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: PHEME Visual Dashboard: Final Version

Deliverable number: D5.2.2

Deliverable date: 31 December 2016

Partners responsible: MODUL University Vienna (MOD)

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: Susan.Fraser@ec.europa.eu	Sent to functional mail box: *CNECT-ICT-*611233*@ec.europa.eu*	On date: 6 January 2017

FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)
Computing Veracity Across Media, Languages, and Social Networks

D5.2.2 PHEME Visual Dashboard – Final Version

Arno Scharl, Tobi Schäfer, Alexander Hubmann-Haidvogel,
Shu Zhu and Tim Lammarsch (MODUL University Vienna)

Abstract

FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)
Deliverable D5.2.2 (WP 5)

This deliverable summarizes work conducted in WP5 of the PHEME project. The goal of T5.4 is to develop a visual analytics dashboard based on a multiple coordinated view approach to explore the veracity intelligence extracted by the content analytics methods from WP2, WP3 and WP4. This requires modular and scalable update mechanisms, advanced query capabilities to reveal supportive and critical voices, charts to track the rise and decay of rumours across media channels and languages, and visual tools to reveal the context and diffusion of emerging rumours.

Keywords: visual analytics, multiple coordinated views, veracity intelligence

Nature: Prototype (P)
Date of Delivery: 06 Jan 2017

Dissemination: PU
Reviewed By: Kalina Bontcheva

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
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: k.Bontcheva@dcs.shef.ac.uk

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

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

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

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

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

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

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

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

Table of Contents

Executive Summary

The software deliverable D5.2.2 summarizes the work conducted in Task T5.4 of the PHEME project, with a focus on the period between July 2015 and December 2016. It represents an updated and extended version of deliverable D5.2.1.

Research and development activities have focused on the Web-based PHEME Visual Dashboard to provide interactive access to a continuously updated news and social media archive of metadata-enriched documents, including a public *Application Programming Interface* (API) for project partners to upload relevant content.

Based on a synchronized ensemble of multiple coordinated views, the dashboard supports the interactive exploration of veracity intelligence extracted by the content analytics methods from WP2, WP3, and WP4 by means of:

- *Advanced query and drill-down capabilities* to reveal supportive and critical voices as well as other metadata dimensions (in D5.2.2, this is conveyed through the colour coding of topics or selected metadata dimensions),
- *Trend charts* to track the rise and decay of rumours across media channels (the deployed prototype includes English-language news media articles and metadata-enriched social media content provided by project partners in T6.1), and
- *Visual tools* to reveal the context and diffusion patterns of emerging rumours (D5.2.2 builds on the portfolio of methods reported in D5.1.2, and also integrates the graph-based network analysis capabilities of D3.3).

Introduction

The overall goal of WP5 is to build the visual analytics tools to interactively explore the veracity intelligence collected in WP6 of PHEME, including visualisations of geospatially and semantically referenced information across news media and social networks. T5.4 integrates these tools and provides a veracity intelligence dashboard with high-performance synchronisation mechanisms to quickly adapt the dashboard configuration according to the specific requirements of a user, and to show context information across the various visualizations.

The dashboard enables gaining insight on popular issues that are being discussed related to the health domain (WP7), with a special focus on *rumours* and *misconceptions*, *mental health* and the *pharmaceutical industry*. The specific context of published documents is being determined based on automatically extracted metadata such as veracity, stance, sentiment, geographic location, or references to other social media authors. This metadata helps us to enrich the display, for example color-coded visualizations that convey these metadata dimensions.

Dashboard Platform Setup and Configuration

Work in T5.4 included the setup and configuration of the PHEME Dashboard using the underlying portal platform. This multi-step process included the following components:

- *Content Feeds*. The PHEME dashboard features data sources enabled by the social media retrieval work from T6.1. Three different additional content feeds are available: a live-updated Twitter feed, a static Twitter dataset around the 2016 Nice attack and a static Twitter dataset around the Germanwings crash in March 2015. All datasets are streamed into the platform through the document API, which allows third-parties to provide both content as well as custom metadata annotations to the platform. All these sources are available in the Data Sources menu of the dashboard so that users can explore the data.
- *Database Configuration*. The dashboard instance needs its own user data schema and corresponding PostgreSQL tables. This is created by importing a general template with customized specifications for the PHEME project. This step includes creating administration accounts.
- *Application Server Configuration*. Each dashboard instance runs on a dedicated application server (*Apache Tomcat*). This includes creating override settings to adjust the PHEME dashboard including (i) title and layout, (ii) PostgreSQL, *ElasticSearch* and other service connection settings, (iii) authentication and user-specific feature configuration, (iv) URL schema adjustments and (v) interface and visualization component configuration.
- *Indexer Setup*. While the application server hosts the dashboard interface itself, it is necessary to configure an additional indexing service for initial batch operations and ongoing data ingestion and indexing.
- *Deployment*. This step concludes the setup with deploying the configured components to the server infrastructure. For incremental feature updates, this step needs to be repeated.

Content Customization and Disambiguation

Selecting relevant news and social media content for the PHEME dashboard, in line with the use case requirements, required effective methods for content filtering and disambiguation. We used a domain-specificity measure based with a combination of blacklists and whitelists to assess the relevance of gathered news media articles in the context of mental health and related medical issues, in addition to the domain-specific social media feeds provided by project partners.

Mental Health Disorders

Many mental health disorders are associated with stigma – a sign of discredit, which sets someone apart from others. Social media play a significant role in diffusing inaccurate portrayals of mental illness in the form of rumours and misinformation. Thereby, these channels perpetuate stereotypes and promote the discrimination against mental health disorders. Failure to identify and deal with this stigma – societal or self-directed – may hinder prospects of recovery and rehabilitation. To customize the dashboard's list of predefined topics, we used the ten mental health disorders including associated terms listed in Table 1, as well as a set of mental health/stigma-related hashtags.

Table 1. Search terms and related keywords for disorders/stigma search

Disorders	**Keywords**
Attention deficit disorder	attention deficit, attentiondeficit, adhd
Bipolar affective disorder	bipolar*, manic depress*, manicdepress*
Psychotic disorder	psychos*, psychot*
Schizophrenia	schizophren*, schizoaffective, schizo affective
Depressive disorder	depress*
Obsessive-compulsive disorder	obsessive compulsive, obsessivecompulsive, ocd
Autism spectrum disorders	*autism*
Alzheimer's disease	alzheimer*
Dementia	dementia
Anxiety disorder	*anxiety*
Mental health / stigma	#mentalhealth*, #mentalillness, #endstigma, #stigma, #whatstigma, #mhstigma, #stigmahurts

Medications

In addition to the mental health orders, the dashboard has been configured to track medications used in mental healthcare, including 30 generic names and associated brand names. The list was developed in three phases:

(i) the initial consideration set was composed of the most commonly referenced medications in the mental healthcare clinical record by number of mentions and number of unique patient records within which they were documented – the 18 most common medications were selected;

(ii) by inspecting the number of individual drug references from 2009 to 2014, we could establish which medications had seen the most increase in mentions – seven medications were selected;

(iii) after consultation with a senior pharmacist, the five newest psychotropic medications were added, irrespective of their mentions in clinical records. All 30 medications were then categorized by type according to the *British National Formulary* (Joint Formulary Committee, 2014).

Table 2. Generic drugs, respective brand names and type of medication

Generic	**Brand**	**Type**
clozapine	Clozaril, FazaClo, Versacloz, Clopine, Zaponex	Antipsychotic
olanzapine	Zyprexa, Zypadhera, Lanzek	Antipsychotic
methadone	Symoron, Dolophine, Amidone, Methadose, Physeptone, Heptadon	Opioid analgesic
risperidone	Risperdal	Antipsychotic
citalopram	Celexa, Cipramil	Antidepressant
diazepam	Valium, Diastat, Diastat AcuDial, Zetran	Anxiolytic
zopiclone	Imovane, Zimovane	Hypnotic
promethazine	Phenergan, Promethegan, Romergan, Fargan, Farganesse, Prothiazine, Avomine, Atosil, Receptozine, Lergigan, Sominex	Hypnotic
valproate	Epilim, Episenta, Epival, Convulex	Drug for mania and hypomania
quetiapine	Seroquel	Antipsychotic
lorazepam	Ativan	Anxiolytic
lithium	Eskalith, Lithobid, Cibalith S	Drug for mania and hypomania
aripiprazole	Abilify	Antipsychotic
mirtazapine	Avanza, Axit, Mirtaz, Mirtazon, Remeron, Zispin	Antidepressant
fluoxetine	Prozac, Sarafem	Antidepressant
sertraline	Zoloft, Lustral	Antidepressant
venlafaxine	Effexor	Antidepressant
haloperidol	Haldol	Antipsychotic
amisulpride	Amazeo, Amipride, Amival, Solian, Soltus, Sulpitac, Sulprix	Antipsychotic
paliperidone	Invega	Antipsychotic
buprenorphine	Subutex, Butrans, Buprenex	Opioid analgesic
donepezil	Aricept	Drug for dementia
memantine	Axura, Akatinol, Namenda, Ebixa, Abixa, Memox	Drug for dementia
rivastigmine	Exelon	Drug for dementia
galantamine	Nivalin, Razadyne, Reminyl, Lycoremine	Drug for dementia
lurasidone	Latuda	Antipsychotic
brexpiprazole	OPC34712, OPC 34712	Antipsychotic
pimavanserin	ACP-103, ACP 103, ACP103, Nuplazid	Antipsychotic
vortioxetine	Brintellix, Trintellix	Antidepressant
cariprazine	RGH188, RGH 188	Antidepressant

