



**THEME [ICT-2013.4.1]
[Content analytics and language technologies]**

Grant agreement for: Collaborative project

Annex I - "Description of Work"

Project acronym: PHEME

Project full title: " Computing Veracity Across Media, Languages, and Social Networks "

Grant agreement no: 611233

Version date: 2016-10-19

Table of Contents

Part A

A.1 Project summary	3
A.2 List of beneficiaries	4
A.3 Overall budget breakdown for the project	5

Workplan Tables

WT1 List of work packages	1
WT2 List of deliverables	2
WT3 Work package descriptions	6
Work package 1.....	6
Work package 2.....	9
Work package 3.....	13
Work package 4.....	17
Work package 5.....	21
Work package 6.....	25
Work package 7.....	29
Work package 8.....	33
Work package 9.....	37
WT4 List of milestones	41
WT5 Tentative schedule of project reviews	42
WT6 Project effort by beneficiaries and work package	43
WT7 Project effort by activity type per beneficiary	44
WT8 Project efforts and costs	45

A1: Project summary

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per project

General information

Project title ³	Computing Veracity Across Media, Languages, and Social Networks		
Starting date ⁴	01/01/2014		
Duration in months ⁵	39		
Call (part) identifier ⁶	FP7-ICT-2013-10		
Activity code(s) most relevant to your topic ⁷	ICT-2013.4.1: Content analytics and language technologies		

Abstract ⁹

Social media poses three major computational challenges, dubbed by Gartner the 3Vs of big data: volume, velocity, and variety. Content analytics methods have faced additional difficulties, arising from the short, noisy, and strongly contextualised nature of social media. In order to address the 3Vs of social media, new language technologies have emerged, e.g. using locality sensitive hashing to detect breaking news stories from media streams (volume), predicting stock market movements from microblog sentiment (velocity), and recommending blogs and news articles based on user content (variety).

PHEME will focus on a fourth crucial, but hitherto largely unstudied, challenge: veracity. It will model, identify, and verify phemes (internet memes with added truthfulness or deception), as they spread across media, languages, and social networks.

PHEME will achieve this by developing novel cross-disciplinary social semantic methods, combining document semantics, a priori large-scale world knowledge (e.g. Linked Open Data) and a posteriori knowledge and context from social networks, cross-media links and spatio-temporal metadata. Key novel contributions are dealing with multiple truths, reasoning about rumour and the temporal validity of facts, and building longitudinal models of users, influence, and trust.

Results will be validated in two high-profile case studies: healthcare and digital journalism. The techniques will be generic with many business applications, e.g. brand and reputation management, customer relationship management, semantic search and knowledge management. In addition to its high commercial relevance, PHEME will also benefit society and citizens by enabling government organisations to keep track of and react to rumours spreading online.

PHEME addresses Objective ICT-2013.4.1 Content analytics and language technologies; a) cross-media analytics.

A2: List of Beneficiaries

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

List of Beneficiaries

No	Name	Short name	Country	Project entry month ¹⁰	Project exit month
1	THE UNIVERSITY OF SHEFFIELD	USFD	United Kingdom	1	39
2	UNIVERSITAET DES SAARLANDES	USAAR	Germany	1	39
3	MODUL UNIVERSITY VIENNA GMBH	MOD	Austria	1	39
4	ONTOTEXT AD	ONTO	Bulgaria	1	39
5	ATOS SPAIN SA	ATOS	Spain	1	39
6	KING'S COLLEGE LONDON	KCL	United Kingdom	1	39
7	I-HUB LIMITED	iHUB	Kenya	1	39
8	SCHWEIZERISCHE RADIO-UND FERNSEHGESELLSCHAFT ASSOCIATION	SWI	Switzerland	1	39
9	THE UNIVERSITY OF WARWICK	UWAR	United Kingdom	1	39

A3: Budget Breakdown

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One Form per Project

Participant number in this project ¹¹	Participant short name	Fund. % ¹²	Ind. costs ¹³	Estimated eligible costs (whole duration of the project)					Requested EU contribution
				RTD / Innovation (A)	Demonstration (B)	Management (C)	Other (D)	Total A+B+C+D	
1	USFD	75.0	T	619,280.00	0.00	116,919.00	0.00	736,199.00	581,379.00
2	USAAR	75.0	T	476,800.00	0.00	31,800.00	0.00	508,600.00	389,400.00
3	MOD	75.0	T	498,496.00	0.00	13,600.00	0.00	512,096.00	387,472.00
4	ONTO	75.0	T	408,064.00	0.00	9,600.00	0.00	417,664.00	315,648.00
5	ATOS	50.0	A	568,500.00	0.00	19,500.00	0.00	588,000.00	303,750.00
6	KCL	75.0	T	432,852.00	0.00	5,600.00	0.00	438,452.00	330,239.00
7	iHUB	50.0	S	319,232.00	0.00	10,440.00	0.00	329,672.00	170,056.00
8	SWI	50.0	A	496,320.00	0.00	20,760.00	0.00	517,080.00	268,920.00
9	UWAR	75.0	T	212,153.00	0.00	10,022.00	0.00	222,175.00	169,136.00
Total				4,031,697.00	0.00	238,241.00	0.00	4,269,938.00	2,916,000.00

Note that the budget mentioned in this table is the total budget requested by the Beneficiary and linked Third Parties.

*** The following funding schemes are distinguished**

Collaborative Project (if a distinction is made in the call please state which type of Collaborative project is referred to: (i) Small of medium-scale focused research project, (ii) Large-scale integrating project, (iii) Project targeted to special groups such as SMEs and other smaller actors), Network of Excellence, Coordination Action, Support Action.

1. Project number

The project number has been assigned by the Commission as the unique identifier for your project, and it cannot be changed. The project number **should appear on each page of the grant agreement preparation documents** to prevent errors during its handling.

2. Project acronym

Use the project acronym as indicated in the submitted proposal. It cannot be changed, unless agreed during the negotiations. The same acronym **should appear on each page of the grant agreement preparation documents** to prevent errors during its handling.

3. Project title

Use the title (preferably no longer than 200 characters) as indicated in the submitted proposal. Minor corrections are possible if agreed during the preparation of the grant agreement.

4. Starting date

Unless a specific (fixed) starting date is duly justified and agreed upon during the preparation of the Grant Agreement, the project will start on the first day of the month following the entry into force of the Grant Agreement (NB : entry into force = signature by the Commission). Please note that if a fixed starting date is used, you will be required to provide a detailed justification on a separate note.

5. Duration

Insert the duration of the project in full months.

6. Call (part) identifier

The Call (part) identifier is the reference number given in the call or part of the call you were addressing, as indicated in the publication of the call in the Official Journal of the European Union. You have to use the identifier given by the Commission in the letter inviting to prepare the grant agreement.

7. Activity code

Select the activity code from the drop-down menu.

8. Free keywords

Use the free keywords from your original proposal; changes and additions are possible.

9. Abstract

10. The month at which the participant joined the consortium, month 1 marking the start date of the project, and all other start dates being relative to this start date.

11. The number allocated by the Consortium to the participant for this project.

12. Include the funding % for RTD/Innovation – either 50% or 75%

13. Indirect cost model

A: Actual Costs

S: Actual Costs Simplified Method

T: Transitional Flat rate

F :Flat Rate

Workplan Tables

Project number

611233

Project title

PHEME—Computing Veracity Across Media, Languages, and Social Networks

Call (part) identifier

FP7-ICT-2013-10

Funding scheme

Collaborative project

WT1

List of work packages

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

LIST OF WORK PACKAGES (WP)

WP Number ⁵³	WP Title	Type of activity ⁵⁴	Lead beneficiary number ⁵⁵	Person-months ⁵⁶	Start month ⁵⁷	End month ⁵⁸
WP 1	Project Management	MGT	1	25.00	1	39
WP 2	Ontologies, Multilinguality, and Spatio-Temporal Grounding	RTD	4	58.00	1	18
WP 3	Contextual Interpretation	RTD	1	57.00	1	30
WP 4	Detecting Rumours and Veracity	RTD	2	62.00	4	35
WP 5	Interactive Visual Analytics Dashboard	RTD	3	55.00	4	36
WP 6	Scalability, Integration, and Evaluation	RTD	5	72.00	1	37
WP 7	Veracity Intelligence in Patient Care	RTD	6	55.00	1	38
WP 8	Digital Journalism Use Case	RTD	8	58.00	1	39
WP 9	Dissemination and Exploitation	RTD	5	25.00	1	39
Total				467.00		

WT2: List of Deliverables

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

List of Deliverables - to be submitted for review to EC

Deliverable Number ⁶¹	Deliverable Title	WP number ⁵³	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D1.1	Self Assessment, Risk Assessment and Management Plan	1	1	0.50	R	CO	4
D2.1	Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages	2	9	8.00	R	PU	6
D2.2	Linguistic Pre-processing Tools and Ontological Models of Rumours and Phemes	2	2	24.00	P	PU	12
D2.3	Spatio-Temporal Algorithms	2	1	10.00	P	PU	18
D2.4	Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages: Final Version	2	9	16.00	R	PU	18
D3.1	Cross-Media and Cross-Language Linking Algorithm	3	2	15.00	P	PU	18
D3.2	Algorithms for Implicit Information Diffusion Networks	3	3	14.00	P	PU	24
D3.3.1	Longitudinal models of users, networks, and trust: Initial Prototype	3	1	12.00	P	CO	12

WT2: List of Deliverables

Deliverable Number ⁶¹	Deliverable Title	WP number ⁵³	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D3.3.2	Longitudinal models of users, networks, and trust: Final Public Release	3	1	16.00	P	PU	27
D4.1.1	LOD-based Reasoning about Rumours: Initial Prototype	4	4	12.00	P	CO	22
D4.1.2	LOD-based Reasoning about Rumours: Final Prototype	4	4	12.00	P	PU	32
D4.2.1	Algorithms for Detecting Disputed Information: Initial Prototype	4	2	5.00	P	CO	14
D4.2.2	Algorithms for Detecting Disputed Information: Final Version	4	2	14.00	P	PU	27
D4.3.1	Algorithms for Detecting Misinformation and Disinformation: Initial Prototype	4	1	6.00	P	CO	18
D4.3.2	Algorithms for Detecting Misinformation and Disinformation: Final Version	4	1	13.00	P	PU	35
D5.1.1	Open-source Visual Analytics Tools: Initial Prototype	5	3	10.00	P	PU	12
D5.1.2	Open-source Visual Analytics Tools: Final Version	5	3	10.00	P	PU	22
D5.2.1	PHEME Visual Dashboard: Initial Prototype	5	3	12.00	P	PU	24

WT2: List of Deliverables

Deliverable Number ⁶¹	Deliverable Title	WP number ⁵³	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D5.2.2	PHEME Visual Dashboard: Final Version	5	3	17.00	P	PU	36
D5.3	Usability Evaluation Report	5	3	6.00	R	PU	36
D6.1.1	PHEME Integrated Veracity Framework - v 0.5	6	5	18.00	P	CO	12
D6.1.2	PHEME Integrated Veracity Framework - v.1.0	6	5	24.00	P	PU	24
D6.1.3	PHEME Integrated Veracity Framework - v.2.0	6	5	22.00	P	PU	37
D6.2.1	Evaluation report - Interim Results	6	1	3.00	R	PU	21
D6.2.2	Evaluation report - Final Results	6	2	5.00	R	PU	37
D7.1	Requirements and design documents	7	6	6.00	R	PU	4
D7.2.1	Annotated Corpus - Initial Version	7	6	6.00	O	RE	14
D7.2.2	Annotated Corpus - Final Version	7	6	6.00	O	PU	24
D7.3	Healthcare Application Prototype and User Evaluation results	7	6	36.00	P	PU	38
D8.1	Requirements and Use Case Design document	8	8	6.00	R	PU	4
D8.2	Annotated Corpus of	8	8	14.00	O	PU	18

WT2: List of Deliverables

Deliverable Number ⁶¹	Deliverable Title	WP number ⁵³	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
	Newsworthy Rumours						
D8.3	Digital journalism prototype	8	7	20.00	P	PU	30
D8.3.1	Digital journalism prototype (v2)	8	7	8.00	P	PU	38
D8.4	Evaluation results: validation and analysis	8	8	10.00	R	PU	39
D9.1	Project Fact Sheet	9	1	0.50	R	PU	1
D9.2	Project Website	9	1	7.50	O	PU	3
D9.3	Dissemination and exploitation Plan (v0.5 M9, v.1 M18)	9	5	6.00	R	CO	9
D9.4	Dissemination and exploitation Report	9	1	3.00	R	PU	39
D9.5.1	Market Watch - Initial Version	9	5	4.00	R	CO	12
D9.5.2	Market Watch - Final version	9	5	4.00	R	CO	30
Total				441.50			

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP1	Type of activity ⁵⁴	MGT
Work package title	Project Management		
Start month	1		
End month	39		
Lead beneficiary number ⁵⁵	1		

Objectives

The Management work package ensures:

- that the project objectives are met within the planned effort and budget,
- that it is effectively and correctly managed financially,
- that its progress, quality, and status are efficiently and effectively monitored,
- that the required reporting is prepared and delivered in a timely manner,
- that all quality aspects of the project are fully and correctly addressed,
- the infrastructure for dissemination and central intra-project communication and cooperation.

