus on the involvement of various geographical locations (cities and regions) from UK and public figures.

2.1 Preprocessing: grouping retweets or similar tweets in clusters

Our datasets inevitably contain retweets, meaning that a micropost is expected to appear one or many times, hence the regions mentioned will also be amplified by retweeting. Whereas the number of retweets hints to the popularity or importance of a topic – therefore of the locations mentioned as well – for some statistics and insights it is better to work with a 'clean' dataset, consisting of tweets that are unique.

However, not always the information that a tweet is a retweet is received via the Twitter API. For example, a user can simply type RT and repost a message as an idependent tweet.

We used a basic strategy to group together almost identical tweets. We first define the distance between the text of two tweets as their *least common sub-sequence* (lcs). Then we empirically observed that the distribution of the distances in our dataset is bimodal, with a gap around the value of lcs = 50. (figure)

We therefore grouped together in clusters tweets having mutual distances smaller than 50.

Cluster of 24 *tweets:*

```
[1] "KEEP THE ENGLISH FLAG ON THE OAK https://t.co/zTQnmaaBM4
#england #freedom #EU #brexit #referendum"
```

```
[2] "RT @Col_Connaughton:  KEEP THE ENGLISH FLAG ON THE
OAK https://t.co/zTQnmaaBM4 #england #freedom #EU #brexit
#referendum"

[3] "KEEP THE ENGLISH FLAG ON THE OAK https://t.co/zTQnmaaBM4
#england #freedom #EU #brexit #referendum"

[4] "KEEP THE ENGLISH FLAG ON THE OAK https://t.co/zTQnmaaBM4
#england #freedom #EU #brexit #referendum"

...

[24] "KEEP THE ENGLISH FLAG ON THE OAK https://t.co/zTQnmaaBM4
#england #freedom #EU #brexit #referendum"
```

To remove redundancy when necessary, we keep only one tweet from each cluster, namely the first posted, in chronological order.

2.2 Analysis of mentions of UK regions

In this section, the *regions* dataset (Section 1.1.1) is analyzed. After preprocessing with the method for grouping redundant tweets, the *regions* dataset shrunk from $19,674$ tweets to $4,917$ clusters.

2.2.1 Alternative ways of referring to UK regions

LOD is a powerful tool for reconciling alternative spellings or names of the same location. Table 2.1 shows for example all forms that are mapped to Northern Ireland `https://en.wikipedia.org/wiki/Northern_Ireland`. Some of these would be retrieved with simple text normalization such as transformation to lower case or removing spaces, but some references are not so trivial, such as "Six Counties", or "Norn Iron". The tagger also makes one mistake, namely it thinks that "NOT" stands for Northern Ireland, in the tweet:

```
  Well I'm sure glad UK is going to impose #Brexit on a
pro-EU Northern Ireland (NOT). https://t.co/98THqPGwQg
```

2.2.2 Regions by number of tweets

We found tweets mentioning all four main administrative regions of UK, namely England, Wales, Scotland and Northern Ireland. Figure 2.1 shows the number of mentions of each region – resulting after excluding retweets, by using the method from Section **??**. The result reflects one of the main post-Brexit political disputes, namely Scotland's referendum outcome opposing the overall vote. Scotland is mentioned most often in Tweets, according to our dataset.

Form	Number of unique tweets
Northern Ireland	253
Ireland	55
N Ireland	34
N. Ireland	17
Northern Irish	13
N.Ireland	9
Ulster	6
Northern Irish politics	5
Norn Iron	4
the Northern Ireland	3
North Ireland	3
the North of Ireland	2
North of Ireland	2
north of Ireland	2
the occupied 6 counties	1
the north of Ireland	1
The Northern Ireland	1
Six Counties	1
Six counties	1
six counties	1
NOTHERN IRELAND	1
Nothern Ireland	1
NOT	1
northern Irish	1
Northern Irelands	1
northern ireland	1
Ire	1

Table 2.1: Ways in which Twitter users refer to Northern Ireland, according to the concept tagger.

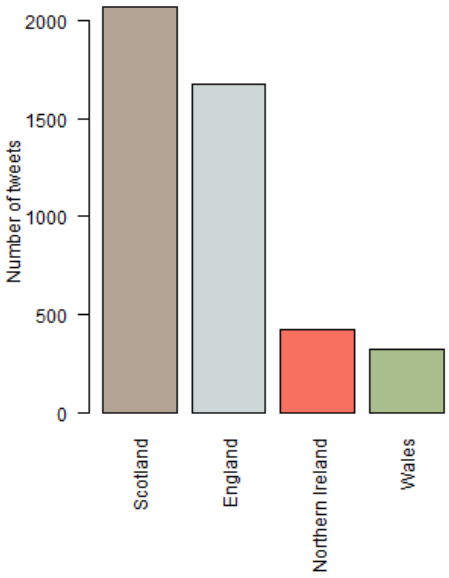

Figure 2.1: Regions mentioned in tweets. Counts are unique tweets, excluding retweets.