Red – medications most commonly mentioned on the mental health record
Green – medications with the most rapid increase in mentions on the mental health record
Blue – most recently introduced mental health medication

Pharmaceutical Industry

To represent the major players of the pharmaceutical industries, the top 10 companies in terms of revenues (sales of prescription medicines, including generics drugs)[1] have been included in the dashboard: *Novartis, Pfizer, Roche, Sanofi, Merck & Co., Johnson & Johnson, GlaxoSmithKline, AstraZeneca, Gilead Sciences, and Takeda.*

To better understand the perceptions of pharmaceutical brands, the dashboard includes topic definition for the well-established "Dimensions of Brand Personality" by Aaker (1997): *Competence, Excitement, Ruggedness, Sincerity,* and *Sophistication*. Using the dashboard's built-in radar chart, analysts can investigate the relative performance along each individual dimension. Since the radar chart supports ad hoc data exploration and topic definition, it is not restricted to brands, but can also be applied to specific products (e.g., different antidepressants) or persons (e.g., the company's managing director).

Dashboard Overview

Rather than relying on simple statistical representations, the visual analytics dashboard supports the real-time synchronization of multiple coordinated views (Hubmann-Haidvogel et al., 2009) to convey context information along various semantic dimensions. When properly disambiguated, such context information helps to investigate the veracity of emerging stories. The dashboard distinguishes three types of context:

1. *lexical context* – specific vocabulary and sequence of words that precede or follow a statement (Fischl and Scharl, 2014; Wattenberg and Viégas, 2008);
2. *geospatial context* – the author's location and the geospatial references contained in a document (Niepold et al., 2008; Scharl and Tochtermann, 2007);
3. *relational context* – frequency distribution of named entities in the content repository, and co-occurrence patterns among these entities (Derczynski et al., 2015; Weichselbraun et al., 2015).

Shown in Figure 1, the dashboard provides visual means to analyse the content repository and the extracted metadata (veracity, stance, sentiment, etc.). It is divided into six main content areas, briefly summarized in the following sections – for additional details about these interface elements, please refer to the online documentation.[2]

Source and Configuration Management

The upper menu provides temporal controls, a selector for the sources to be analysed (news media, social media, etc.), and access to a sidebar for exporting search results and other datasets in various formats.

Topic Management

The upper left window of the dashboard contains the topic management section, which provides one-click access to the topics defined in the *Content Customization and Disambiguation* section above. Clicking on a topic management element triggers one of

[1] www.pmlive.com/top_pharma_list/global_revenues

[2] www.weblyzard.com/interface

the following actions: (i) *topic labels* trigger a full-text search based on the topic's specific configuration; (ii) *topic markers* (= small rectangles) select topics to be shown in the charts; (iii) the *settings symbol* allows to configure or delete topics. This section can be switched to the drill-down view to explore custom metadata dimensions.

Trend Charts

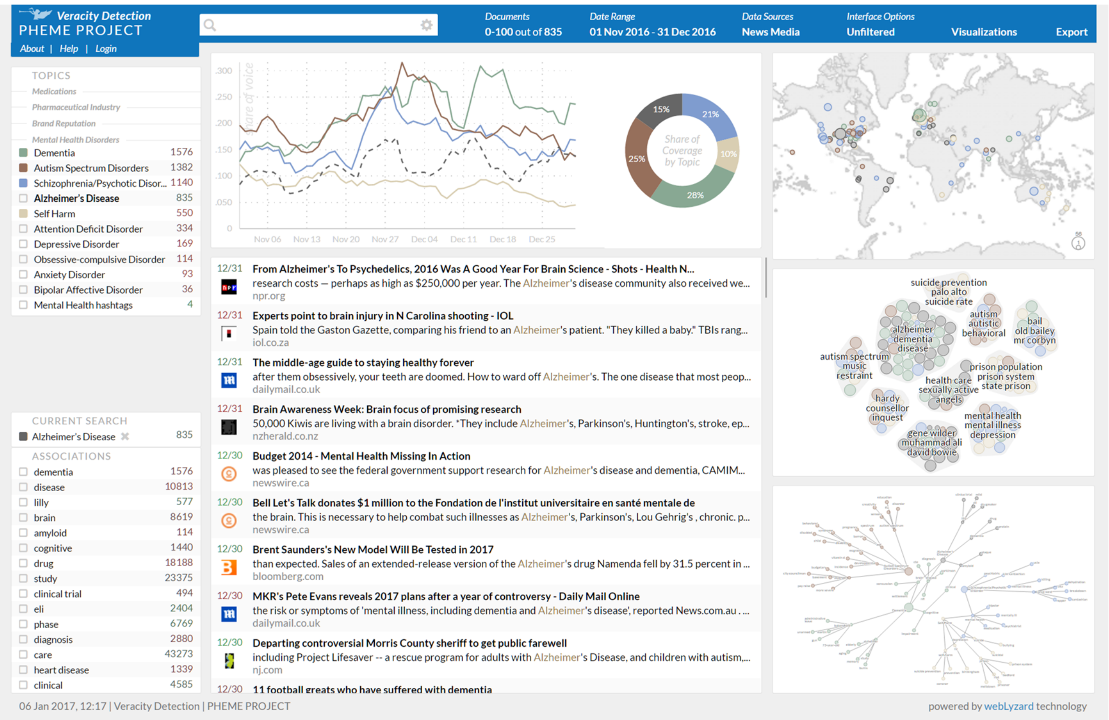
A trend chart with optional donut chart shows one of the following times series, selectable via a floating menu: (i) the *share of voice*, a comparative measure of attention based on the relative number of mentions, (ii) the *frequency* of selected topics in the specified time interval, (iii) the average *sentiment* regarding these topics, or (iv) the level of *disagreement* (standard deviation of sentiment). The charts support either the display of the raw values or can be switched to running averages of 7, 30 or 60 days.

Content View

The floating menu provides seven different representations to show the search results: documents, sentences, word tree, entities, relation tracker, sources, source map. Interactive controls in the 'document' and 'sentence' views include: (i) mouse-over to preview documents; (ii) click to select a document and shows its content in extended form; and (iii) second click to switch to full text view, which reveals the document's annotations including veracity, stance, sentiment, keywords, and location.

Semantic Associations

The lower left view displays a list of associations with the search term, based on the selected source and data range. The means to analyse and visualise these associations, as well as other metadata patterns (right sidebar), will be presented in the next sections.



TOPICS

Medications

Pharmaceutical Industry

Brand Reputation

Mental Health Disorders

🟩 Dementia	1576
🟫 Autism Spectrum Disorders	1382
🟦 Schizophrenia/Psychotic Disor...	1140
□ **Alzheimer's Disease**	835
🟨 Self Harm	550
□ Attention Deficit Disorder	334
□ Depressive Disorder	169
□ Obsessive-compulsive Disorder	114
□ Anxiety Disorder	93
□ Bipolar Affective Disorder	36
□ Mental Health hashtags	4

CURRENT SEARCH

■ Alzheimer's Disease ✖	835

ASSOCIATIONS

□ dementia	1576
□ disease	10813
□ lilly	577
□ brain	8619
□ amyloid	114
□ cognitive	1440
□ drug	18188
□ study	23375
□ clinical trial	494
□ eli	2404
□ phase	6769
□ diagnosis	2880
□ care	43273
□ heart disease	1339
□ clinical	4585

12/31 **From Alzheimer's To Psychedelics, 2016 Was A Good Year For Brain Science - Shots - Health N...**
research costs — perhaps as high as $250,000 per year. The Alzheimer's disease community also received we...
npr.org

12/31 **Experts point to brain injury in N Carolina shooting - IOL**
Spain told the Gaston Gazette, comparing his friend to an Alzheimer's patient. "They killed a baby." TBIs rang...
iol.co.za

12/31 **The middle-age guide to staying healthy forever**
after them obsessively, your teeth are doomed. How to ward off Alzheimer's. The one disease that most peop...
dailymail.co.uk

12/31 **Brain Awareness Week: Brain focus of promising research**
50,000 Kiwis are living with a brain disorder. *They include Alzheimer's, Parkinson's, Huntington's, stroke, ep...
nzherald.co.nz