The project management will entail strategic, project-wide, as well as day-to-day, central management and coordination activities.

Description of work and role of partners

Prior to the start of the project, the consortium will put in place a Consortium Agreement following the EC recommended template. The overall project management will be carried out by the coordinator, USFD. There will also be a project administrator at USFD, Lucy Moffatt, with a proven track record in administrative and financial administration of EC projects. USAAR will assist by co-chairing the scientific and technical management board, whereas ATOS will lead the exploitation board (see Section 2.1.1).

Task 1.1 Management of the consortium (M1 – M39) - USFD, USAAR, ATOS

Monitoring the application of the Consortium Agreement and the compliance of all partners with the regulations put forth in it. Planning of consortium meetings according to the projects' work plan and EC guidelines. Ensuring effective communication and cooperation between work packages.

Task 1.2 Communication with the EC and periodic reporting - USFD

Preparing and creating the necessary management reports according to EC guidelines. Contacting the EC services for administrative purposes.

Task 1.3 Risk assessment and contingency planning (M1 – M39) - USFD

A comprehensive Risk Assessment and Management Plan will be implemented within six months of project start, addressing the different kinds of risk (external, internal, strategic, operational, others). We will identify risks of any nature that might occur and assess their probability and potential impact on the project. Concrete actions will be planned, to prevent the occurrence of the risk. If problems do occur, then the associated contingency measures will be implemented swiftly to minimize the impact (D1.1).

Task 1.4 Financial Management (M1 – M39) - USFD

WT3: Work package description

Preparing and compiling of Periodic Financial Report. Obtaining Audit Certificates where applicable. Establishing and maintaining an internal tool to monitor resource expenditure per partner and per WP.

Task 1.5 Internal communication (M1 – M39) - USFD

Establishing and maintaining of mailing lists according to the needs of the project. Establishing and maintaining the internal project document repository.

Task 1.6 Quality assurance of deliverables (M1 – M39) - USFD

Establishing a quality assurance procedure that every deliverable has to undergo before submission to the EC.

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
1	USFD	12.00
2	USAAR	3.00
3	MOD	1.00
4	ONTO	1.00
5	ATOS	3.00
6	KCL	1.00
7	iHUB	1.00
8	SWI	2.00
9	UWAR	1.00
Total		25.00

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D1.1	Self Assessment, Risk Assessment and Management Plan	1	0.50	R	CO	4
Total			0.50			

Description of deliverables

D1.1) Self Assessment, Risk Assessment and Management Plan: This document will detail internal procedures to ensure the project is being governed according to plan and meeting its own targets for the level of user involvement, results evaluation, and technical delivery. The assessment plan will also look to track partner satisfaction with deliverables and that deliverables are peer reviewed and verified. This deliverable will form part of T1.3. [month 4]

WT3: Work package description

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS1	Inception	1	4	Inception Check & Risk analysis and plan maintenance
MS2	Initial Data and Requirements Analysis	1	6	
MS3	First Development and Delivery Cycle	1	12	Year 1 Completion, Progress Review, and Y2 Planning
MS4	Project Mid-Point Evaluation	1	18	M18 Quantitative and user-based, qualitative evaluation results
MS5	Second Development and Delivery Cycle	1	24	Year 2 Completion, Progress Review, and Y3 Planning
MS6	Project Pre-Completion Evaluation	1	30	Preparation for final technical releases and user experiments
MS7	Project Completion	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP2	Type of activity ⁵⁴	RTD
Work package title	Ontologies, Multilinguality, and Spatio-Temporal Grounding		
Start month	1		
End month	18		
Lead beneficiary number ⁵⁵	4		

Objectives

- Undertake qualitative social science analysis of rumours, across media and languages
- Build ontological models of veracity, misinformation, social and information diffusion networks, rumours, disputed claims, temporal validity of statements, and user online behaviour
- Compare, adopt and adapt the necessary linguistic pre-processing tools for the three target languages, including language identification, POS tagging, chunking, dependency parsing, entity and relation recognition, LOD-based entity disambiguation, and sentiment and opinion mining. These will be used to generate linguistic and semantic features for the WP3 and WP4 methods.
- Develop multilingual spatio-temporal annotation tools, adapted also to noisy social media content

Description of work and role of partners

T2.1 Qualitative Social Science Analysis of Rumour across Media and Languages (M1 – M18) – UWAR, SWI

T2.1 will carry out qualitative analysis of the four types of rumours, central to PHEME: speculation, controversy, misinformation, and disinformation, using the methodology of Procter et al. (2013) (UWAR). Initial results will be delivered in D2.1, which will then be extended and elaborated in D2.4. First, candidate topics will be identified and examples for each rumour type extracted from the corpus created in T8.2. Following initial analysis, one example representing each of the four rumour types will be selected for in-depth analysis. An actor code frame will be developed for annotating the type of source (e.g., mainstream media, businesses, bloggers, government agencies, etc.). Information flow analysis across media will be conducted and the information flows extracted then annotated using the actor type code frame and annotation schemas defined in T8.2. Links between media types will be unpacked and followed to establish cross-media links and iteratively expand the corpora. Finally, for each rumour, the role of each media element and that of the actor associated with it will be analysed to gain in-depth understanding of the contribution of factors such as trust, credibility and evidence to its propagation and its 'life history'.

The actor code frame and annotation schema will be validated on test corpora using standard measures of inter-coder agreement as they are developed. The annotated rumour corpora will be validated in a similar way. The analysis of the rumour corpora will be reviewed by an independent panel of experts (D2.1, D2.4).

T2.2 Ontological modelling (M1 – M12) – ONTO, USAAR

This task will build new and extend existing ontologies to model veracity, misinformation, social and information diffusion networks, rumours, disputed claims and temporal validity. The results will be delivered in D2.2. We will distinguish between content authors (wrote the content originally), receivers (users who received a given content, e.g. in their Twitter stream), and diffusers (users who propagated that content). T2.2 will also model the temporal validity of statements (e.g. "Lenin was born in the Soviet Union" vs. "in Russia") and lexicalisations (e.g. Kaliningrad vs. Königsberg), based on USAAR's work on adding temporal arguments to RDF triples (Krieger, 2010). The formal model will also distinguish between posts supporting, denying, or questioning a given rumour/statement. ONTO's PROTON ontology will be used to model the entities, events, relations mentioned in a given document, as well as their grounding in LOD resources. For modelling interlinked authoritative sources, UGC, social networks, and online sharing practices T2.2 will evaluate (and if necessary extend)

WT3: Work package description

relevant ontologies, such as DLPO (The LivePost Ontology). DLPO models social media posts, going beyond Twitter (Scerri et al, 2012) and is strongly grounded in FOAF, SOIC, and the Simple Knowledge Organisation System (SKOS). The ontology captures six main types of knowledge: online posts, different kinds of posts (e.g. retweets), microposts, online presence, physical presence, and online sharing practices (e.g. liking, favouriting). T2.2 will also model user behaviour and user discussions, where a good candidate is the User Behaviour Ontology (Angeletou et al, 2011). It captures the impact of posts (replies, comments, etc), user behaviour, user roles (e.g. popular initiator, supporter, ignored), temporal context (time frame), and other interaction information. The spatio-temporal grounding is of particular importance to PHEME, especially for modelling changes over time (e.g. new online friends, recently discussed topics).

We will involve social media domain experts in the development and evaluation process in T2.2. They will help to create questionnaires with respect to the coverage and the quality of the resulting models. Another aspect is an evaluation against a corpus of social media posts coded by two manual annotators. Firstly, any discrepancies between the manual annotators would indicate areas for improvement. Secondly, a comparison between the manually coded annotations and the domain modelling shall be discussed and handled appropriately. Wikipedia/DBPedia will also be used as valuable resource to evaluate the T2.2 models in terms of coverage. The above mentioned methods of quality management will be incorporated in the lifecycle of the ontology (D2.2).

T2.3 Multilingual Pre-processing (M1 – M12) – USAAR, USFD, ONTO

T2.3. will compare and, if necessary, carry out low-effort adaptation to user-generated content, of existing NLP tools, needed for pre-processing of the diverse media and language content, to be processed in PHEME. The results will be delivered in D2.2.

Three languages will be addressed: English (USFD), German (USAAR), Bulgarian (ONTO). Firstly, the highly-scalable low-level tools from the TRENDMINER project will be reused (language identification, spam detection, boilerplate removal, normalisation, tokenisation and POS tagging). Secondly, relevant named entity recognition, relation extraction and LOD-based entity disambiguation tools will be reused (especially for EN and DE), including those from the TRENDMINER, OpeNER, and X-Like projects, as well as commercial web services (e.g. AlchemyAPI, OpenCalais). Thirdly, T2.3 will identify, compare and reuse the most suitable sentiment and opinion mining tools from Limosine, EuroSentiment, SentiStrength, and others. Lastly, for entailment and contradiction detection (WP3) we will also run dependency parsing (e.g. Minipar, the StanfordParser), and shallow parsing (e.g., the GATE NP and VP Chunker). For pre-processing the WP7 medically-related content, T2.3 will use relevant biomedical NLP tools (e.g. ABNER, MetaMap, BADREX, GENIA). All this lexical, syntactic, and semantic information will be used as document-intrinsic features by the context interpretation (WP3) and rumour detection and verification algorithms (WP4). The detected entities, events, and opinions will also be stored in the ontological models (T2.2), based on PROTON and opinion ontologies.

State-of-the-art systems and tools (e.g. those from TrendMiner) will be used as a baseline. Performance will be evaluated in terms of precision, recall, and f-score. T2.3 will also benefit from Bulgarian manually annotated corpora of social media (developed as part of the QLeap project), which will be interlinked to the social and rumor ontologies developed as part of PHEME. That new extension of the BulTreeBank will support the development and evaluation of tools for disambiguation of concepts mentions in social media texts in Bulgarian (D2.2).

T2.4. Spatio-Temporal (ST) Grounding and Content Geolocation (M3 – M18) – USFD, USAAR

T2.4 will create general-purpose tools for projecting spatio-temporal annotations across languages, given parallel texts and re-using existing corpora (e.g. TimeBank, the multilingual TempEval, ACE2 temporal annotations, WikiwarsDE). These resources will be used to develop multilingual temporal annotation tools, based on their state-of-the-art techniques, developed for longer texts. T2.4 will also address the problem of geolocating documents. Our method will go beyond features based on words in the document and will use disambiguated URIs (e.g. against GeoNames) and additional knowledge from the LOD resource (e.g. NUTS subdivisions, latitude/longitude, neighbouring locations).