	Tweets mention England	Tweets mention Scotland
tweets contain #indyref2	34	233
tweets don't contain #indyref2	2070	1839

Table 2.2: Association betweet hashtag #indyref2 and tweets mentioning either England or Scotland.

2.2.3 Key-terms significantly associated with a region

An interesting question is: what topics are being discussed in relation to one region, as opposed to another region? We considered the interesting comparison of tweets that mention Scotland and tweets that mention England. We tokenized the tweets and looked at the terms that most often occur in the combined set of tweets mentioning either England or Scotland.

For a particular term, say hashtag '#indyref2', we can compare the frequency in tweets mentioning England and tweets mentioning Scotland. A Fisher's exact test (Fisher, 1922) can quantify the significance of association of the term with one of the regions. Eg. the distribution of the hashtag '#indyref2' in the two sets of tweets is given by the contingency table 2.2 and the significance p-value of the Fisher exact test is $1.058687e-40$. This means that the term is significantly associated with Scotland, which is expected as the hashtag stands for Scotland independence referendum.

The following list gives the terms most significantly polarized between England and Scotland, sorted by Fisher's test p-value, adjusted for multiple testing (Holm, 1979). Note that we haven't removed stopwords, as some of them may be informative, such as 'in',

'out', 'for', 'against', 'before', 'after', etc.

scotland ($9.613994e-311$)**, england** ($1.890835e-292$)**, #england** ($4.937512e-144$)**, #scotland** ($2.115173e-139$)**, #indyref2** ($2.085614e-38$)**, scotland's** ($8.005337e-38$)**, england's** ($1.476495e-25$)**, @nicolasturgeon** ($1.887104e-19$)**, sturgeon** ($4.184612e-16$)**, independence** ($1.140539e-14$)**, #uk** ($2.671011e-11$)**, #remain** ($5.556592e-11$)**, #trident** ($6.300122e-11$)**, eu** ($1.235730e-10$)**, bank** ($4.228756e-10$)**, #snp** ($9.103106e-10$)**, nicola** ($3.417163e-07$)**, scottish** ($5.039970e-07$)**, @thesnp** ($3.118258e-06$)**, independent** ($5.812025e-06$)**, voted** ($6.198238e-06$)**, could** ($4.216669e-05$)**, vote** ($6.183181e-05$)**, remain** ($1.301554e-04$)**, be** ($2.869511e-04$)**, #northernireland** ($6.334449e-04$)**, rt** ($9.029809e-04$)**, to** ($9.885050e-04$)**, last** ($1.179732e-03$)**, stay** ($1.945785e-03$)**, #wales** ($2.503500e-03$)**, leave** ($3.567595e-03$)**, in** ($7.699819e-03$)**, wales** ($1.888429e-02$)**, scots** ($2.055371e-02$)**, now** ($2.424857e-02$)**, world** ($2.918002e-02$)**.**

The list of keywords makes much sense, as it includes topics like Scotland's Independence, Scotland's First Minister Nicola Sturgeon, the Scottish National Party (SNP), the voting for leaving or staying in EU, etc.

2.3 Analysis of mentions of cities in Brexit follow-up

In this section, the *cities* dataset (Section 1.1.2) is analyzed. After preprocessing with the method for grouping of redundant tweets, the dataset shrunk from $21,130$ tweets to $8,331$.

Figure 2.2 shows on the map the locations that are most frequently mentioned in our dataset. Mapping to geographical coordinates is only possible via LOD annotations. The size of the points is proportional to the frequency of mentions. London is the most mentioned city.

2.4 Analysis of mentions of people in Brexit follow-up

We explore the full potential of LOD by investigating the age of people mentioned in the post Brexit Twitter posts. Figure 2.3 shows the distribution of year of birth of person mentions, as resulting from the *People* dataset (1.1.3).