12/30 **Budget 2014 - Mental Health Missing In Action**
was pleased to see the federal government support research for Alzheimer's disease and dementia, CAMIM...
newswire.ca

12/30 **Bell Let's Talk donates $1 million to the Fondation de l'institut universitaire en santé mentale de**
the brain. This is necessary to help combat such illnesses as Alzheimer's, Parkinson's, Lou Gehrig's , chronic. p...
newswire.ca

12/30 **Brent Saunders's New Model Will Be Tested in 2017**
than expected. Sales of an extended-release version of the Alzheimer's drug Namenda fell by 31.5 percent in ...
bloomberg.com

12/30 **MKR's Pete Evans reveals 2017 plans after a year of controversy - Daily Mail Online**
the risk or symptoms of 'mental illness, including dementia and Alzheimer's disease', reported News.com.au . ...
dailymail.co.uk

12/30 **Departing controversial Morris County sheriff to get public farewell**
including Project Lifesaver -- a rescue program for adults with Alzheimer's Disease, and children with autism,...
nj.com

12/30 **11 football greats who have suffered with dementia**

06 Jan 2017, 12:17 | Veracity Detection | PHEME PROJECT

powered by webLyzard technology

Figure 1. Screenshot of the PHEME Dashboard with online coverage about *Alzheimer's Disease* between November and December 2016

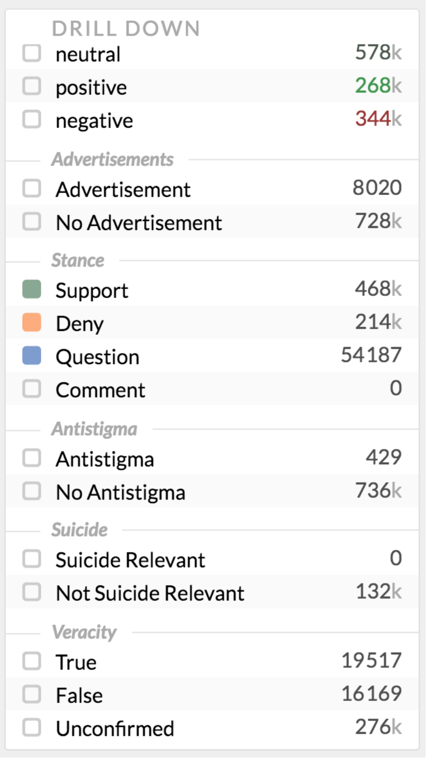
Metadata Integration

Drill Down Sidebar

The primary interface to select metadata attributes makes use of the dashboard's search drill down component. This component enables the selection of distinct metadata attributes for further analysis with visualization components within the dashboard. The drilldown sidebar has the additional advantage of clearly separating attribute values, e.g. *support, deny, question* and *comment* in the case of *stance*.

The drill down sidebar is enabled by using the menu in the topic sidebar header to switch between different sidebar types. This menu consists of metadata categories relating to the current search. It allows using the trend chart to compare elements within a metadata category. In addition to the generic attributes *language, source type* and *sentiment*, the PHEME dashboard supports the metadata attributes *veracity, stance, stigma,* and *advertisements*. The content streams provided through the document API are pre-annotated along these dimensions:

(i) Boolean flags for advertisements, anti-stigma and suicide relevant documents,
(ii) a stance classification (support, deny, comment, question),
(iii) a veracity score, and
(iv) a cluster identifier.



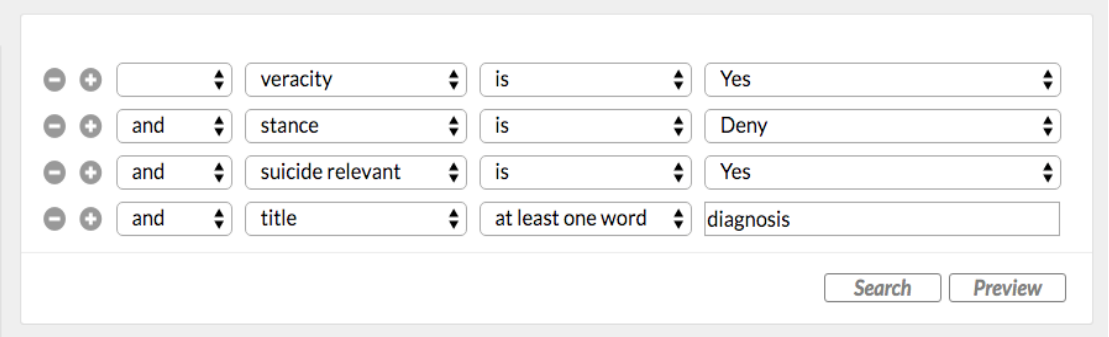
DRILL DOWN	
□ neutral	578k
□ positive	268k
□ negative	344k
Advertisements	
□ Advertisement	8020
□ No Advertisement	728k
Stance	
🟩 Support	468k
🟧 Deny	214k
🟦 Question	54187
□ Comment	0
Antistigma	
□ Antistigma	429
□ No Antistigma	736k
Suicide	
□ Suicide Relevant	0
□ Not Suicide Relevant	132k
Veracity	
□ True	19517
□ False	16169
□ Unconfirmed	276k

Figure 2. The search drill down menu including newly introduced metadata attributes

These metadata annotations are selectable in the drill-down sidebar, allowing users to visualize the number of documents tagged as e.g. denying a claim (*stance information*) or being suicide-relevant (*domain specifity*).

Figure 2 shows a screenshot of the search drill down menu. The various metadata attributes are organized in customized categories. The most left column is used to select individual attributes for further analysis. On selection, a colour from a predefined palette gets dynamically assigned to the attribute which is then used in other dashboard components to identify the attribute. A click on the attribute name itself triggers a new search to only show documents that comply with the clicked attribute (e.g., only documents that were tagged to be *false*).

The right column shows frequency values in context of the current search filters like the time span and content sources, additionally colour-encoded with sentiment values. On hovering an attribute, the frequency value changes to a gear icon. This can be clicked to open a contextual menu with specific options for each attribute.

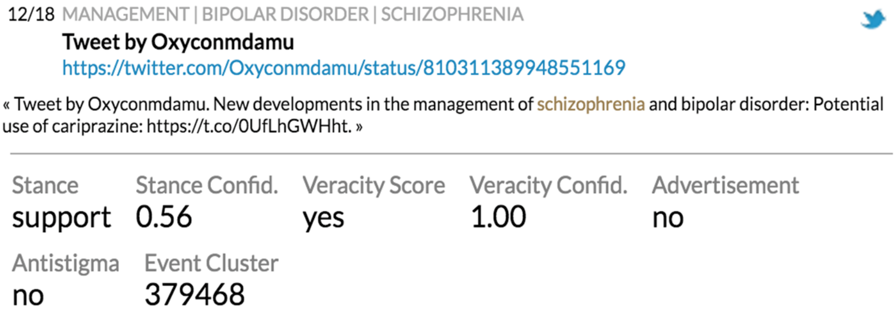


	veracity	is	Yes
and	stance	is	Deny
and	suicide relevant	is	Yes
and	title	at least one word	diagnosis

Figure 3. The advanced search dialog including selected options to filter by specific metadata attributes and a search for "diagnosis" in the title field

Advanced Search

For further analysis, as shown in Figure 3, metadata annotations are also available in both the advanced search as well as user-defined topics (e.g. to search for all documents tagged with a "deny" stance which are suicide-relevant and have a confirmed veracity status). Figure 4 shows the content view's support for the detailed display of metadata attributes for individual documents including veracity scores within search results.



12/18 MANAGEMENT | BIPOLAR DISORDER | SCHIZOPHRENIA

Tweet by Oxyconmdamu

https://twitter.com/Oxyconmdamu/status/810311389948551169

« Tweet by Oxyconmdamu. New developments in the management of schizophrenia and bipolar disorder: Potential use of cariprazine: https://t.co/0UfLhGWHht. »

Stance	Stance Confid.	Veracity Score	Veracity Confid.	Advertisement
support	0.56	yes	1.00	no

Antistigma	Event Cluster
no	379468

Figure 4. Detailed document view of a tweet including metadata

Component Term Filtering for Keyword Labels

Improved filtering algorithms avoid the redundant listing of the search term itself or its component words as a label in the *association list*, the *keyword graph*, the *tag cloud*, or the *cluster map*. Additionally, component words are combined if a term would be listed as part of a longer n-gram and the difference in frequency of the two associations is below 10% - i.e., this combines "mental" and "mental health" to "mental health" if both associations appear equally often (within a 10% threshold).