The results of T2.4 will be delivered in D2.3 (D2.3).

WT3: Work package description

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
1	USFD	12.00
2	USAAR	6.00
4	ONTO	16.00
8	SWI	6.00
9	UWAR	18.00
Total		58.00

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D2.1	Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages	9	8.00	R	PU	6
D2.2	Linguistic Pre-processing Tools and Ontological Models of Rumours and Phemes	2	24.00	P	PU	12
D2.3	Spatio-Temporal Algorithms	1	10.00	P	PU	18
D2.4	Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages: Final Version	9	16.00	R	PU	18
Total			58.00			

Description of deliverables

D2.1) Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages: This report will detail the results of the social science analysis of naturally occurring rumours and their propagation across social networks, news media, and other information sources. The early version at M6 will contain initial findings, to verify the validity of the rumour linguistic annotation schemas, used in the corpora from the healthcare (T7.2) and digital journalism (T8.2) use cases. [month 6]

D2.2) Linguistic Pre-processing Tools and Ontological Models of Rumours and Phemes: This deliverable will detail the results from the comparison of existing multilingual pre-processing tools and specify which tools will be reused in PHEME to pre-process the contradictions, rumour, and disputed claims corpora and produce lexical, syntactic, and semantic features for the WP3 and WP4 algorithms. USFD will address English tools, USAAR – German, and ONTO – Bulgarian. While some tools are already adapted to the noisiness of user-generated content (e.g. the TRENDMINER language identification, normalisation, and LOD-based entity disambiguation tools), others have not (e.g. the dependency and shallow parsers). Therefore, this deliverable will also include limited adaptation effort for such tools on user-generated content. ONTO will also deliver the ontologies developed in T2.3, freely available to others to reuse, in RDF/OWL format [month 12]

D2.3) Spatio-Temporal Algorithms: This software deliverable will include open-source tools and algorithms arising from T2.4, as well as the automatically projected multilingual annotations. [month 18]

D2.4) Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages: Final Version: This is the completed, final version of the report, detailing the social science research in T2.1. This report will contain

WT3: Work package description

all information from D2.1 plus the in-depth analysis of the four types of rumours studied in PHEME: speculation, controversy, misinformation and disinformation. The analysis methodology is detailed in Task 2.1. [month 18]

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS2	Initial Data and Requirements Analysis	1	6	
MS3	First Development and Delivery Cycle	1	12	Year 1 Completion, Progress Review, and Y2 Planning
MS4	Project Mid-Point Evaluation	1	18	M18 Quantitative and user-based, qualitative evaluation results

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP3	Type of activity ⁵⁴	RTD
Work package title	Contextual Interpretation		
Start month	1		
End month	30		
Lead beneficiary number ⁵⁵	1		

Objectives

This workpackage is focused on interpreting the wider cross-media and social graph context, within which a given rumour appears, i.e. posteriori knowledge. More concretely:

- Topically related stories are identified automatically, as well as the dialogue threads inside
- User-generated content is aligned to authoritative content, via cross-media and language linking
- Implicit information diffusion networks across media are identified
- Longitudinal models of users, trust and influence are computed

Description of work and role of partners

T3.1 Cross-Media and Cross-Language Linking (M7 – M18) – USAAR

The aim of T3.2 is to align automatically user-generated content to authoritative content (across media and languages), e.g. tweets and blog posts to contemporaneous mainstream news, patient forum posts to relevant MEDLINE abstracts. The results will be delivered in D3.1.

This task will build on a monolingual entity- and keyphrase-based linking method for aligning news video segments with news web pages (Dowman et al, 2005). The algorithm will be extended to make use of multilingual lexicalisations from LOD resources, temporal and spatial knowledge, and deeper semantics (i.e. disambiguated organisations, locations, and other entities, events grounded in ontologies, and predicate-argument structures from T2.3). T3.2 will also investigate cross-document and cross-language event coreference using existing corpora (e.g. AQUAINT TimeML).

For evaluation, we will take as baseline the results of a simple IR approach for linking events described in mainstream news and twitter described in Hristo Tanev, Maud Ehrmann, Jakub Piskorski, and Vanni Zavarella. Enhancing event descriptions through twitter mining. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, ICWSM. The AAAI Press, 2012. They reported a precision of 75% in relating tweets to news clusters. The method described in this paper was using word co-occurrence clustering, domain-specific keywords and named entity recognition. Our approach will make use additionally of LOD and will deal with multilingual documents, a fact which can affect performance (D3.1).

T3.2. Implicit Information Diffusion Networks Across Media (M7 – M24) - MOD

This task will identify implicit information diffusion networks (Scharl et al, 2007), which provide the basis for the identification (WP4) and visualisation of rumours (WP5), across media and over time. The results will be delivered in D3.2.

Task 3.2 will deal with the identification of similar contagions, based on methods from text reuse (e.g. sketches of shingles of tokens). Next these contagions will be tracked across different media (authoritative sources, web content, social networks) and used to infer the implicit diffusion network, using temporal meta-data to determine the optimum time window size (c.f. Belák et al., 2012). T3.2 will draw upon spreading activation, integrating internal (e.g., contact with the contagion through a neighbouring node) and external stimuli (e.g., influence by out-of-network sources, e.g. other media), to provide a holistic model of the information diffusion process (D3.2).

WT3: Work package description

T3.3. Identifying Stories and Conversations (M1 – M12) - USFD

The T3.3 algorithms will cluster content streams into topically-related stories and then analyse conversations within these clusters to identify threads of information flow. The results will be delivered in D3.3.1.

Since story detection through topic modelling is computationally expensive, recent work has used locality-sensitive hashing (Petrovic et al, 2012). T3.3 will compare this approach against a PMI-based event detection approach, developed by USFD in TRENDMINER. However, where entities are mentioned, instead of the words, the corresponding LOD URIs will be used. This will enable the algorithms to distinguish between occurrences of the word apple from mentions of Apple Inc. Within each story cluster, we will identify the threads of the information flows (also referred to as dialogue threads), i.e. who replied what to whom in blog post comments, how are twitter conversations around a shared hashtag structured. These will be used as input to T3.2 on implicit diffusion networks and also Task 3.4 on longitudinal models (D3.3.1).

T3.4. Longitudinal Models of Users, Networks, Trust, and Influence (M13 – M27) - USFD

This task is motivated by the observation that the trustworthiness of a user/web site depends on the veracity of past content. The opposite is also true and is investigated in WP4 below, i.e. that the veracity of a given message depends, amongst other things, on the trustworthiness of its author. The results will be delivered in D3.3.2.

T3.4 will use USFD's historical Twitter data (7 TB dating back to 2009), the SWI, METER and APA news corpora, and any other linked blogs, forums, and web content. These will be searched for known false rumours, acquired from fact-checking websites, in order to create automatically large amounts of training data on past rumours and their spread.

The longitudinal models in T3.4 will be based on pioneering research on detecting epidemics on Twitter (Lampos et al, 2010) (Lampos is now a member of the USFD team). This task will develop predictive models of the overall prevalence of specific rumours in social media, and characterise their temporal profiles using time-series prediction techniques such as auto-regression. Information about the user geographic location will be taken into account (T2.4).

As a second stage we will extend the model to consider individual users and how they react to incoming rumours, critically seeking to characterise what types of incoming message are likely to be propagated. We will model rumour spread using contact process models of epidemiology (Barbour, 1990), whereby a disease is said to spread over a graph according to a Markov process. At each moment in time an ill individual (i.e. a user who currently believes a rumour) may infect a neighbour (i.e. share the rumour with them) or recover from the illness (i.e. realise they were mistaken due to new information), both with given probability. The model will use rich feature based models for each of these probabilities, parameterised by content- and network-level features, e.g., based on user profiles, post history and social network structure. In such a way the model will be used to predict which users are most important for rumour spread, find trusted sub-networks and predict the continuing evolution of a rumour.

The longitudinal models will be evaluated in terms of their predictive performance, i.e. whether a beginning rumour will "go viral" by obtaining broad coverage of the social network, or else die out. A second evaluation scenario is given incomplete historical data about the rumour, predicting who was involved in spreading the rumour. Performance will be measured using prediction accuracy and subjective human evaluation of usefulness for test cases from WP7 and WP8. Baselines will be drawn from the literature, e.g., using simpler feature sets, constant rumour prevalence, or generic occurrence profiles (D3.3.2).

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
1	USFD	28.00
2	USAAR	15.00
3	MOD	14.00

WT3: Work package description

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
	Total	57.00

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D3.1	Cross-Media and Cross-Language Linking Algorithm	2	15.00	P	PU	18
D3.2	Algorithms for Implicit Information Diffusion Networks	3	14.00	P	PU	24
D3.3.1	Longitudinal models of users, networks, and trust: Initial Prototype	1	12.00	P	CO	12
D3.3.2	Longitudinal models of users, networks, and trust: Final Public Release	1	16.00	P	PU	27
		Total	57.00			

Description of deliverables

D3.1) Cross-Media and Cross-Language Linking Algorithm: This deliverable contains the results of T3.1. It will comprise a software deliverable, accompanied by a short documentation and evaluation report. The algorithm will be tested and evaluated on a gold standard subset from the journalism data collected in WP8, including multilingual cross-media content (EN, DE). [month 18]

D3.2) Algorithms for Implicit Information Diffusion Networks: This software deliverable will provide open-source methods for tracking the flow of rumours, contradictions, contentious claims, and disinformation, across media and social networks, arising from Task 3.2. The corpora from D2.1 will be used for development, parameter tuning, and initial evaluation, whereas those from WP7 and WP8 - for final evaluation. [month 24]

D3.3.1) Longitudinal models of users, networks, and trust: Initial Prototype: This software deliverable has an initial prototype at M12, which will include open source algorithms for story detection and the identification of dialogue threads, developed in T3.3. These will be evaluated on the rumour corpora collected in T7.2 and T8.2, as appropriate. Identifying stories and conversations is a pre-requisite for the final version of the software, which will be the longitudinal models of users, networks, and trust. [month 12]

D3.3.2) Longitudinal models of users, networks, and trust: Final Public Release: The final public release of this deliverable will elaborate on the initial prototype (D3.3.1). The final public release will implement the predictive models described in Task 3.4. The algorithms will be evaluated on the unseen gold-standard corpora created in the use cases (WP7 and WP8). [month 27]

WT3: Work package description

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS3	First Development and Delivery Cycle	1	12	Year 1 Completion, Progress Review, and Y2 Planning
MS4	Project Mid-Point Evaluation	1	18	M18 Quantitative and user-based, qualitative evaluation results
MS5	Second Development and Delivery Cycle	1	24	Year 2 Completion, Progress Review, and Y3 Planning

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP4	Type of activity ⁵⁴	RTD
Work package title	Detecting Rumours and Veracity		
Start month	4		
End month	35		
Lead beneficiary number ⁵⁵	2		

Objectives

- Investigate new methods for detecting the different kinds of rumour
- Implement reasoning about rumours to determine veracity

Description of work and role of partners

T4.1 LOD-Based Reasoning about Rumours (M13 – M32) – ONTO.

T4.1 requires the creation of a domain-independent model of the four kinds of rumours addressed here: misinformation, disinformation, unverified information, and disputed information. Initial results will be delivered in D4.1.1 and then final software release in D4.1.2.

Modelling and reasoning with rumours is particularly challenging, due to the need to represent multiple possible truths (e.g. superfoods may cause vs. prevent cancer). The reasoning will be parameterised further in accordance to the domains of the two use cases (healthcare and digital journalism). Reasoning will also take into account the temporal validity of information, to accommodate that two otherwise contradictory statements can be both valid at different points in time (e.g. The President of the US is a different person at different times). Provenance models and stream reasoning approaches will be reused from TRENDMINER. For the knowledge base, ONTO's OWLIM will be adapted as a semantic repository with scalable light-weight reasoning. It will support efficient inference on tractable dialects of RDF(S) and OWL against billions of facts (RDF triples).