2.4.1 Historical figures

We noted the long tail towards the early years, which is somewhat unexpected. Those are mentions of historical figures, such as:

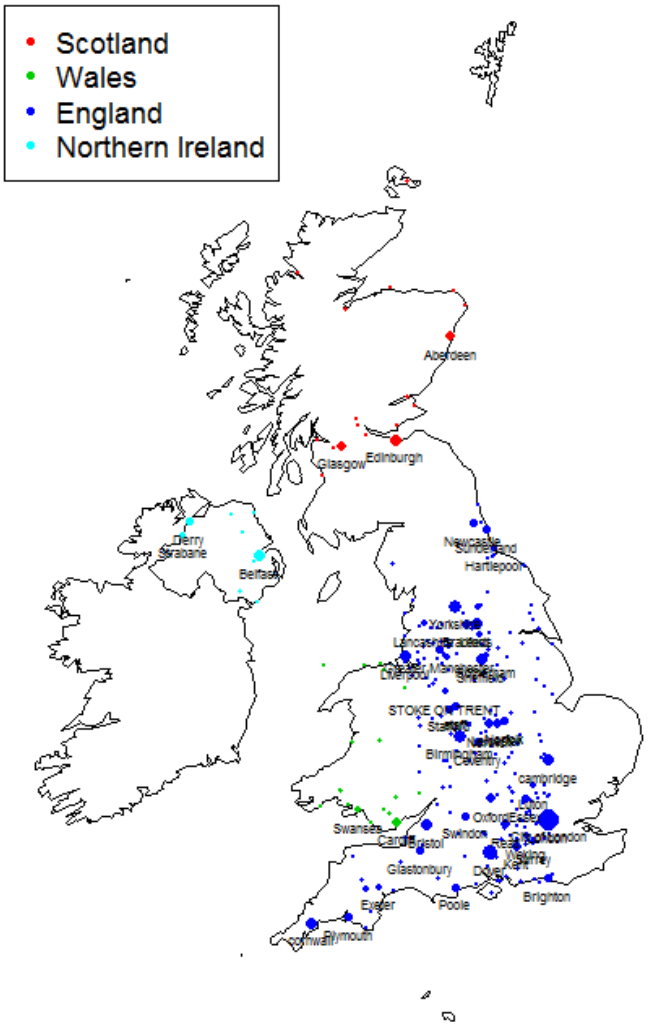

Figure 2.2: Locations mentioned in tweets.

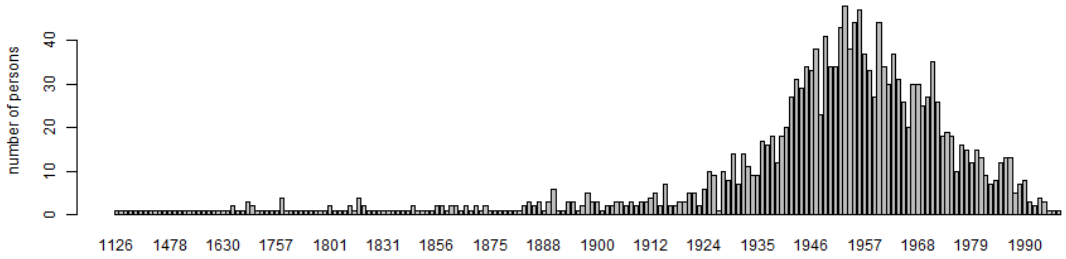

Figure 2.3: Year of birth of people mentioned in the Brexit topic.

Sir Winston Churchill https://en.wikipedia.org/wiki/Winston_Churchill,

Henry VIII https://en.wikipedia.org/wiki/Henry_VIII_of_England,

Charles de Gaulle https://en.wikipedia.org/wiki/Charles_de_Gaulle,

Adam Smith https://en.wikipedia.org/wiki/Adam_Smith,

Adolf Hitler https://en.wikipedia.org/wiki/Adolf_Hitler,

Sir Arthur Harris https://en.wikipedia.org/wiki/Sir_Arthur_Harris,_1st_Baronet,

Ralph Vaughan Williams https://en.wikipedia.org/wiki/Ralph_Vaughan_Williams,

George Santayana https://en.wikipedia.org/wiki/George_Santayana,

Richard III https://en.wikipedia.org/wiki/Richard_III_of_England,

Aldous Huxley https://en.wikipedia.org/wiki/Aldous_Huxley,

Isaac Newton https://en.wikipedia.org/wiki/Isaac_Newton and others (147 mentions in our dataset).