Analytics Components and Visualization Tools

The PHEME dashboard integrates a number of analytics components and visualization tools to convey the documents' geospatial, semantic and temporal context. The components are synchronized and the visualization in the right column of the dashboard can be re-positioned using drag-and-drop operations. All the visual tools of the dashboard are based on the *D3.js* library (Bostock et al., 2011), the graph-based representations use the *Graphyte* library reported in D5.1.2, which has been released under an BSD (Berkeley Software Distribution) open source license on github.com.[3]

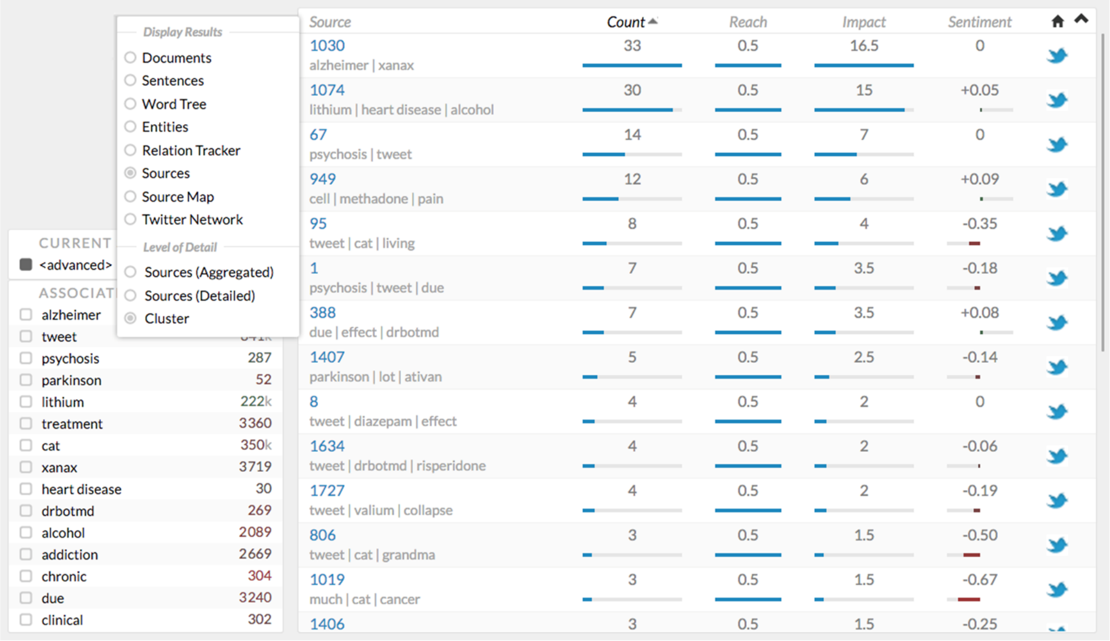
Considering the research goals of the PHEME project in general and the use case requirements in particular, special emphasis has been placed on the analysis and visualization of document clusters to represent emerging stories *(“Cluster Map”),* and graph-based visualizations of topics associated with the search query *(“Keyword Graph”).*

Document Clustering

The PHEME dashboard features several possibilities to analyse search results with clustering techniques. This enables the user to explore and understand large search results without the need to skim through endless lists of single documents.

Source List

The *source list* component has been extended to allow aggregation on a pre-annotated event cluster level in addition to the existing "aggregated" and "detailed" views, shown in Figure 5.



Display Results

○ Documents
○ Sentences
○ Word Tree
○ Entities
○ Relation Tracker
● Sources
○ Source Map
○ Twitter Network

Level of Detail

○ Sources (Aggregated)
○ Sources (Detailed)
● Cluster

CURRENT

■ <advanced>

ASSOCIAT	
□ alzheimer	
□ tweet	
□ psychosis	287
□ parkinson	52
□ lithium	222k
□ treatment	3360
□ cat	350k
□ xanax	3719
□ heart disease	30
□ drbotmd	269
□ alcohol	2089
□ addiction	2669
□ chronic	304
□ due	3240
□ clinical	302

Source	Count ▲	Reach	Impact	Sentiment
1030 alzheimer | xanax	33	0.5	16.5	0
1074 lithium | heart disease | alcohol	30	0.5	15	+0.05
67 psychosis | tweet	14	0.5	7	0
949 cell | methadone | pain	12	0.5	6	+0.09
95 tweet | cat | living	8	0.5	4	-0.35
1 psychosis | tweet | due	7	0.5	3.5	-0.18
388 due | effect | drbotmd	7	0.5	3.5	+0.08
1407 parkinson | lot | ativan	5	0.5	2.5	-0.14
8 tweet | diazepam | effect	4	0.5	2	0
1634 tweet | drbotmd | risperidone	4	0.5	2	-0.06
1727 tweet | valium | collapse	4	0.5	2	-0.19
806 tweet | cat | grandma	3	0.5	1.5	-0.50
1019 much | cat | cancer	3	0.5	1.5	-0.67
1406	3	0.5	1.5	-0.25

Figure 5. Screenshot of the source list with the newly introduced capability to aggregate results by cluster annotation

The first column features the cluster id as well as up to three significant keywords representing the clusters. The list can be sorted by name, frequency of mentions (count), reach based on the Alexa traffic rank, impact (frequency multiplied by impact), or average sentiment. Clicking on a cluster name triggers a search for documents that belong to this cluster, and list all these documents.

[3] www.github.com/weblyzard/graphyte

Source Map

The *source map* is an interactive scatter plot that translates the table of the *source list* into visual form. The version used in the PHEME Visual Dashboard is capable of visualising the pre-annotated event clusters. The data points of the scatter plot reflect the following source characteristics: (i) topic frequency (horizontal axis), (ii) average sentiment towards the topic (vertical axis, color-coding), and (iii) reach of the source (size of the marker). Whenever a user triggers a new search, animated transitions show the resulting changes in the data point distribution.

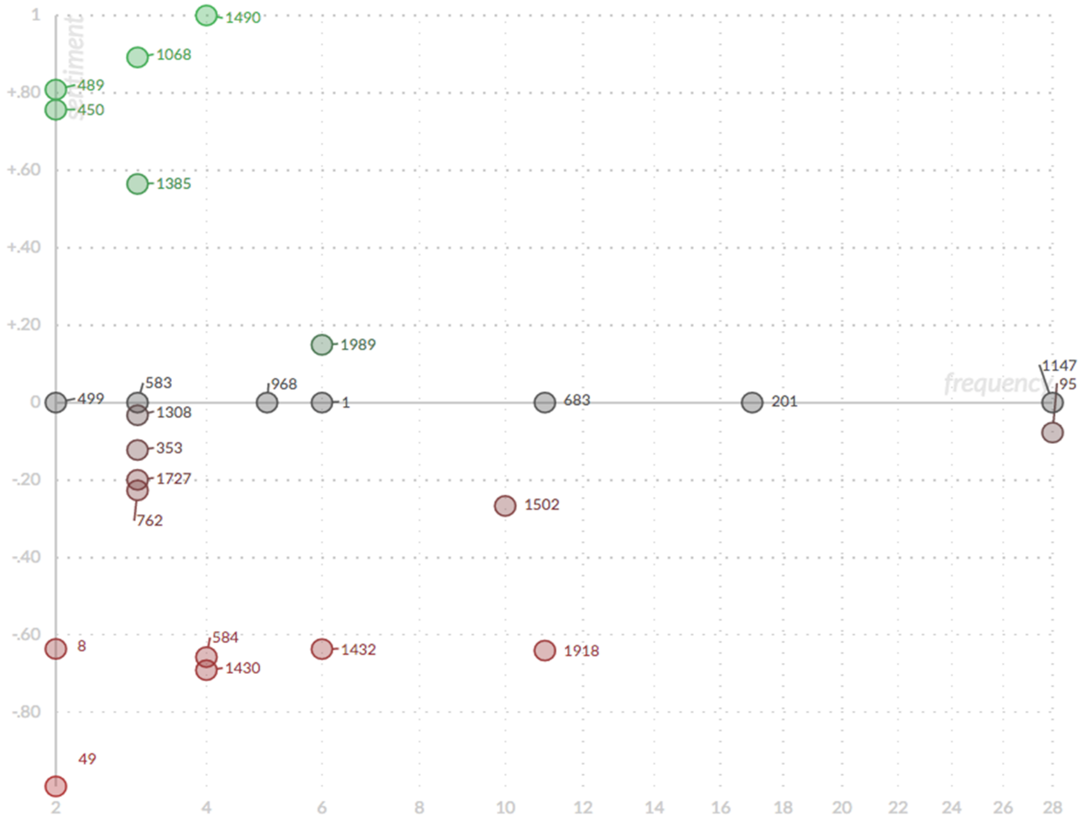


Figure 6. The *source map* is a scatterplot and enables the visualization of cluster distribution across the frequency and sentiment axis

Cluster Map

The *cluster map* uses a variation of force-directed placement and node grouping to arrange search results by topic, visualizing documents by their semantic similarity using three alternative methods: (i) *K-means*, which divides the collection of documents into a fixed amount of clusters where each document belongs to the cluster with the nearest centroid, (ii) *agglomerative hierarchical clustering*, a deterministic approach that pairs clusters into a tree-like structure and (iii) *Louvain graph community detection* (Blondel et al., 2008). The last approach aims to partition a graph of document to keyword relations into subsets which share more edges connecting the subset compared to others. The Louvain algorithm separates the network in communities by optimizing greedily a modularity score after trying various grouping operations on the network. By using this simple greedy approach the algorithm is computationally very efficient.