Knowledge from relevant Linked Open Data (LOD) resources will be used, such as ONTO's FactForge, DBpedia, and OpenCyc for journalism and ONTO's LinkedLifeData (including PubMed, Uniprot, etc.) for the medical use case. These datasets will provide invaluable large-scale world-knowledge of meronymy, synonymy, antonymy, hypernymy, and functional relations, which are all essential features for classifying entailment and contradictions. In addition, functional relations, semantic constraints, and temporal validity will also be taken into account when detecting contradictions. Contradictions arising from numeric values will be cross-checked also against relevant use-case specific Linked Government Data (D4.1.1, D4.1.2).

T4.2 Detecting Disputed Information: Entailment and Contradictions (M4 – M27) – USAAR.

The detection of controversies will be modelled as a textual entailment and contradiction detection task. Initial results will be delivered in D4.2.1 and then final software release in D4.2.2.

Since textual entailment methods are better understood and more developed (including open-source tools developed in the Excitement project, the GATE-based predicate-argument extractor (Krestel et al, 2010) and a GATE-based textual entailment system (Krestel et al, 2008)), these will be reused. The multilingual pre-processing tools from T2.3 will be used to create the necessary linguistic, syntactic, and semantic features, including dependency parses and hypothesis graphs (Marneffe et al, 2008).

A supervised machine learning classifier for contradiction detection will be trained not just on the small corpora currently available, but also bootstrapped from automatically created training examples. These will be collected from the web and the corpora from WP7 and WP7, using Hearst-like patterns to detect statements like "<Entity> claims/alleges that..." or "scientists claim that...", as well as seeded with known use case specific controversies (T7.1 and T8.1) (e.g. whether or not aluminium causes Alzheimer's).

WT3:

Work package description

The contradiction detection algorithms will take into account document-intrinsic features (e.g. negation, meronymy, synonymy, antonymy, modalities, and factive verbs). A priori LOD knowledge will be used to facilitate entity type and relation matching. Problems with ambiguity of entities, especially locations, will be addressed through the use of the LOD-based entity disambiguation tools (T2.3).

Posteriori knowledge from historical data and cross-media context will be added, including information velocity (e.g. number of tweets/blog/forum posts in the past 1 hr/12hrs/24hrs asserting the same or the opposite pHEME); cross-media features (e.g. links to other sources, trustworthiness and authority); historical features (T3.4) (e.g. has this user/online source previously propagated rumours, about what and when).

The new T4.2 algorithms will be evaluated and benchmarked against any contradiction detection algorithms released as part of the forthcoming Excitement open-source RTE framework. We will re-use here mainly the evaluation procedure that are in place in the RTE (Recognizing textual entailment) campaigns (see also IDO DAGAN, BILL DOLAN, BERNARDO MAGNINI and DAN ROTH, Recognizing textual entailment: Rational, evaluation and approaches, Natural Language Engineering 15, Cambridge University Press 2009). It will consist in re-using existing annotated data sets or developing new ones, an activity foreseen in PHEME, and computing precision and recall measure on results of our systems compared to the gold standard (D4.2.1, D4.2.2).

T4.3 Detecting Misinformation through Credibility Assessment (M7 – M35) – USFD.

Misinformation and disinformation tend to be questioned more than facts, attract more affirmations and denials/refutations, and result in deeper conversation threads (Mendoza et al, 2011). T4.3 will build supervised machine learning classifiers for detecting new misinformation, based on document-intrinsic stylistic and content features (e.g. text quality e.g. emoticons, shouting (Weerkamp & de Rijke, 2012); sentiment, opinions, and entities (from T2.3); mood; personality; hashtags; URLs), posteriori contextual features (e.g. alignment to news event, semantic similarity to credible sources, comments attracted (or RTs and replies)); historical user data (e.g. age, location, frequency of postage, trustworthiness); information spread features (e.g. depth of re-tweets; number of originating sources). Some of the document-intrinsic features are motivated by linguistics-based cues for detecting deception (Zhou et al, 2004); social network features: directly connected users who recently agreed, disagreed or questioned the suspected misinformation). When comparing the pHEME to mainstream media sources, we will distinguish between broadsheets and tabloids (more controversial).

Since training data will be imbalanced between positive and negative examples (e.g. the majority of tweets are not misinformation), T4.3. will experiment with the uneven margins SVM and perceptron algorithms, originally developed by USFD for dealing with imbalanced data in named entity recognition (Li et al, 2009). Performance evaluation will be carried out in Task 6.4.

Resources permitting, T4.3 will also experiment with rumour discovery by comparing the longitudinal models of users and networks (T3.4) against the trends and topics currently spreading across media, in order to identify candidates for rumours or newly emerging facts.

All experimental results and algorithms will be delivered in D.4.3.1 (D4.3.1).

T4.4 Detecting Disinformation (M18 – M35) – USFD.

This task builds on the results of T4.3 and is concerned with classifying disinformation, i.e. malicious users spreading wrong information with intent to deceive. One such example is political abuse during election campaigns (Ratkiewicz et al, 2011). The longitudinal model of user, their trustworthiness, and the trustworthiness of their network connections (T3.4) will be used as features, as will be the credibility score (T4.3) and the reasoning-based validity (T4.1). We will also explore different users' motivation for posting and spreading disinformation. This task will also look at patient forums and flag false statements, posted by unreliable users.

With respect to evaluation, the biggest available dataset (Qazvinian et al, 2011) includes 10,000 manually annotated tweets with respect to five pre-identified rumours, whereas the Truthy dataset (Ratkiewicz et al, 2011) has 61 false claims and 305 legitimate ones. Performance evaluation of T4.4 (and Task 4.3) will be undertaken in Task 6.4, also using the corpora newly developed in the PHEME use cases (WP7 and WP8).

All experimental results and algorithms will be delivered in D.4.3.2 (D4.3.1).

WT3: Work package description

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
1	USFD	18.00
2	USAAR	18.00
4	ONTO	24.00
8	SWI	2.00
Total		62.00

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D4.1.1	LOD-based Reasoning about Rumours: Initial Prototype	4	12.00	P	CO	22
D4.1.2	LOD-based Reasoning about Rumours: Final Prototype	4	12.00	P	PU	32
D4.2.1	Algorithms for Detecting Disputed Information: Initial Prototype	2	5.00	P	CO	14
D4.2.2	Algorithms for Detecting Disputed Information: Final Version	2	14.00	P	PU	27
D4.3.1	Algorithms for Detecting Misinformation and Disinformation: Initial Prototype	1	6.00	P	CO	18
D4.3.2	Algorithms for Detecting Misinformation and Disinformation: Final Version	1	13.00	P	PU	35
Total			62.00			

Description of deliverables

D4.1.1) LOD-based Reasoning about Rumours: Initial Prototype: D4.1 implements the rumour reasoning algorithms within OWLIM. An initial software prototype v 0.5 is released early to enable integration and use within T4.2, T4.3 and T4.4. [month 22]

D4.1.2) LOD-based Reasoning about Rumours: Final Prototype: This is the final, public prototype of the rumour reasoning algorithms. It will provide the complete reasoning support, as detailed in T4.1. [month 32]

D4.2.1) Algorithms for Detecting Disputed Information: Initial Prototype: The early software release v0.5 will be the first prototype of the algorithms developed in T4.2, based on relevant open-source entailment tools, including those from the Excitement project. This early prototype will underpin the development of the misinformation and disinformation detection algorithms and enable early integration and scalability testing in the PHEME architecture. [month 14]

D4.2.2) Algorithms for Detecting Disputed Information: Final Version: This is the final, public prototype of the algorithms developed in T4.2. The complete and tested version v.1 will contain the fully evaluated and refined algorithms, as well as software documentation. [month 27]

WT3: Work package description

D4.3.1) Algorithms for Detecting Misinformation and Disinformation: Initial Prototype: The initial prototype (v.0.5) will contain the first prototype of the algorithm for detecting misinformation through credibility assessment (T4.3). It will enable early integration into the PHEME architecture (WP6) and the related WP5 visualisations. [month 18]

D4.3.2) Algorithms for Detecting Misinformation and Disinformation: Final Version: This is an updated version of D4.3.1. The final, public version will include the refined prototype of the misinformation detection algorithm, tested across media. V.1 will also deliver the disinformation detection algorithm, where the interlinked content and networks will be taken into account, as well as the implicit diffusion network (D4.2) and features based on historical user behaviour and reputation (D4.3). [month 35]

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS4	Project Mid-Point Evaluation	1	18	M18 Quantitative and user-based, qualitative evaluation results
MS6	Project Pre-Completion Evaluation	1	30	Preparation for final technical releases and user experiments
MS7	Project Completion	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP5	Type of activity ⁵⁴	RTD
Work package title	Interactive Visual Analytics Dashboard		
Start month	4		
End month	36		
Lead beneficiary number ⁵⁵	3		

Objectives

Understanding longitudinal datasets at different levels of granularity and analysing complex information spaces are crucial elements of successful decision-making. WP5 will provide an interactive visual analytics framework, supporting several levels of granularity, from tracking threaded dialogs between individuals to revealing evolving community structures and large-scale diffusion patterns within and across network structures. While temporal activity and temporal intensity views are suitable for discovering temporal patterns and temporal behaviour, they generally cannot express complex, manifold relations and patterns in interlinked media and social networks. PHEME will therefore develop new methods to help users visualise and analyse contradictory and misleading information, spreading across multiple media and social networks, as well as the temporal-semantic connections. For this purpose, WP5 will draw upon (i) graph-based visualizations of social networks and ontological structures, (ii) topographical rendering of dynamic content clusters, and (iii) geospatial projections to reveal the patterns of spread of facts, misinformation, contradictions, rumours.

Description of work and role of partners

T5.1 Visualisation of Rumour Conversations over Time - M4 - M12 - MOD

This task will implement methods to show the stories and conversations (discovered in T3.3) in a compact visual map, identifying the participants of the discussion and its chronology. Colour coding to identify authors, quick navigation within and across threads, and a flexible mechanism to select appropriate timescales represent the most granular elements of the PHEME visualisation framework. For rapid prototyping of the desired visual methods, T5.1 and T5.2 will use the D3 JavaScript library (Bostock et al., 2011) and extend MOD's open-source ThreadVis component (threadvis.mozdev.org/) As opposed to other similar libraries, D3 is not focused on a new grammar for graphics, but rather on how to integrate existing standards to create effective visualisations.

These algorithms will be delivered in D.5.1.1 (D5.1.1).

T5.2 Visualising Information Exchanges and Rumour Propagation in Explicit and Implicit Networks - M13-M24 - MOD

This task will go beyond the individual interactions of T5.1 and provide aggregate view on information/rumour distribution patterns, in terms of both node structure and communicated content. The visualisation is underpinned by the results of T3.3 on information diffusion. The methods developed in T5.2 will allow analysts to track longitudinal changes of the network (including the spread of rumours and misconceptions) and visually identify specific types of diffusion patterns - e.g. spikes after external events, resonance patterns as a result of distributed interactions around a common theme, etc. These patterns will manifest themselves along multiple dimensions, to describe associations and functional dependencies: semantic (similarity), geospatial (proximity), and temporal (dependency, correlation and causality)

These algorithms will be delivered in D5.1.2 (D5.1.2).

T5.3 Geospatial Projections of Author Distribution and Sphere of Influence (M25 - M36) - MOD

This task will be responsible for creating dynamic overlays on top of selectable base topographies to reveal the spatio-temporal distribution of authors on a given topic, track misinformation/rumour flows across regions,

WT3: Work package description

and display metadata such as type of source (individual author, news media outlet, corporate Web site, etc.), authority (e.g., number of followers, re-tweets), productivity (i.e., numbers of publications in a given timeframe), and sphere of influence (i.e., geographic area of the recipients, colour-coded by the expected impact of messages posted by this author). Content and user geolocation data (T2.4) and the longitudinal models of users, trustworthiness and influence (T3.4) will underpin these visualisations.