Tweets that mention these historical figures bring up particularly insightful interpretations of the Brexit events, in historical context:

```
Long time since I've seen signs of Combat 18 - 18 for AH
Adolf Hitler - in #Greenwich.  #Brexit has made Nazis cocky
https://t.co/LcofBT5njb
```

```
RT @LiberalLeave:  Benjamin Franklin:  "Those who surrender
freedom for security deserve neither".  Let's vote for our
freedom and democracy #
```

```
RT @DailyAgendaUK: "They sowed the wind and now they are
going to reap the whirlwind" - Sir Arthur Harris #VoteLeave
https://t.co/Ec7M9u8p66
```

```
RT @bridgettcooper:  The arts must rise above the #Brexit
fallout https://t.co/9piMEb9Hul Ralph Vaughan Williams one of
my favorite composer
```

```
"Those who cannot remember the past are condemned to repeat
it." George Santayana #brexit #history #trump #war...
https://t.co/DhSxuqjtHM
```

```
#Brexit will truly re-shape Europe.  Article by @srs2_
Was de Gaulle Right on Britains Role in Europe?
https://t.co/ZmrzjMMrkR
```

```
RT @DailyAgendaUK: "Never in the field of human conflict was
so much owed by so many to so few" - Sir Winston Churchill
#VoteLeave https://
```

```
Time to write the grant application for my comparative study of
Henry VIII's break with Rome and #Brexit https://t.co/2DWFDSA23f
```

```
@cjsnowdon @GarthGodsman What would Isaac Newton, Rutherford,
Darwin and Maxwell have done without the European Union?
#science #brexit
```

As a side comment, there are wrong assignments, too: eg. businessman John Longworth `https://en.wikipedia.org/wiki/John_Longworth_(businessman)` is mistaken by a lawyer that lived during the 19th century `https://en.wikipedia.org/wiki/John_Longworth`.

```
John Longworth - "The moment we leave the repatriation of the
money is immediate, I mean within a year, lets put it that way."
#Brexit
```

2.4.2 Contemporary actors of Brexit

Another insight revealed by the year of birth distribution (Figure 2.3) is the peak around personalities currently aged 65 (born around 1950).

The top most mentioned actors post Brexit, not surprisingly, are (in decreasing order of number of mentions):

Theresa May `https://en.wikipedia.org/wiki/Theresa_May` $(11,452)$,

Boris Johnson `https://en.wikipedia.org/wiki/Boris_Johnson` $(3,961)$,

Nigel Farage `https://en.wikipedia.org/wiki/Nigel_Farage` $(2,926)$,

David Cameron `https://en.wikipedia.org/wiki/David_Cameron` $(2,868)$,

Andrea Leadsom `https://en.wikipedia.org/wiki/Andrea_Leadsom` $(1,388)$,

Angela Merkel `https://en.wikipedia.org/wiki/Angela_Merkel` (922),

Jeremy Corbyn `https://en.wikipedia.org/wiki/Jeremy_Corbyn` (820),

Nicola Sturgeon https://en.wikipedia.org/wiki/Nicola_Sturgeon (780),

Stephen Hawking https://en.wikipedia.org/wiki/Stephen_Hawking (746).

There are surprisingly few younger people. We looked at those mentions that are born after 1975. From them (183 in our dataset), many are famous actors or sporstmen and comments on them are without direct connection to Brexit. We selected a few young politicians and journalists below:

Ruth Davidson https://en.wikipedia.org/wiki/Ruth_Davidson,

Will Straw https://en.wikipedia.org/wiki/Will_Straw,

Paul Nuttall https://en.wikipedia.org/wiki/Paul_Nuttall,

Max Schrems https://en.wikipedia.org/wiki/Max_Schrems,

Tim Stanley https://en.wikipedia.org/wiki/Tim_Stanley,

Tulip Siddiq https://en.wikipedia.org/wiki/Tulip_Siddiq,

Julia Reda https://en.wikipedia.org/wiki/Julia_Reda,

Tom Cotton https://en.wikipedia.org/wiki/Tom_Cotton,

Chuka Umunna https://en.wikipedia.org/wiki/Chuka_Umunna,

Laura Kuenssberg https://en.wikipedia.org/wiki/Laura_Kuenssberg,

Faisal Islam https://en.wikipedia.org/wiki/Faisal_Islam,

Nicola Blackwood https://en.wikipedia.org/wiki/Nicola_Blackwood,

Kezia Dugdale https://en.wikipedia.org/wiki/Kezia_Dugdale,

Jonathan Arnott https://en.wikipedia.org/wiki/Jonathan_Arnott, and others.

2.4.3 Controversiality of most mentioned people

We ran query from Figure 1.4 for the most prominent post-Brexit figures. Table 2.3 summarizes the number of support/ deny and question tweets for each personality, together with an overall controversiality score, which is computed as follows:

$$\text{controversiality}(s,d,q) = 1 - \frac{9}{2}\frac{(s-\frac{1}{3})^2 + (d-\frac{1}{3})^2 + (q-\frac{1}{3})^2}{3},$$

where $s = \frac{\text{\# support tweets}}{\text{\# tweets}}$, $d = \frac{\text{\# deny tweets}}{\text{\# tweets}}$, $q = \frac{\text{\# question tweets}}{\text{\# tweets}}$.