The *cluster map* groups similar documents by a convex hull shape that visually holds its nodes together. The size of this shape is dynamic and depends on the number of contained nodes. Each node represents a document returned by the search function, shown as a circle shape of variable size and colour. While node size is proportional to the reach of the document's media source, node colour reflects normalized document sentiment – ranging from green (positive) to grey (neutral) and red (negative). Sentiment is shown with variable saturation, depending on the degree of polarity – vivid colours indicate emotional articles or postings, and lower saturation a more factual online coverage.

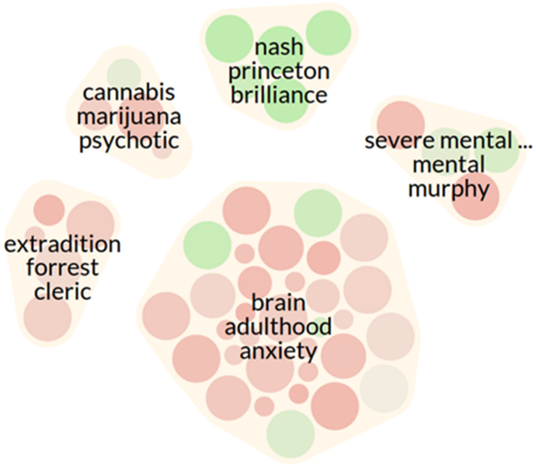


Figure 7. Document clusters based on the search query "schizophrenia" (using the December 2015 version of the *cluster map* component)

Three keywords per cluster are used as a label to describe its contents. The hull shapes of nodes and clusters are rendered with reduced opacity to decrease the visual load and increase the readability of cluster labels. These labels are extracted from the ordered list of all document keywords within the cluster, considering the reach of the documents' sources. While Figure 7 shows the previous version of the *cluster map* reported in D5.2.1, Figure 8 and Figure 9 show screenshots of the latest *cluster map* version using colour coding to classify documents by *topic* (in this case, different mental health disorders) or by *stance* (i.e., support, deny, question). In addition to the ability to colour code by different metadata attributes, the underlying clustering technique, label detection as well as the design have been improved as well:

- *Clustering.* Complementing the previously available methods, the *Louvain* community detection can be used as an option to detect clusters. For this approach to work, internally a graph of document to keyword relations is created. Compared to K-means, the results are more stable, and the algorithm is more performant because only one iteration is necessary.

  Additionally, the clustering framework has been extended to be able to consider additional dimensions besides document keywords: The specific metadata available in the PHEME dashboard like Stance or Stigma can be used to be considered by the cluster detection. While these options can be configured by an administrator to adapt the dashboard's settings, we avoid exposing these technical configuration options to end users.

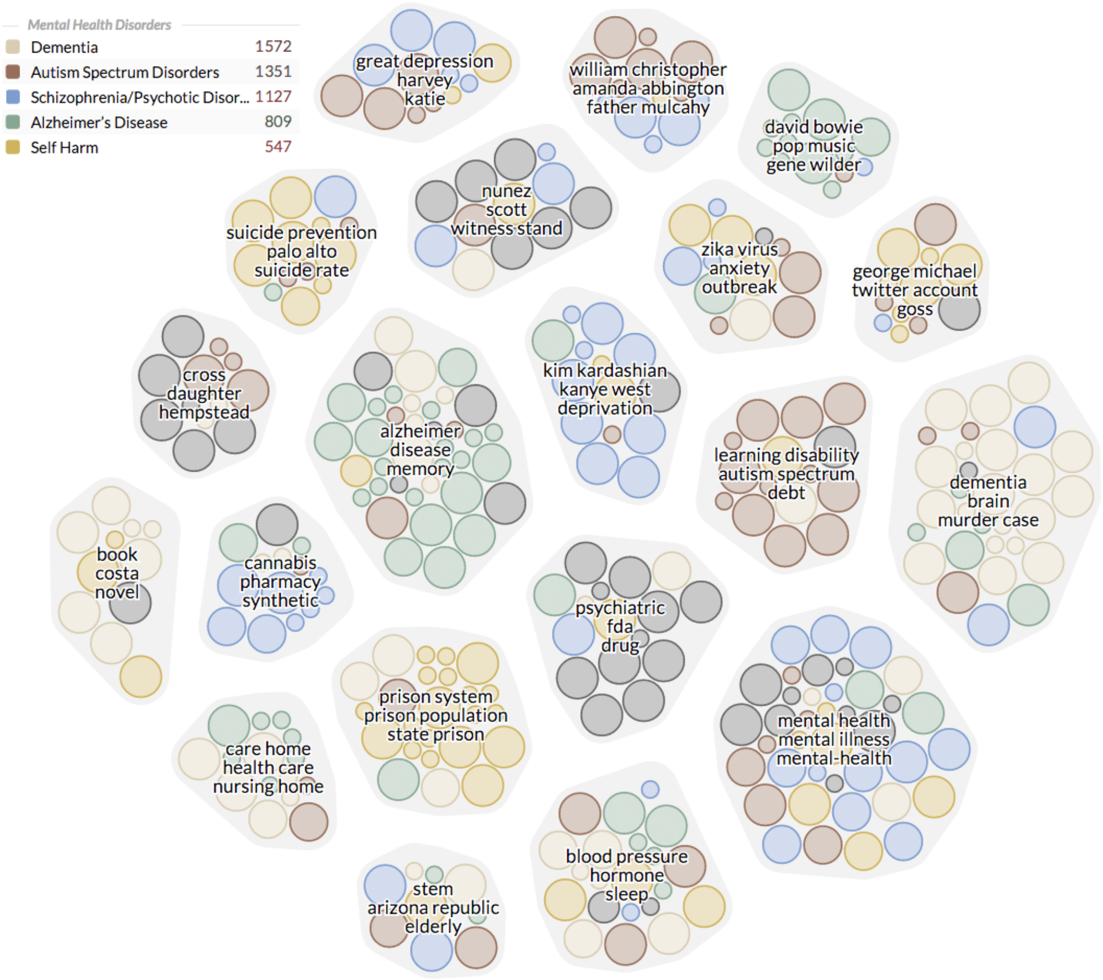
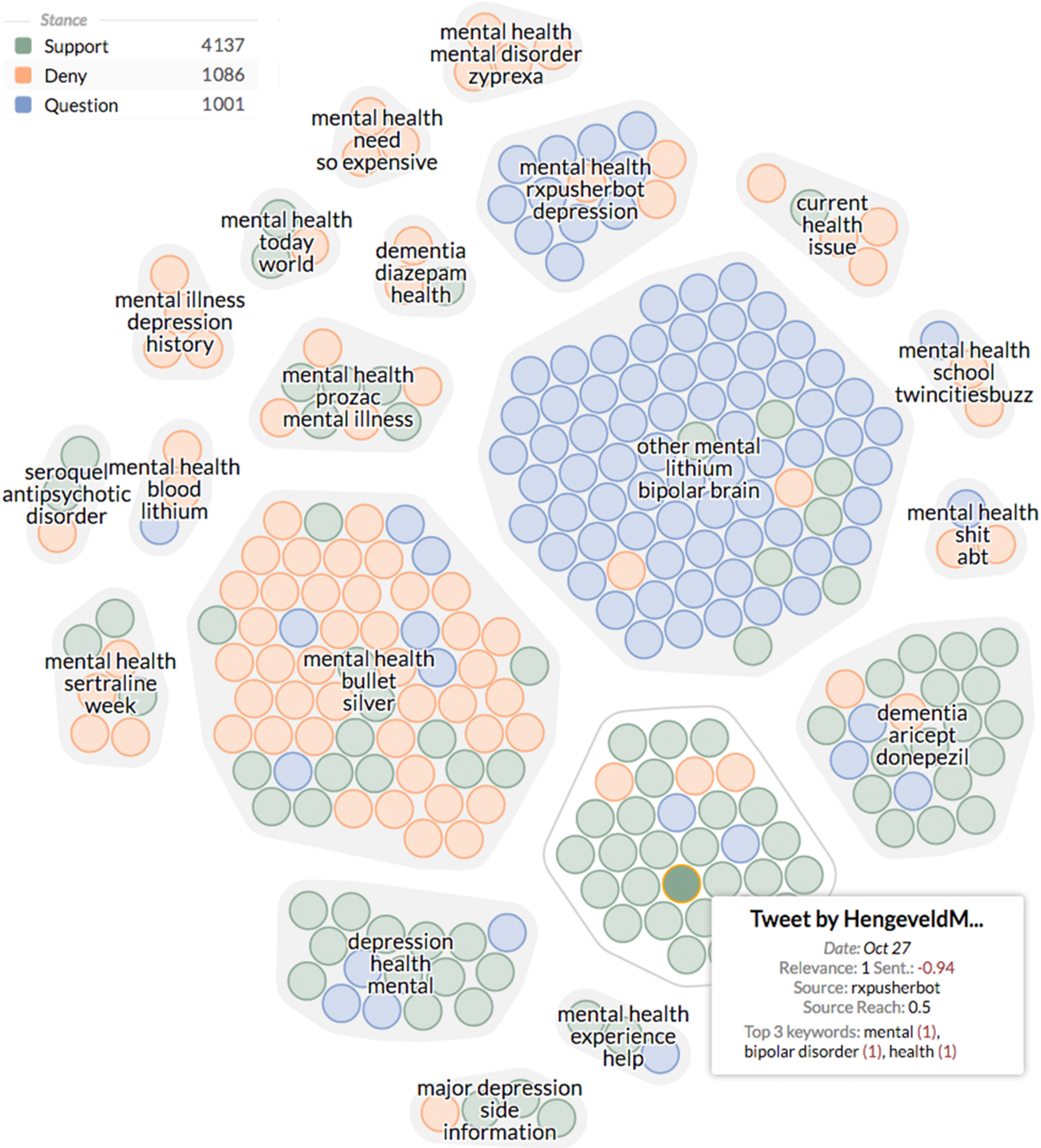