The results of this task will be integrated in D5.2.2 (D5.2.2).

T5.4 PHEME Visualisation Dashboard (M19 - M36) - MOD, ATOS

The dashboard will display an ensemble of tightly coupled views (T5.1, T5.2, T5.3), allowing users to explore the veracity intelligence extracted by the content analytics methods from WP2, WP3, and WP4. The dynamics and complexity of the chosen domains (health, journalism) will need a granular and scalable update mechanism based on a general coordination model that can handle any number of linked views. Whenever views need to communicate to specific PHEME content analytics services, in order to obtain new data, this will be achieved via the integrated server-side system architecture (T6.3). The dashboard will include media maps to reveal supportive and critical voices and observe the rise and decay of rumours across media and languages; rumour and knowledge landscapes to reveal the context of emerging rumours; impact maps of the spread of phemes in social networks and spheres of influence.

The initial prototype will be deliverable D5.2.1. and then the final release in D5.2.2 (D5.2.1, D5.2.2).

T5.5 Usability Evaluation (M20 - M36) - MOD, SWI, KCL

The two use cases of WP7 and WP8 will provide excellent opportunities to improve the interface in rapid cycles of evolutionary development. The evaluation activities will put special emphasis on the visualisations' dynamic aspects. Usability inspections are low-overhead methods to analyze and assess the interface, to ensure effective results (e.g. risk reduction and value increase). PHEME will conduct heuristic evaluation, where a team of experts systematically investigates the interface design against recognised usability principles (= "heuristics"). The evaluation will be performed periodically during the dashboard design (M21) and implementation phases (M26), so that improvements can be integrated into the prototype early in the development cycle. Later project stages focus on summative usability evaluation (M33 onwards). Formal summative evaluation experiments with test users will collect quantitative performance measurements (e.g., time required to successfully complete a given task) and statistically analyse the collected data. We will also measure scalability in terms of number of threads (in the case of dialogs) and nodes (in the case of network data). The following baselines will be used: the current ThreadVis open source library and the dashboard of the Media Watch on Climate Change.

All usability evaluation results will be reported in D5.3 (D5.3).

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
3	MOD	39.00
5	ATOS	6.00
6	KCL	2.00
7	iHUB	4.00
8	SWI	4.00
	Total	55.00

WT3: Work package description

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D5.1.1	Open-source Visual Analytics Tools: Initial Prototype	3	10.00	P	PU	12
D5.1.2	Open-source Visual Analytics Tools: Final Version	3	10.00	P	PU	22
D5.2.1	PHEME Visual Dashboard: Initial Prototype	3	12.00	P	PU	24
D5.2.2	PHEME Visual Dashboard: Final Version	3	17.00	P	PU	36
D5.3	Usability Evaluation Report	3	6.00	R	PU	36
			Total	55.00		

Description of deliverables

D5.1.1) Open-source Visual Analytics Tools: Initial Prototype: The initial prototype (D5.1.1) will deliver the extension of the ThreadViz open-source library with visualisations of discussions over time (T5.1). It will be released early, to enable integration and testing. [month 12]

D5.1.2) Open-source Visual Analytics Tools: Final Version: The final, updated version of D5.1.1 will add the network-based visualisations, showing the dynamics of information exchanges and content propagation across explicit and implicit social networks. The software will be accompanied by documentation, defining the API and giving usage examples. [month 22]

D5.2.1) PHEME Visual Dashboard: Initial Prototype: The initial prototype (v 0.5) of this software deliverable will integrate the visualisation tools from D5.1 into a multiple-coordinated views dashboard, as described in T5.4. The deliverable will comprise a public, web-based prototype, accompanying documentation, and a public API. [month 24]

D5.2.2) PHEME Visual Dashboard: Final Version: The final version of the PHEME Visual dashboard will extend the functionality from D5.2.1, to include the spatio-temporal visualisations from Task T5.3 and also provide updated versions of the tools from D5.1, based on the first integrated usability evaluation of v1 of the dashboard, carried out at M26. The deliverable will comprise a public, web-based prototype, accompanying documentation, and a public API. [month 36]

D5.3) Usability Evaluation Report: This report details the results of the usability evaluation experiments, including those from the first evaluation cycle at M26, and the second one (from M33). [month 36]

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS3	First Development and Delivery Cycle	1	12	Year 1 Completion, Progress Review, and Y2 Planning
MS5	Second Development and Delivery Cycle	1	24	Year 2 Completion, Progress Review, and Y3 Planning

WT3: Work package description

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS7	Project Completion	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP6	Type of activity ⁵⁴	RTD
Work package title	Scalability, Integration, and Evaluation		
Start month	1		
End month	37		
Lead beneficiary number ⁵⁵	5		

Objectives

The objectives of WP6 are to deliver:

- Data collection tools for capturing content over time, from diverse media and social networks
- Scalable integration of content, metadata, social network data, extra-linguistic knowledge, and extracted information
- Integrated content analytics methods and their coupling to the PHEME visual dashboard
- Quantitative evaluation of the social semantic intelligence methods, developed in WP2, WP3, and WP4, as well as comparison against other approaches.

Description of work and role of partners

T6.1. Data Collection Tools (M1 – M9) – ATOS

User-generated content and social networks are highly dynamic, thus the PHEME data collection tools will track over time the content created by a given user, exchanges with other users, as well as changes in their profiles and social networks. ATOS already have in place web crawling components, including RSS feeds for downloading new content from web sites over time. Web crawling components from TRENDMINER will also be considered, although they are currently not openly available outside of the project consortium. Task T6.1 will extend these existing web crawling tools with continuously running data collection tools for social networks, e.g. Twitter, patient forums. For example, given a Twitter user handle, the tools will download continuously the user's tweets, their profile, lists, favourites, and followers/followees. If required, the tweets and profiles of the connected users will also be acquired over time. In this way the dynamics of the user-generated content and their network will be captured over time. The results will be stored in the integrated repository of T6.2.

The data collection tools will be released as part of D6.1.1 (D6.1.1).

T6.2. Content and Knowledge Integration (M3-M24) - ATOS, ONTO

Task 6.2 will investigate how PHEME's diverse kinds of content and knowledge can be stored and accessed. This task will deliver a distributed, scalable, and integrated storage, indexing and retrieval component, for textual content (both authoritative and user-generated content), social network data, semantic metadata, and Linked Open Data knowledge. The idea is to provide a set of well-suited storage system ranging from NoSQL databases (i.e. MongoDB, HBase, etc.) for textual content, and a semantic repository for semantic metadata and LOD resources (semantic data and metadata will be stored and accessed via OWLIM - ONTO's highly scalable semantic repository, using the ontological models defined in T2.2). This storage layer will provide access to the raw data. However, to make the data available for querying, the raw data should be processed alongside the extracted social semantic intelligence coming from the integration done in T6.3 of the WP2, 3 and 4 components. This aggregated data will be exposed in an intermediate indexed layer, making use of tools such as Apache Solr (for indexing), or even traditional SQL databases for fast access of aggregated data. The integrated storage layer will power the visual analytics (WP5) and will be populated with social media, news, scientific articles, and other relevant content, as required by each case study (WP7 and WP8).

WT3: Work package description

The initial content and knowledge integration tools will be released as part of D6.1.1. The final versions will be released as part of D6.1.2 (D6.1.1, D6.1.2).

T6.3. System Architecture and Integration (M13 - M37) - ATOS

The various multilingual content analytics tools from WP2, WP3, and WP4 will be integrated in T6.3. The integration will be designed and tested for on-demand scalability, using distributed computing based on MapReduce and the Hadoop framework for batch analysis of historical data. The Storm framework will be used for real-time analysis of incoming content, which is complementary to the Hadoop-based batch processing. ONTO's experience and results from the TRENDMINER real-time stream processing architecture will also be taken into account. The outputs from the content analytics tools (e.g. semantic metadata and associated LOD knowledge; extracted statements, rumours, and contradictions; user authority; and implicit user networks) will be stored in the integrated repository (T6.2). The architecture will be designed to run efficiently on a powerful server, and come with a web-service API to enable the web-based visualisation dashboard (WP5) to call on demand the content analytics services and to query the content and semantics store (T6.2).

A number of architecture and integration metrics will be applied, including complexity, criticality, reliability, message rates, network load, response time, CPU usage, memory usage, and effective throughput.

Interim results will be released as part of D6.1.2 at M24, followed by the final version of the PHEME framework in D6.1.3 at M37 (D6.1.2, D6.1.3).

T6.4 Accuracy and Scalability Evaluation (M14 - M37) – USAAR, USFD, ONTO, ATOS

The datasets created in Task 2.1 will be used for iterative development and parameter tuning of the PHEME content analytics methods from WP3 and WP4, as well as for testing their integration into a processing pipeline. The first iteration will start from month 14 to help improve the methods for their M24 deliverables in WP3 and WP4, whereas the second iteration will start from month 26 to underpin the final WP3 and WP4 deliverables. Evaluation on unseen data in the first iteration will be based on half of the use case-specific corpora (D7.2 and D8.2), while the rest will be used in the second iteration from M26. The scalability of the integrated tools (T6.3) will be evaluated on the large-scale datasets collected in PHEME, as well as on historical data (e.g. USFD's 10% Gardenhose Twitter feed, USAAR's APA corpus). The usability of the visual analytics tools (WP5) is evaluated in T5.5, whereas T7.3 and T8.3 will trial the use case prototypes with end users.

Results from the first evaluation cycle will be reported in D6.2.1, whereas results from the second cycle - in D6.2.2 (D6.2.1, D6.2.2).

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
1	USFD	2.00
2	USAAR	3.00
4	ONTO	5.00
5	ATOS	56.00
7	iHUB	4.00
8	SWI	2.00
	Total	72.00

WT3: Work package description

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D6.1.1	PHEME Integrated Veracity Framework - v 0.5	5	18.00	P	CO	12
D6.1.2	PHEME Integrated Veracity Framework - v.1.0	5	24.00	P	PU	24
D6.1.3	PHEME Integrated Veracity Framework - v.2.0	5	22.00	P	PU	37
D6.2.1	Evaluation report - Interim Results	1	3.00	R	PU	21
D6.2.2	Evaluation report - Final Results	2	5.00	R	PU	37
Total			72.00			

Description of deliverables

D6.1.1) PHEME Integrated Veracity Framework - v 0.5: This software deliverable is the integrated framework for veracity intelligence. It consists of several key components, developed in the respective tasks above. A preliminary version 0.5 will be released early in the project and will contain software tools for continuous data collection from mainstream media web sites, other authoritative sources (e.g. MEDLINE, government web sites), blogs, patient forums, other relevant user-generated content, microblogs, and social networks. Since data collection underpins all technical and use case WPs, it is essential to address it early in the project. [month 12]

D6.1.2) PHEME Integrated Veracity Framework - v.1.0: This software deliverable is the integrated framework for veracity intelligence. It consists of several key components, developed in the respective tasks above. Version 1.0 will be the first integrated prototype of key technologies and tools. It will contain the software tools, implementing the integrated content and knowledge repository developed in T6.2, as well as an API for storage and access for the use case applications in WP7 and WP8. Scalability of the algorithms from WP2 and WP3 will be achieved by utilising Hadoop to process historical content, and Storm or a similar streaming framework for continuous processing of the new, incoming information streams. [month 24]

D6.1.3) PHEME Integrated Veracity Framework - v.2.0: This software deliverable is the integrated framework for veracity intelligence. It consists of several key components, developed in the respective tasks above. Version 2.0 will integrate further results from WP3 and WP4, as well as enable cross-media processing and result storage, as well as integration of the algorithms from WP5. The framework will be made open-source, subject to tool licensing restrictions. [month 37]

D6.2.1) Evaluation report - Interim Results: Report detailing the results of T6.4. This deliverable will cover the first internal evaluation cycle. [month 21]

D6.2.2) Evaluation report - Final Results: Report detailing the results of T6.4. This public and final version will be an update on D6.2.1 and will cover the second internal evaluation cycle. [month 37]

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS3	First Development and Delivery Cycle	1	12	Year 1 Completion, Progress Review, and Y2 Planning

WT3: Work package description

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS5	Second Development and Delivery Cycle	1	24	Year 2 Completion, Progress Review, and Y3 Planning
MS7	Project Completion	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP7	Type of activity ⁵⁴	RTD
Work package title	Veracity Intelligence in Patient Care		
Start month	1		
End month	38		
Lead beneficiary number ⁵⁵	6		

Objectives

The aim of this work package is to turn the project technologies toward practical applications in the healthcare domain, to enable clinicians, public health professionals, and health policy makers to analyse the high volume, high-variety, and high-velocity internet content for emerging medically-related rumours, patient complaints, and other health-related related misconceptions. This analysis can in turn be used (i) to develop educational materials for patients and the public, by addressing concerns and misconceptions, and (ii) to link to analysis of the Electronic Patient Record (EPR).