The controversiality score yields a value between 0 and 1, with non-controversial sets of tweets having score closer to 0 and controversial sets having score close to 1.

The intuition behind the score is that a set of tweets (eg. mentioning Theresa May) that have 100% support tweets, 0% deny and 0% questions is non-controversial (there's agreement), thus its score is 0. Same goes for 0% support, 100% deny and 0% questions. On the other hand, a set of tweets that have 33% support, 33% deny and 34% questions are highly controversial (there isn't agreement), so its score is very close to 1.

Name	Support	Deny	Question	Controversiality score
Theresa May	7665 (67%)	2766 (24%)	1021 (9%)	0.73
Boris Johnson	3732 (94%)	109 (3%)	120 (3%)	0.17
Nigel Farage	2839 (97$)	69 (2%)	18 (1%)	0.09
David Cameron	2664 (93%)	150 (5%)	54 (2%)	0.20
Andrea Leadsom	1261 (91%)	102 (7%)	25 (2%)	0.25
Angela Merkel	646 (70%)	11 (1%)	265(29%)	0.64
Jeremy Corbyn	679 (83%)	120 (15%)	21 (3%)	0.44
Nicola Sturgeon	731(94%)	39 (5%)	10 (1%)	0.18
Stephen Hawking	508 (68%)	233(31%)	5 (1%)	0.66

Table 2.3: Number of Support / Deny /Question tweets for the most prominent post-Brexit personalities.

Our results show that discussions involving Theresa May are most controversial (0.73 score) and Nigel Farage the least (0.09 score, as people tend to agree in tweets mentioning him). Naturally, post-Brexit Angela Merkel is subject to many questions, with 29% of the tweets mentioning her being interrogative. And perhaps most surprisingly, Twitter users feel like they know better than Stephen Hawking, 31% of the tweets mentioning him are deny.

In fact, at a closer look, the large number of deny tweets are themselves quotes by Stephen Hawking – a Brexit opposer – shared by Twitter users. The most viral quote is:

```
"Our attitude towards wealth played a crucial role
in Brexit.  We need a rethink" - Stephen Hawking
https://t.co/IA0tr0l8Jm #Brexit #UK
```

2.5 Relevance to PHEME

2.5.1 Relevance to project objectives

We showed that LOD annotations are a powerful tool for semantic enrichment of Social Media microposts, allowing for reasoning with information that is not transmitted directly through the Social Media channels, but available in rich, hopefully unbiased Knowledge Bases. Given the very short text of tweets for example, such enrichment provide with the necessary context, which is crucial for understanding opinions, trends, veracity in Social

Media. LOD enrichment allows for the computer algorithms to 'understand' tweets in a way that a human would, by referring to common knowledge external to the micropost.

The showcase that we chose is relevant for the world political scene at the moment of the deliverable. Post-Brexit discussions on Twitter provide with insights on the mixed feelings, attitude, propaganda, interests that follow the referendum and precede the political actions that need to be taken.

We showed that one can automatically mine the general opinion of the main UK administrative regions and cities. We also identified the main actors of the political scene, with side comments on their age - an aspect that has been so many times brought to the public attention and even used for manipulating the opinion of the voters. Reasoning about age or birth year is only possible via LOD annotations. Controversiality of the main political figures also becomes easily available. A fascinating result is the automatic retrieval of comments mentioning historical figures, which relate the current events to lessons of the past. We can only imagine that historians and journalists would greatly benefit from this collection of quotes and analogies.

2.5.2 Relation to other work packages

The post-Brexit analysis is in fact tightly connected to almost all work packages, as it is an example of exploitation of the entire Pheme infrastructure. It relies on the Ontology modeling developed in WP2, on the algorithms for detecting disputed information from WP4, such as support/deny inference. The real-time streaming and annotation of Brexit-related tweets is possible due to the large efforts for integration of all components into a live pipeline, described in WP6. The selection of the Brexit topic has been done during hackathons and fruitful discussions with the partners involved in the development of the journalism dashboard (WP8).

Bibliography

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165.

Fisher, R. A. (1922). On the Interpretation of 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Lukasik, M., Cohn, T., and Bontcheva, K. (2015). Classifying tweet level judgements of rumours in social media. In *The 2015 Conference on Empirical Methods on Natural Language Processing*.

Toloşi, L. (2016). #brexit twitter analysis. *Whitepaper,* `http://www.ontotext.com/documents/white_papers/brexit-twitter-analysis.pdf`.