Figure 8. Revised *cluster map* representation, using the final version of the component to classify documents by topic

- *Cluster Labels.* The labelling technique has been improved: The version reported in D5.2.1 used aggregations of document keywords to create cluster labels. Because this didn't consider semantic relations between keywords, redundant labels or partial duplicates could occur. The updated technique in D5.2.2 solves these issues by matching significant document keywords to distinct entity annotations. The results especially for compound words are greatly improved by this as the labels shown in Figure 8 demonstrate.

- *Design.* Each cluster's hull color has been changed from yellow to gray to put more emphasis on the sentiment encoded document nodes and avoid skewing color perception of these sentiment values. The labels feature a white outline to improve readability. Additionally, the labels font size gets adjusted dynamically depending on the dynamic zoom level of the whole visualization. Hovering a document cluster hides the keywords and allows the user to focus on individual nodes. Hovering a node shows a tooltip with extended information on the document, as shown in Figure 9.



Stance	
🟩 Support	4137
🟧 Deny	1086
🟦 Question	1001

Tweet by HengeveldM...
Date: Oct 27
Relevance: 1 Sent.: -0.94
Source: rxpusherbot
Source Reach: 0.5
Top 3 keywords: mental (1), bipolar disorder (1), health (1)

mental health mental disorder zyprexa
mental health need so expensive
mental health rxpusherbot depression
current health issue
mental health today world
dementia diazepam health
mental illness depression history
mental health prozac mental illness
other mental lithium bipolar brain
mental health school twincitiesbuzz
seroquel antipsychotic disorder
mental health blood lithium
mental health shit abt
mental health sertraline week
mental health bullet silver
dementia aricept donepezil
depression health mental
mental health experience help
major depression side information

Figure 9. *Cluster map* with colour coding to *convey* stance information (support, deny, question), with activated tooltip to inspect additional metadata

Graph Visualisations

Within PHEME, the *Graphyte* library reported in D5.1.2 plays a key role in the rendering of (i) associated topics as part of a *Keyword Graph,* and (ii) the structure of Twitter communication as part of the *Social Network Analysis,* which depicts information flows surrounding a specific topic (as reported in D3.3 *Algorithms for Implicit Information Diffusion Network).*

Keyword Graph

Moving beyond document clusters as described in the previous sections, the second application of the force-directed layout algorithm of D5.1.2 focuses on the visualization of graphs as a flexible method to show semantic associations among extracted topics.

The computation of associations considers the selected source(s) and time interval. Changing these settings, therefore, triggers an immediate update of the graph (which can also be enlarged to full screen size).

To visualize the relations between the strongest associations (keywords) detected in a search query, an undirected graph is computed. This graph consists of vertices (nodes) representing the keywords, and edges (connections) representing the relations between the nodes. The version reported in D5.2.1 supported visualization of the current search only: At the graph's centre resides the original term of the query – i.e., the root node for which the associations were evaluated for.

D5.2.2 introduces support for multiple seed nodes and therefore enables detailed analysis of topic relations. The keyword graph considers the same topics or metadata dimensions a user has selected for the trend charts. Overall, this provides a consistent data display across different visualization components. When only one topic or metadata dimension is selected in the sidebar, the single seed node is shown in the centre of the graph. Outgoing nodes are colour coded by an aggregated sentiment value in context of the seed nodes search and filtering parameters as shown in Figure 10.

The size of a node reflects the term's frequency of occurrence, while its colour represents its average sentiment. Each node is labelled with its corresponding term. The lengths of the edges (i.e., the distances between nodes) are dynamically calculated by the *Graphyte* algorithms reported in D5.1.2.

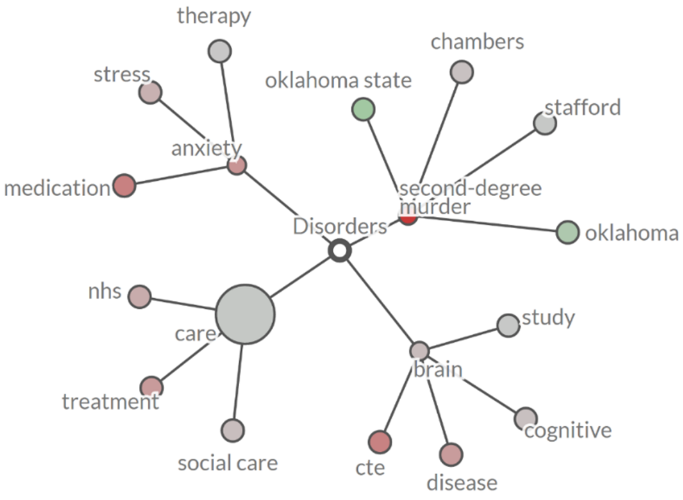


Figure 10. Associated keywords based on the search query "mental disorders"′

When multiple elements are selected in the sidebar, the *Keyword Graph* arranges seed nodes with fixed positions in a circular layout. Outgoing and interconnecting nodes of each seed node are then positioned using *Graphyte's* layout techniques. Figure 11 shows an instance of the *Keyword Graph* showing relation between multiple mental health related topics.

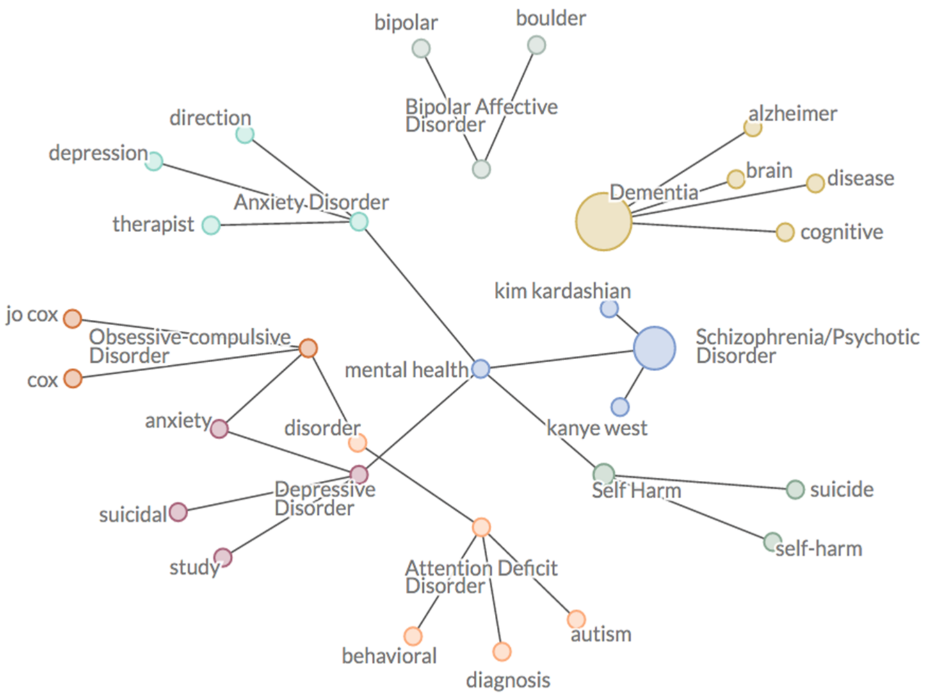


Figure 11. *Keyword Graph* with multiple root nodes to show relations between all the selected items of the topic management section

Note that for example the node "mental health" is not part of the seed nodes: It is among the most significant keywords across topics and therefore establishes connections between several topics. When using the drill-down menu, the *Keyword Graph* can also be used to examine the relations of metadata attributes such as *stance* or *veracity* within a single search topic, as shown in Figure 12.