The objective of this work package is therefore to carry out research towards a PHEME-based platform for the medical domain (specifically mental health care at our partner KCL's BRC unit at the South London and Maudsley NHS trust), for multi-channel media monitoring, extraction, verification, and visualisation of automatically extracted knowledge across media and languages. PHEME will create the application and resources needed to monitor and inter-link social media and patient records, specifically targeted at health professionals and meeting the quality required by both the profession and its regulators. This case study provides the integration of PHEME's technology into a (hospital-based) patient records application, and methodological and user verification for the ultimate goals of monitoring health related rumours and misinformation in social media.

The outcomes will be (1) an analysis of medically-related social media, including veracity; (2) a tagged corpus of medically-related social media; (3) a PHEME-based medical platform, as described below; (4) an evaluation of this application.

The PHEME medical application will be a web-based platform to enable monitoring social media for mental health related events, and the linkage of these to the electronic patient record, at a population level. For example, geospatial incidence of a new form of substance abuse or a new environmental stressor, may be linked to secondary care statistics derived from the electronic patient record. This will enable better micro-level characterisation of a catchment area, and better response to emerging trends in population mental health.

Description of work and role of partners

Task 7.1: Case study design, and application mock ups (M1 - M18) – KCL, USFD, USAAR

This task will identify and outline test use cases and the requirements and determine a blend of algorithms that will meet the objective of monitoring medically-related rumours, misconceptions, misinformation and self-reported health-related data in social media, and linkage of these to the EPR. The task will focus on achieving various quality dimensions within both the authoritative resources and the streamed social media, in order to ensure uptake within healthcare applications. In addition to this quality dimension, we must address others specific to social media. First, social media may be dirty, in that it may contain spam or otherwise irrelevant material. Second, the veracity of social media content is unproven (unlike traditionally published material), and it often contains rumours and misinformation.

The results of Task 7.1 will be reported in D7.1 (D7.1).

WT3: Work package description

Task 7.2: Test data collection and corpus annotation (M1 – M24) – KCL, USFD

The availability of a properly tagged corpus has previously allowed the quantitative benchmarking of software solutions aiming to automatically detecting potential drug discussions (e.g. [Toldo et al; 2013]) and we will extend this approach to the identification of more general medical discussions, and particularly those relevant to mental healthcare.

Currently there are no appropriately constructed corpora of medically-related rumours in social media. This necessitates its development in PHEME, to enable quantitative benchmarking of the project's new rumour detection and veracity analysis tools against the state-of-the-art methods (Task 7.3).

The results of Task 7.2 will be released first as D7.2.1, followed by a larger dataset in D7.2.2 (D7.2.1, D7.2.2).

Task 7.3: Application construction (M19-M38) – KCL, USFD, MOD, ONTO, USAAR

The social media data collected will be analysed with the content analytics tools from WP2, WP3 and WP4 and visualised using the PHEME visualisation dashboard (WP5). The rumour and misinformation detection algorithms (WP3 and WP4) will be adapted to make use of Life Science linked data for reasoning (specifically, ONTO's Linked Life Data resource of close to 10 billion triples). The PHEME visualisation dashboard (Task 5.4) will be customised for this case study, to support plotting over time, by geolocation, by linkage to the aggregated patient record, and by allowing for real-time observation and evaluation and feedback of the results as they occur. This will enable a user to capitalise on successes and mitigate negative aspects. The web-based application design will allow for delivery to users anywhere and across a range of devices and platforms and is therefore the most appropriate form of development.

The application prototype will be delivered as part of D7.3 (D7.3).

Task 7.4: User-based Evaluation (M25 - M38) – KCL, USFD, USAAR

This task will involve the deployment of the application with a set of analysts within KCL, as well as other relevant users (in particular, outreach towards policy makers and government agencies will be sought, through KCL's already established position in health policy formation and USFD's ongoing cooperation with the UK Health Protection Agency and the WHO). In particular, KCL is associated with the UK NHS centre for biomedical research in mental health and leads initiatives to develop mental health informatics at a national level. The KCL team includes funded academic psychiatrists, epidemiologists, social scientists, statisticians and informaticians

Qualitative evaluation and feedback will be elicited in an iterative fashion, to aid in the development of the interfaces to the data and to test its effectiveness on real-world datasets. Quantitative evaluation will be carried out using the corpus developed in Task 7.2 (which will be adjusted as necessary according to results from this task in the early stages of each iteration of its definition and construction).

The application constructed in Task 7.3 will be evaluated using the corpus collected in Task 7.2. For each mental health case study, required Electronic Health Record outcomes will be defined and validated. The distribution of these will be measured prior to any social media linkage. This will act as a baseline. Distribution will be re-measured at each phase of development and use, in order to quantify the change brought about by PHEME. In addition to standard metrics of precision and recall, we will, where appropriate, use the closely related diagnostic test metrics of sensitivity and specificity. Additionally, system usability will be measured with the SUS scale, and appropriateness rated using Likert-scaled questionnaires. Baselines for both of these will be measured from either existing practice, or from initial application versions. We will also measure uptake amongst mental health researchers.

All evaluation results will be made public, as part of the D7.3 report (D7.3).

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
1	USFD	4.00
2	USAAR	4.00
3	MOD	3.00

WT3: Work package description

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
4	ONTO	5.00
5	ATOS	4.00
6	KCL	35.00
Total		55.00

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D7.1	Requirements and design documents	6	6.00	R	PU	4
D7.2.1	Annotated Corpus - Initial Version	6	6.00	O	RE	14
D7.2.2	Annotated Corpus - Final Version	6	6.00	O	PU	24
D7.3	Healthcare Application Prototype and User Evaluation results	6	36.00	P	PU	38
Total			54.00			

Description of deliverables

D7.1) Requirements and design documents: D7.1 will contain the proposed scope of the application, mock-ups of interfaces and planning for server capacity and the kinds of content analytics and visualisations needed for the health domain (and their difference from the generic PHEME content analytics tools). Specifically, the way in which linked data in the life sciences may be used to aid rumour and misconception in health-related social media will be delineated, and ways in which it may be linked to the patient record. [month 4]

D7.2.1) Annotated Corpus - Initial Version: A corpus of de-identified, quality assured, manually annotated texts for social media monitoring, developed to an established, rigorous methodology to meet quality requirements. A preliminary, smaller set of data will be released at M14, to aid the development and evaluation of the technical WPs and the first evaluation cycle in WP6. [month 14]

D7.2.2) Annotated Corpus - Final Version: A corpus of de-identified, quality assured, manually annotated texts for social media monitoring, developed to an established, rigorous methodology to meet quality requirements. The complete deliverable will contain the larger gold-standard dataset. The newly added data will be used as gold-standard data for evaluation of the technical algorithms in the second evaluation iteration, as described in T6.4. [month 24]

D7.3) Healthcare Application Prototype and User Evaluation results: This deliverable will have two parts: the healthcare application prototype and a report on quantitative and qualitative evaluation of the prototype with the users. We will start iterative testing from month 24 as well as gathering user feedback for integration with the application over that time. Therefore, following the agile software development paradigm, there will be continuous, short implementation and evaluation cycles, thus ensuring that evaluation results are fully integrated and impact prototype implementation. The M36 deliverable will report on the final outcomes, as well as on key earlier problems encountered in user evaluation and how they were addressed in the final prototype version. [month 38]

WT3: Work package description

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS1	Inception	1	4	Inception Check & Risk analysis and plan maintenance
MS3	First Development and Delivery Cycle	1	12	Year 1 Completion, Progress Review, and Y2 Planning
MS5	Second Development and Delivery Cycle	1	24	Year 2 Completion, Progress Review, and Y3 Planning
MS7	Project Completion	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP8	Type of activity ⁵⁴	RTD
Work package title	Digital Journalism Use Case		
Start month	1		
End month	39		
Lead beneficiary number ⁵⁵	8		

Objectives

The objective of this work package is to turn the project technologies toward practical applications in the digital journalism domain. More specifically, WP8 will support newsroom journalists involved in the news gathering and verification process through automatic methods for detecting rumours and misinformation in user-generated content (UGC); modelling authority in social networks; and tracking information diffusion across social and traditional media.

The specific objectives are to:

- Provide the expertise needed to understand the requirements of newsroom journalists towards PHEME's automated methods for collection, verification, aggregation, and visualisation of socio-semantic intelligence, operating in multiple languages and across media.
- Create an experimental digital journalism application, which supports collaborative, semi-automatic monitoring, analysis, and visualisation of social semantic intelligence, gathered from interlinked news, online content, and social networks.
- Provide a tightly integrated base of data and carry out user evaluation with newsroom journalists.

Description of work and role of partners

The primary tasks of this package break down across the expertise of the partners with SWI handling the requirements gathering and manual data annotation required for training the testing the algorithms as well as being early adopters and end-user evaluators in the final phases. The integrated PHEME content analytics and veracity services (WP6) will be connected to iHub's SwiftRiver open-source platform. ONTO will carry out domain-specific customisations of PHEME's content analytics tools, to meet the requirements of newsroom journalists; identify and tailor relevant Linked Open Data resources; and ensure compatibility with existing standards for news and content such as IPTC's G2 standards and Schema.org vocabularies.

Task 8.1 Requirements gathering, use case design and interface mock-ups (M1 – M6) - SWI, UWAR, iHub, ONTO

The objective of this task is to identify and outline cross-lingual digital journalism use case scenarios at SWI. UWAR will also gather the requirements through observational fieldwork with the user group (see Section 2.3.5 for a full list of participants), which includes all key stakeholders: (i) traditional news media (e.g. BBC World Service, the Guardian); (ii) news providers (e.g. the Press Association); (iii) open news initiatives (e.g. Open society media programme, the Knight Mozilla Open News program). Multilinguality is a key requirement (e.g. SWI already target 9 languages), so the journalist dashboard (task 8.3) will be designed for easy expansion with content analytics for additional languages. The task will also create user interface mock-ups which can be used to most fruitfully monitor, analyse, and verify information coming from multiple channels. These mockups will be used then in the journalism dashboard (Task 8.2) as well as in the generic PHEME visualisation dashboard (T5.4).

All results will be reported in D8.1 (D8.1).

Task 8.2. Journalism Corpus Collection and Annotation (M1 – M18) – SWI, USAAR, UWAR

WT3: Work package description

The journalism corpus (D8.2) will be bootstrapped from corrections, clarifications, and retractions in mainstream media; claims and their truthfulness on fact-checking sites (e.g. factcheck.org); Twitter hashtags (e.g. #false); and Hearst-like patterns, involving words like “hoax”, “conspiracy”, “rumour”, and phrases like “<Entity> claims (that) alleges X”. Annotation schemas for the targeted linguistic phenomena will be defined (i.e., rumours, claims, contradictions), e.g. statements will be labelled as affirming, denying, questioning, or not about a given rumour (Mendoza et al, 2010). The annotation will be created partly by SWI and USAAR, and partly through crowdsourcing (e.g. CrowdFlower and MOD’s Facebook linguistic games with a purpose). Additionally, PHEME will recruit volunteer annotators from within the NLP research community, by offering researchers early access to the new corpora, in exchange for annotation time (D8.2).

Task 8.3. Open-source digital journalism showcase (M13 - M38) - iHub, ATOS, ONTO

This task will create a digital journalism showcase (D8.3) by extending iHub’s open source SwiftRiver platform with the PHEME integrated tools for veracity and socio-semantic intelligence (D6.3).. At present SwiftRiver supports users working collaboratively to curate and filter real-time content from multiple channels, including Twitter, SMS, email and RSS feeds. The plugin architecture of SwiftRiver will enable its configuration and customisation, to the specific requirements of this use case.

The PHEME content analytics tools will significantly enhance SwiftRiver, resulting in an open-source digital journalism dashboard, supporting the cross-linking, verification, aggregation, and visualisation of multilingual media streams. Moreover, SwiftRiver’s built-in support for crowdsourcing and collaborative data curation will enable journalists to correct text processing mistakes, which can then be fed back as new training data to the PHEME learning algorithms for veracity detection (D8.3).

Task 8.4: Iterative Evaluation (M20 - M39) – SWI, ONTO, iHub.

This task will involve the deployment of the application with set of newsroom journalists from SWI, who can provide feedback in an iterative fashion to aid in the development of the algorithms and to test effectiveness in a real-world situation. The evaluation will also involve the user group, described in T8.1. The steps of the evaluation will be: (i) Definition of the qualitative evaluation framework; (ii) SWI-based evaluation (M20 and M34); (iii) User-group evaluation (M25 and M36) (iv) Quantifying the improved efficiency and time saved by newsroom journalists by using PHEME’s automated tools.

In addition, iHub will carry out quantitative data assessments between previous data generated through older iterations of crowdsourcing methodologies against new data from PHEME tools and the relationship of these filtered channels to largely manual processes of similar subject matter and region. Additionally, iHub will examine the correlation of large datasets of incoming data against a focus group of journalists familiar with our platforms and crowdsourcing methodologies to measure the impact of PHEME tools in their workflow and output. This measurement will encompass several key metrics including number of reports received, processed, and published within a set amount of time, complimented by interviews with said journalists to measure their experience using both raw data and PHEME-assisted data.

The users involved in the testing and evaluation will be primarily newsroom journalists and editors. SWI has 75 journalists working in 10 languages, not including translators and freelancers (e.g. English, German, French, Russian). SWI has an audience of 715, 710 Swiss people living abroad. Users will also be drawn from the associated partners: BBC, BBC World Service, the Guardian, SWR (Sudwestrundfunk), Open Society Foundations Media Program. iHub will involve also some of its customers, especially those using their SwiftRiver platform.

All evaluation results will be reported in D8.4 (D8.4).

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
2	USAAR	2.00
4	ONTO	4.00
5	ATOS	4.00

WT3: Work package description

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
7	iHUB	18.00
8	SWI	27.00
9	UWAR	3.00
Total		58.00

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D8.1	Requirements and Use Case Design document	8	6.00	R	PU	4
D8.2	Annotated Corpus of Newsworthy Rumours	8	14.00	O	PU	18
D8.3	Digital journalism prototype	7	20.00	P	PU	30
D8.3.1	Digital journalism prototype (v2)	7	8.00	P	PU	38
D8.4	Evaluation results: validation and analysis	8	10.00	R	PU	39
Total			58.00			

Description of deliverables

D8.1) Requirements and Use Case Design document: This deliverable will detail the requirements of newsroom journalists, defined the scope of the digital journalism application, and provide mock-ups of interfaces and definitions of the domain adaptations required for the WP2, 3, 4, and 5 methods. [month 4]

D8.2) Annotated Corpus of Newsworthy Rumours: A corpus of the types of social media, blog, and news content that will be analysed and verified will be gathered and annotated for the project. Annotations will include linking UGC content to authoritative news and marking entities mentioned (e.g names of people and locations), true and false information, and contradictions. A preliminary dataset will be released internally to the consortium, at M12 to enable the technical partners to evaluate their algorithms. At M18 this dataset will be enriched with more detailed rumour, authority, and other social network and user data (e.g. geolocation). [month 18]

D8.3) Digital journalism prototype: This is the final output of task 8.3 - the open-source application assisting newsroom journalists with content verification and curation. The content analytics and visualisation tools from WP2, 3, 4, and 5 will be customised for the needs of this use case. Following the user evaluation results (D8.4), the prototype will be updated for final release at M38. [month 30]

D8.3.1) Digital journalism prototype (v2): This prototype deliverable will be an update on D8.3, containing the final developments in the open-source journalism dashboard, as well as some new automatic fact-checking tools from ONTO, developed in response to journalist feedback. D8.3. v.2 will be a joint work of iHUB, ONTO, and ATOS. It will also include the results of D4.3.2 and the optimised Kafka-based data collection and integration pipeline (D6.1.3). [month 38]

D8.4) Evaluation results: validation and analysis: This deliverable includes the judgement of the SWI journalists and user groups which start testing M24, as well as suggestions from them incorporated into the application over that time for deployment in the second version. [month 39]

WT3: Work package description

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS1	Inception	1	4	Inception Check & Risk analysis and plan maintenance
MS4	Project Mid-Point Evaluation	1	18	M18 Quantitative and user-based, qualitative evaluation results
MS6	Project Pre-Completion Evaluation	1	30	Preparation for final technical releases and user experiments
MS7	Project Completion	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT3: Work package description

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

One form per Work Package

Work package number ⁵³	WP9	Type of activity ⁵⁴	RTD
Work package title	Dissemination and Exploitation		
Start month	1		
End month	39		
Lead beneficiary number ⁵⁵	5		

Objectives

This work package will plan and carry out business modelling, dissemination and exploitation activities, using appropriate tools, methods and channels. The main focus of the work package is on activities aimed at preparing the successful future exploitation of project results and achievements (e.g. market analysis, business planning, license definition). Dissemination activities are targeted at the scientific community, the business use cases communities (media and life sciences), text analytics providers, and the general public.

Description of work and role of partners

Task 9.1 Dissemination activities (USFD, All)

The objective of this task is to identify, define, co-ordinate and organise the activities to be performed for the promotion and dissemination of the project, including addressing:

- **Promotion:** carrying out promotional activities for the dissemination of project work and achievements (examples include, but are not limited to, the production of the project factsheet (D9.1), project brochures, newsletters, project videos, the project web presence (D9.2), etc.). The consortium commits also to issuing press releases within one month of the start of the project (ATOS will create the original in Spanish, to be translated by partners into English (USFD), Bulgarian (ONTO), German (USAAR), French (SWI)). Further press releases will be issued on all important milestones and events. More specifically, ONTO will issue a press release regarding the PHEME ontologies and knowledge repository; USFD - on the public release of the open-source veracity intelligence tools in WP4; MOD - on the public release of the PHEME visual dashboard in WP5; ATOS - on the public release of the PHEME framework in WP6; KCL - on the healthcare application and its user evaluation; iHub - on the integration of PHEME tools in the SwiftRiver platform and the release of the journalism application. A PDF scan of the collection of such press releases will be included in the dissemination and exploitation plan/report deliverables annually.
- **Participation:** the consortium will be informed regularly about relevant events (workshops, national and international events / conferences, coordination with external organisations, special interest groups etc.). Then, individual organisations, a combination of partners or the entire consortium will participate, representing the project or individual parts thereof.

Scientific conferences to be targeted are:

- Language Processing and Computational Linguistics: (E)ACL, COLING, CONLL, EMNLP
- Social Media and social science: ICWSM, WSDM, WebScience, Social Computing
- Reasoning and Linked Data: WWW, ISWC, ESWC, EKAW, Web Science
- Intelligent User Interfaces and Information Visualisation: CHI, IUI, Inf. Visualisation

Complementing this, there will be a programme of papers and articles in the information technology and general business literature, as well as presentations at IT and business seminars and conferences. We will also target key stakeholder groups for the two case studies. Again, this will be directed towards European community and associated countries. Conferences targeted will include:

- General IT, language technologies, knowledge management: Online Information, SemTech, Text Analytics World, Text Analytics Europe, European Semantic Technology Conference, European Data Forum, SemTech, CeBIT, ICT Event, PopTech, OSCON, ICCM, B2BSD (Austrian), eDay (Austrian), EMC Forum, IDC Forum;

WT3: Work package description

-- Biomedical Informatics: IHI, ICCABS, IEEE BIBM, iHealth
-- Digital media and journalism: Online Information, FT Digital Media, DigitalMedia Europe

- Community building: the partners will actively contribute to the building up of an active user community, starting from the use case focus groups, existing collaborators, USFD's open-source GATE NLP user community, iHub's open-source SwiftRiver and Ushahidi platforms, and growing outwards. It will be aimed at involving a critical mass around the activities of the project and its initiatives. Key tools to be used in the process of community building will be the social networks such as LinkedIn, Wikipedia, and Twitter.

The beneficiaries undertake – when invited – to contribute to and participate in focused concertation actions, themed seminars or special interest groups.

All presentations, contributions and publications even partially funded by the project shall include the project logo and prominently acknowledge PHEME grant funding (611233).

Details of all publications even partially funded by the project shall be uploaded to some specific and agreed bibliographic social networks such as <http://www.citeulike.org/>, <http://www.mendeley.com/> or <http://www.bibsonomy.org/>. Every such publication must be tagged with an agreed tag specific to the project, for example "PHEME- 611233".

Whenever such bibliographic social networks allow for a catalogue of publications to be retrieved by tag or published as an RSS feed, the project's web site should expose such a catalogue.

All presentation materials for which this is appropriate shall be published on the project's web site under a Creative Commons licence (<http://creativecommons.org/>) or another appropriate license.

All open source software produced by the consortium shall be published on publicly available software repositories such as <http://sourceforge.net/>, <http://github.com/> or <http://osor.eu>.

All data sets for which this is appropriate shall be published on the project's web site under a Creative Commons licence (<http://creativecommons.org/>) or another appropriate license.

Dissemination and impact will be measured in multiple ways, including access statistics for the PHEME web site and the download sections of the software tools; number of published papers; number of organised events (both technical and industry-oriented); presentations given and stakeholders contacts; and social network presence (D.9.1, D9.2, D9.3, D9.4).

Task 9.2 Market analysis and exploitation planning (ATOS, All)

In this task, all the issues related to planning exploitation of project results will be addressed. Thus, through a continuous 'market watch', this task will analyse and explore different business opportunities in the targeted markets, which will be turned into particular action in the exploitation plan, viable for commercialising the results after the end of the project. A common exploitation strategy will be decided, but plans for individual exploitable products will also be drawn up, to enable individual partners to commercialise their part of the work separately. Specifying the targeted markets, exploring the competition and identifying potential commercial opportunities, will enable the exploitation activities. The results of these activities will be detailed in a Dissemination and Exploitation Plan that is continuously evolving. Activities towards standardisation of the project results and collaboration with other projects and relevant initiatives will also be explored and coordinated through this task, so that the project can have the best possible impact both in the scientific and commercial communities.

This task will also identify appropriate licenses for all software built/extended in this project, taking into account the licenses of any pre-existing plugins/components. We will also define licensing terms and conditions for crowd-sourced datasets and address any related copyright/IP ownership concerns.

ATOS will coordinate the production of a business plan for the commercial exploitation of the language resources and text analysis services arising from the project. As part of the strategic planning, we will carry out SWOT analysis (Strengths, Weaknesses, Opportunities, and Threats).

The results of these activities will be delivered in D9.3, D9.4, D9.5.1 and D9.5.2 (D9.3, D9.4, D9.5.1, D9.5.2).