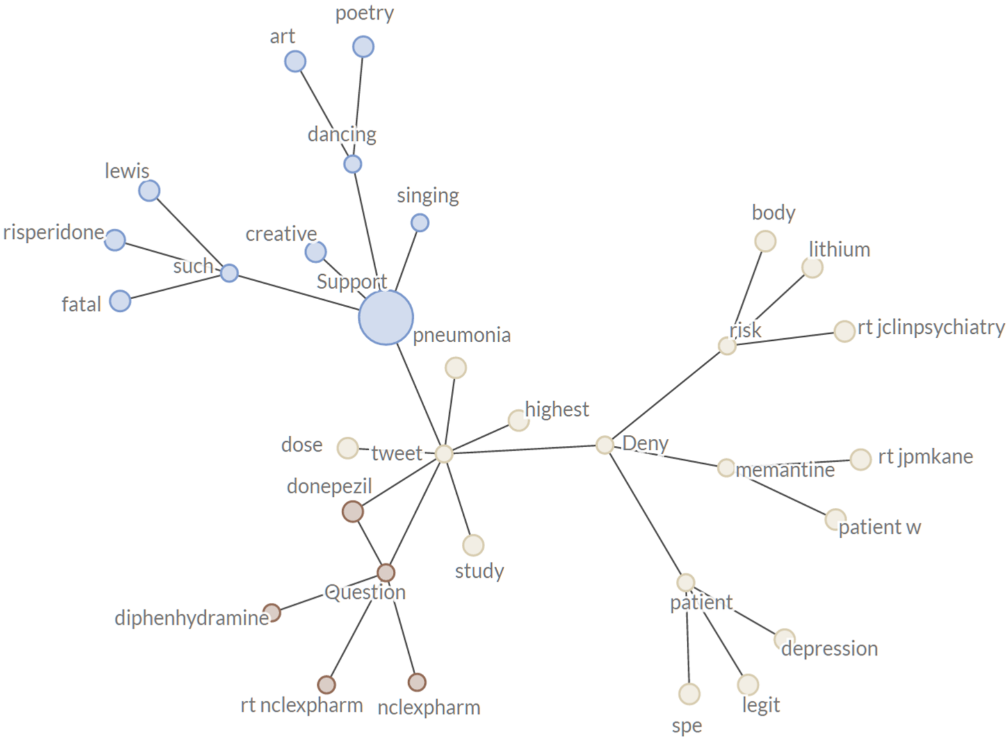


Figure 12. Keyword Graph using metadata attributes as seed nodes (stance information 'support' vs. 'deny' vs. 'question')

Social Network Analysis

While the initial integrated version reported in D5.2.1 supported a single root node to show related keywords in the *Keyword Graph*, D5.2.2 adds support for multiple seed nodes as well as graphs without particular seed nodes. This feature is used in the *Social Network Analysis* component. The visual display itself in each component is handled by the *Graphyte* library. The features described in this deliverable rely on improvements to the data aggregation and transformation capabilities of the PHEME dashboard platform developed in months 24-36 of the PHEME projects.

The Social Network Analysis (SNA) component shown in Figure 13 enables the investigation of complex interactions between individuals and organisations – for example, events that trigger the rapid spread of a rumour in social media, and Twitter accounts that contributed to the dissemination of the story (measured in terms of the number of re-tweets). The visualization helps to track such information diffusion processes and assess the impact of events. The SNA component based on Graphyte significantly extends the analytic framework of the PHEME Visual Dashboard by not only visualizing what is being communicated, but also who is driving the dialogue.

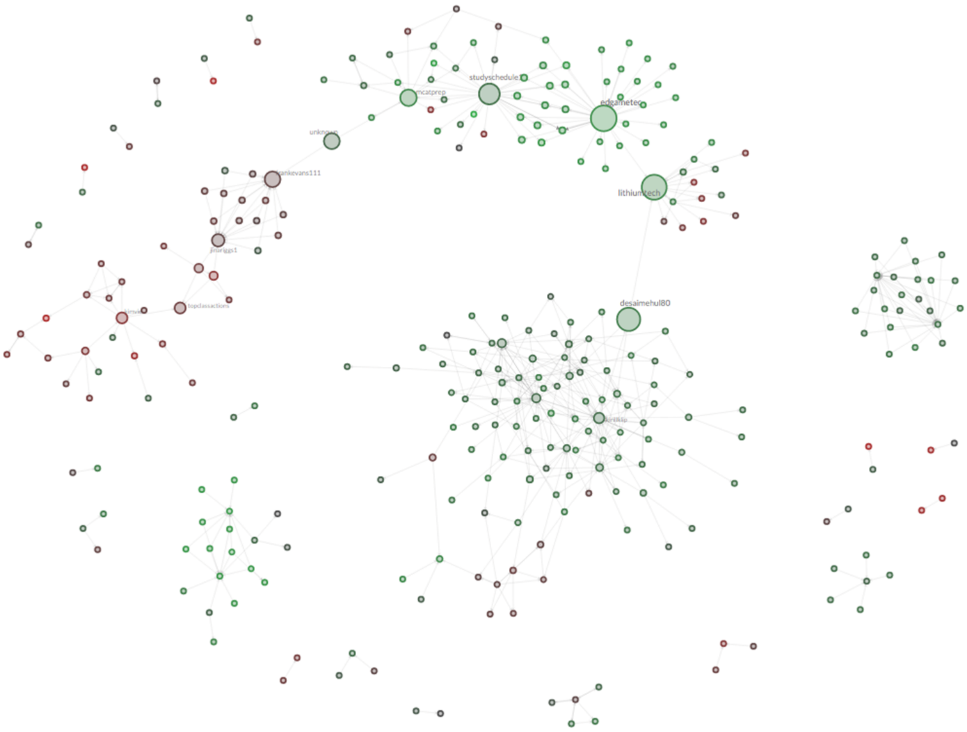


Figure 13. *Social Network Analysis (SNA)* component showing communication patterns among Twitter users within the PHEME social media sample

The data transformation process takes the current search context of the dashboard into account. The graph is the result of real-time data aggregations of communication patterns. *Graphyte's* layout algorithms in conjunction with collision detection optimize the graph layout, considering the varying size of nodes according to the selected network metric – e.g., betweenness centrality. The nodes of the graph show Twitter authors with the highest number of relevant postings. The (directed) edges of the graph indicate "@" references to other Twitter accounts.

The SNA graph visualizes metadata attributes depending on its configuration: Node color shows the average sentiment of all postings of a Twitter user that match the search query in the selected time interval. Additionally, if available, nodes may display the Twitter user's avatar icon. Node size reflects the relative importance in the public dialogue about the selected topic. The configuration menus offer options to select the amount of nodes to display and to switch between different centrality measures:

- *In-Degree Centrality*. Number of incoming links, indicating how many others are mentioning this node (= default metric).
- *Out-Degree Centrality*. Number of outgoing links, which shows the number of references to other nodes.
- *Betweenness Centralit*y. Number of shortest paths that go through a node, reflecting its role as a communication hub in the network structure.
- *Eigenvector Centralit*y. The importance of a node is determined by the importance of its neighbours, similar to Google's PageRank algorithm.

The SNA component includes several interactive display features:

- *Tooltips.* Hovering over a node emphasizes its outgoing links and shows a tooltip with additional information about the Twitter account including user name, number of relevant tweets about the topic, and the average sentiment of the tweets.
- *Highlight Mode.* Clicking on a node greys out the rest of the network and expands the tooltip with three search options (replace, restrict, expand). A single click on additional nodes adds them to the highlighted part of the network, while double clicking disables highlight mode and restores the whole network structure.
- *Zoom Operations* are triggered either via the mouse wheel, or by double clicking a node when the graph is not in highlight mode.
- *Node Positioning.* Users can drag and drop nodes to different positions. The "freeze" button in the upper right corner deactivates the adaptive layout process.

Geographic Map

Similar to how the updated *Keyword Graph* has been extended to support multiple seed nodes, the *Geographic Map* has been extended to support visualizing the geospatial distribution of location references referring to specific topics or metadata attributes (veracity, stance, sentiment, etc.). For this feature, the topic and metadata drill-down sidebar act as a filtering interface to select multiple dimensions for the *Geographic Map.* In combination with the dashboard's date filtering options, this enables the analysis of spatial-temporal distributions of veracity scores within search results of annotated social media data streams.

For this to work, the data transformation and metadata matching of the *PHEME Visual Dashboard* have been improved: In D5.2.1, the *Geographic Map* supported aggregating location data for the current search context only. In D5.2.2 it is possible to select either multiple topics or metadata dimensions via the sidebar as described in the section on metadata integration.

Even without directly interacting with the component, the GeoMap has been designed as an intuitive display providing ambient information about the geographic distribution of news and social media coverage. The GeoMap dynamically updates the circles and arcs whenever a user triggers a new search in the portal. This helps analysts understand the geographic context of their queries without interrupting their workflow.

The *Geographic Map* supports custom base layers. Their full potential unfolds in conjunction with the drill-down capabilities of adaptive tooltips. Based on the user's current context (country shape, point of interest, etc.), the tooltip displays the most relevant information in a local context, and the option to restrict or extend the search.

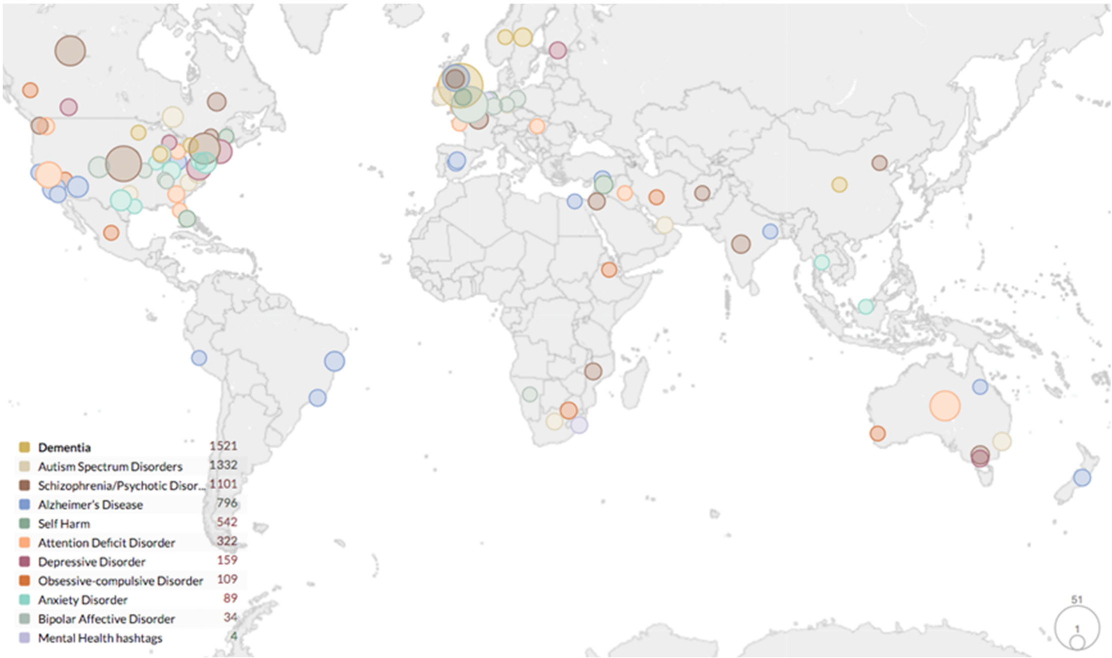

Dementia 1521
Autism Spectrum Disorders 1332
Schizophrenia/Psychotic Disor... 1101
Alzheimer's Disease 796
Self Harm 542
Attention Deficit Disorder 322
Depressive Disorder 159
Obsessive-compulsive Disorder 109
Anxiety Disorder 89
Bipolar Affective Disorder 34
Mental Health hashtags 4
51
1

Figure 14. *Geographic Map* showing geospatial topic distribution throughout 2016

Application Programming Interface (API)

To foster collaboration and leverage synergies between PHEME and the ASAP FP7 project,[4] the data interchange between the dashboard and other PHEME work packages is based on an extended version of the webLyzard API (originally published as part of the ASAP Deliverable 6.2). From the API components, the following are relevant in the context of PHEME:

- *Document API* – ingests unstructured data from social media sources from T6.1 The main API objects are *Documents*, *Sentences* and *Annotations* – the latter are provided by WP2-3, using PHEME-specific API extensions to support the required metadata elements.

[4] ASAP = *A Scalable Analytics Platform* (www.asap-fp7.eu); The use of the webLyzard API not only avoids redundant efforts, but also supports joint exploitation efforts; specific opportunities arise from pursuing a *Visualization-as-a-Service* (VaaS) approach, where PHEME components such as the cluster map and the keyword graph can be offered as part of an integrated framework.

- *Search API* – returns a set of query results in the form of unstructured text documents. The main object is *Query.*
- *Embeddable Visualization API* – represents a standardized way to integrate individual visualizations in third-party applications (as compared to using the full dashboard), which allows using dashboard components in external applications as well, for example the digital journalism showcase (T8.3). The main object is *Visualization,* which is typically rendered based on the results of a *Query*.

To model cross-referencing between documents and represent threaded dialogs, we have extended the document model to support document relations of various types, and have exposed the new structure via the *Document API.* With this addition, one can specify linkage information related to a document to be ingested into the PHEME dashboard by providing the relation type and the target URL through the JSON payload. On request, the API tries to match the provided target URL against existing documents contained in the PHEME metadata repository, resolving link targets to internal IDs where available. The document relation extension has been designed in a generic manner, enabling reasoning on the metadata document level – e.g., outgoing and incoming links for opinion mining, information diffusion paths via temporal linkage analysis, etc.

To facilitate API usage, the documentation[5] has been published using the *Swagger*[6] toolkit. This page represents a central hub to connect the documentation for all API endpoints in a standardized way, providing clear and reliable guidance for technical partners. In addition, code examples allow developers to quickly test and validate client code against the API endpoints.

Summary

This deliverable D5.2.2 summarizes the research and development work conducted in T5.4 of PHEME, implementing a visual analytics dashboard for the medical and health care domain. The dashboard, tailored and extended to meet the requirements of PHEME, was activated in Year 2 and then continuously improved in Year 3 of the project. It integrates *news media articles* with the *social media* feeds of T6.1, and provides interactive tools to analyse the aggregated content along multiple metadata dimensions, including *veracity*, *stance* and *sentiment*.

Several of the dashboard's visual tools use the PHEME open source library *Graphyte,* for example to represent keyword associations, analyse the structure of social networks, and increase the transparency of information flows between authors. The labelling techniques for the *keyword graph* and *cluster map* use keyword filtering techniques that avoid redundancy across clusters and identify references to extracted named entities. The clustering and visualization components can be customized either by pre-defined dashboard property settings (available for administration users), or via interactive dashboard elements, depending on whether these settings should be exposed to end users.

[5] api.weblyzard.com

[6] www.swagger.io

In terms of *geospatial visualization*, the map component has been extended to support colour coding of topic classifications or metadata dimensions such as *veracity, stance, sentiment*, and *source type*. Depending on the structure of the provided dataset, this functionality allows tracking of rumour flows across regions, track the location of social media authors that discuss a certain topic, or reveal the spatio-temporal distribution of specific metadata attributes.

References

Aaker, J.L. (1997). "Dimensions of Brand Personality", *Journal of Marketing Research,* 34(3): 347-356.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). "Fast Unfolding of Communities in Large Networks", *Journal of Statistical Mechanics Theory and Experiment,* 10: 10008.

Bostock, M., Ogievetsky, V. and Heer, J. (2011). "D3: Data-Driven Documents", *IEEE Transactions on Visualization and Computer Graphics,* 17(12): 2301-2309.

Derczynski, L., Maynard, D., et al. (2015). "Analysis of Named Entity Recognition and Linking for Tweets", *Information Processing and Management,* 51: 32-49.

Fischl, D. and Scharl, A. (2014). Metadata Enriched Visualization of Keywords in Context. *6th ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS-2014)*. Italy, Rome: ACM Press: 193-196.

Hubmann-Haidvogel, A., Scharl, A. and Weichselbraun, A. (2009). "Multiple Coordinated Views for Searching and Navigating Web Content Repositories", *Information Sciences,* 179(12): 1813-1821.

Joint Formulary Committee (2014). *British National Formulary (BNF) 67* London: Pharmaceutical Press.

Niepold, F., Herring, D. and McConville, D. (2008). "The Role of Narrative and Geospatial Visualization in Fostering Climate Literate Citizens", *Physical Geography,* 29(6): 529-544.

Scharl, A. and Tochtermann, K., Eds. (2007). *The Geospatial Web - How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society*. London: Springer.

Wattenberg, M. and Viégas, F.B. (2008). "The Word Tree, an Interactive Visual Concordance", *IEEE Transactions on Visualization and Computer Graphics,* 14(6): 1221-1228.

Weichselbraun, A., Streiff, D. and Scharl, A. (2015). "Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence", *International Journal on Artificial Intelligence Tools,* 24(2): 1540008 | 1-31.