WT3: Work package description

Person-Months per Participant

Participant number ¹⁰	Participant short name ¹¹	Person-months per participant
1	USFD	3.00
2	USAAR	2.00
3	MOD	2.00
4	ONTO	2.00
5	ATOS	9.00
6	KCL	2.00
7	iHUB	2.00
8	SWI	2.00
9	UWAR	1.00
Total		25.00

List of deliverables

Deliverable Number ⁶¹	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature ⁶²	Dissemination level ⁶³	Delivery date ⁶⁴
D9.1	Project Fact Sheet	1	0.50	R	PU	1
D9.2	Project Website	1	7.50	O	PU	3
D9.3	Dissemination and exploitation Plan (v0.5 M9, v.1 M18)	5	6.00	R	CO	9
D9.4	Dissemination and exploitation Report	1	3.00	R	PU	39
D9.5.1	Market Watch - Initial Version	5	4.00	R	CO	12
D9.5.2	Market Watch - Final version	5	4.00	R	CO	30
Total			25.00			

Description of deliverables

D9.1) Project Fact Sheet: The Fact Sheet will outline the project's rationale and objectives, specify its technical baseline and intended target groups and application domains, and detail intermediate and final outputs. The Fact Sheet will be used by the Commission for its own dissemination and awareness activities throughout the project lifecycle, and will be published on EC and EC sponsored websites. The Fact Sheet has to be maintained and updated until the end of the project; this will be documented in the regular management reporting. [month 1]

D9.2) Project Website: The web site will provide project overviews and highlights; up-to-date information on intermediate and final project results, including public reports and publications as well as synthesis reports drawn from selected confidential material; project events, including e.g. user group meetings, conference and workshop presentations; contact details, etc. The project website address will be <http://pheme.eu> (already reserved by USFD). It will be kept alive for at least 2 years after the end of the project. The project website's first point of access will describe the goals of the project in simple jargon free language. The project's website shall contain an RSS-enabled news or blog section. This section should be used to advertise project related events, to describe its progress for an interested but not specialised public; to comment on how societal or technology developments in the world at large demonstrate the importance of or open opportunities for the technologies developed under the project. The project's main website shall prominently indicate a link to the repositories of

WT3: Work package description

open source software produced in the project and, whenever possible, download statistics. All open source components published shall be extensively documented by means of textual documents and screencasts of professional quality illustrating how to download, install and operate the components in question. Documentation manuals and screencasts shall be specifically identified as project deliverables and prominently published on the project's website. [month 3]

D9.3) Dissemination and exploitation Plan (v0.5 M9, v.1 M18): This deliverable will detail the plan for the project dissemination strategy to be adopted throughout the project lifetime and will contain first exploitation ideas and action plan. It will be updated at M18, to reflect the findings of the market watch. [month 9]

D9.4) Dissemination and exploitation Report: This deliverable will provide a report on the dissemination activities and will contain a section describing the consortium intentions and activities for the exploitation of the project results. It will prepare for the successful exploitation of project results towards the end, as well as in the post-project period. [month 39]

D9.5.1) Market Watch - Initial Version: This report will collect information on market opportunities and competitors. First version will be released at M12, then updated at M30. [month 12]

D9.5.2) Market Watch - Final version: This report will collect information on market opportunities and competitors. This will be the final, updated version of D9.5.1. [month 30]

Schedule of relevant Milestones

Milestone number ⁵⁹	Milestone name	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS1	Inception	1	4	Inception Check & Risk analysis and plan maintenance
MS3	First Development and Delivery Cycle	1	12	Year 1 Completion, Progress Review, and Y2 Planning
MS4	Project Mid-Point Evaluation	1	18	M18 Quantitative and user-based, qualitative evaluation results
MS5	Second Development and Delivery Cycle	1	24	Year 2 Completion, Progress Review, and Y3 Planning
MS6	Project Pre-Completion Evaluation	1	30	Preparation for final technical releases and user experiments
MS7	Project Completion	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT4: List of Milestones

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

List and Schedule of Milestones

Milestone number ⁵⁹	Milestone name	WP number ⁵³	Lead beneficiary number	Delivery date from Annex I ⁶⁰	Comments
MS1	Inception	WP1, WP7, WP8, WP9	1	4	Inception Check & Risk analysis and plan maintenance
MS2	Initial Data and Requirements Analysis	WP1, WP2	1	6	
MS3	First Development and Delivery Cycle	WP1, WP2, WP3, WP5, WP6, WP7, WP9	1	12	Year 1 Completion, Progress Review, and Y2 Planning
MS4	Project Mid-Point Evaluation	WP1, WP2, WP3, WP4, WP8, WP9	1	18	M18 Quantitative and user-based, qualitative evaluation results
MS5	Second Development and Delivery Cycle	WP1, WP3, WP5, WP6, WP7, WP9	1	24	Year 2 Completion, Progress Review, and Y3 Planning
MS6	Project Pre-Completion Evaluation	WP1, WP4, WP8, WP9	1	30	Preparation for final technical releases and user experiments
MS7	Project Completion	WP1, WP4, WP5, WP6, WP7, WP8, WP9	1	39	Completion of all deliverables, achievement of key indicators, dissemination results and exploitation planning, and final project review

WT5: Tentative schedule of Project Reviews

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

Tentative schedule of Project Reviews

Review number ⁶⁵	Tentative timing	Planned venue of review	Comments, if any
RV 1	13	Luxembourg	Period 1 Review
RV 2	25	Luxembourg	Period 2 Review
RV 3	40	Luxembourg	Final Review

Project Effort by Beneficiary and Work Package

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

Indicative efforts (man-months) per Beneficiary per Work Package

Beneficiary number and short-name	WP 1	WP 2	WP 3	WP 4	WP 5	WP 6	WP 7	WP 8	WP 9	Total per Beneficiary
1 - USFD	12.00	12.00	28.00	18.00	0.00	2.00	4.00	0.00	3.00	79.00
2 - USAAR	3.00	6.00	15.00	18.00	0.00	3.00	4.00	2.00	2.00	53.00
3 - MOD	1.00	0.00	14.00	0.00	39.00	0.00	3.00	0.00	2.00	59.00
4 - ONTO	1.00	16.00	0.00	24.00	0.00	5.00	5.00	4.00	2.00	57.00
5 - ATOS	3.00	0.00	0.00	0.00	6.00	56.00	4.00	4.00	9.00	82.00
6 - KCL	1.00	0.00	0.00	0.00	2.00	0.00	35.00	0.00	2.00	40.00
7 - iHUB	1.00	0.00	0.00	0.00	4.00	4.00	0.00	18.00	2.00	29.00
8 - SWI	2.00	6.00	0.00	2.00	4.00	2.00	0.00	27.00	2.00	45.00
9 - UWAR	1.00	18.00	0.00	0.00	0.00	0.00	0.00	3.00	1.00	23.00
Total	25.00	58.00	57.00	62.00	55.00	72.00	55.00	58.00	25.00	467.00

Project Effort by Activity type per Beneficiary

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

Indicative efforts per Activity Type per Beneficiary

Activity type	Part. 1 USFD	Part. 2 USAAR	Part. 3 MOD	Part. 4 ONTO	Part. 5 ATOS	Part. 6 KCL	Part. 7 iHUB	Part. 8 SWI	Part. 9 UWAR	Total
---------------	-----------------	------------------	----------------	-----------------	-----------------	----------------	-----------------	----------------	-----------------	-------

1. RTD/Innovation activities										
WP 2	12.00	6.00	0.00	16.00	0.00	0.00	0.00	6.00	18.00	58.00
WP 3	28.00	15.00	14.00	0.00	0.00	0.00	0.00	0.00	0.00	57.00
WP 4	18.00	18.00	0.00	24.00	0.00	0.00	0.00	2.00	0.00	62.00
WP 5	0.00	0.00	39.00	0.00	6.00	2.00	4.00	4.00	0.00	55.00
WP 6	2.00	3.00	0.00	5.00	56.00	0.00	4.00	2.00	0.00	72.00
WP 7	4.00	4.00	3.00	5.00	4.00	35.00	0.00	0.00	0.00	55.00
WP 8	0.00	2.00	0.00	4.00	4.00	0.00	18.00	27.00	3.00	58.00
WP 9	3.00	2.00	2.00	2.00	9.00	2.00	2.00	2.00	1.00	25.00
Total Research	67.00	50.00	58.00	56.00	79.00	39.00	28.00	43.00	22.00	442.00

2. Demonstration activities										
Total Demo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

3. Consortium Management activities										
WP 1	12.00	3.00	1.00	1.00	3.00	1.00	1.00	2.00	1.00	25.00
Total Management	12.00	3.00	1.00	1.00	3.00	1.00	1.00	2.00	1.00	25.00

4. Other activities										
Total other	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Total	79.00	53.00	59.00	57.00	82.00	40.00	29.00	45.00	23.00	467.00
--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	---------------

WT8: Project Effort and costs

Project Number ¹	611233	Project Acronym ²	PHEME
-----------------------------	--------	------------------------------	-------

Project efforts and costs

Beneficiary number	Beneficiary short name	Estimated eligible costs (whole duration of the project)						Requested EU contribution (€)
		Effort (PM)	Personnel costs (€)	Subcontracting (€)	Other Direct costs (€)	Indirect costs OR lump sum, flat-rate or scale-of-unit (€)	Total costs	
1	USFD	79.00	417,050.00	1,719.00	42,000.00	275,430.00	736,199.00	581,379.00
2	USAAR	53.00	293,000.00	3,000.00	23,000.00	189,600.00	508,600.00	389,400.00
3	MOD	59.00	280,560.00	4,000.00	37,000.00	190,536.00	512,096.00	387,472.00
4	ONTO	57.00	224,040.00	0.00	37,000.00	156,624.00	417,664.00	315,648.00
5	ATOS	82.00	410,000.00	0.00	55,000.00	123,000.00	588,000.00	303,750.00
6	KCL	40.00	239,187.00	24,000.00	19,846.00	155,419.00	438,452.00	330,239.00
7	iHUB	29.00	243,600.00	0.00	31,127.00	54,945.00	329,672.00	170,056.00
8	SWI	45.00	389,250.00	0.00	41,650.00	86,180.00	517,080.00	268,920.00
9	UWAR	23.00	118,360.00	0.00	20,500.00	83,315.00	222,175.00	169,136.00
Total		467.00	2,615,047.00	32,719.00	307,123.00	1,315,049.00	4,269,938.00	2,916,000.00

1. Project number

The project number has been assigned by the Commission as the unique identifier for your project. It cannot be changed. The project number **should appear on each page of the grant agreement preparation documents (part A and part B)** to prevent errors during its handling.

2. Project acronym

Use the project acronym as given in the submitted proposal. It cannot be changed unless agreed so during the negotiations. The same acronym **should appear on each page of the grant agreement preparation documents (part A and part B)** to prevent errors during its handling.

53. Work Package number

Work package number: WP1, WP2, WP3, ..., WPn

54. Type of activity

For all FP7 projects each work package must relate to one (and only one) of the following possible types of activity (only if applicable for the chosen funding scheme – must correspond to the GPF Form Ax.v):

- **RTD/INNO** = Research and technological development including scientific coordination - applicable for Collaborative Projects and Networks of Excellence
- **DEM** = Demonstration - applicable for collaborative projects and Research for the Benefit of Specific Groups
- **MGT** = Management of the consortium - applicable for all funding schemes
- **OTHER** = Other specific activities, applicable for all funding schemes
- **COORD** = Coordination activities – applicable only for CAs
- **SUPP** = Support activities – applicable only for SAs

55. Lead beneficiary number

Number of the beneficiary leading the work in this work package.

56. Person-months per work package

The total number of person-months allocated to each work package.

57. Start month

Relative start date for the work in the specific work packages, month 1 marking the start date of the project, and all other start dates being relative to this start date.

58. End month

Relative end date, month 1 marking the start date of the project, and all end dates being relative to this start date.

59. Milestone number

Milestone number: MS1, MS2, ..., MSn

60. Delivery date for Milestone

Month in which the milestone will be achieved. Month 1 marking the start date of the project, and all delivery dates being relative to this start date.

61. Deliverable number

Deliverable numbers in order of delivery dates: D1 – Dn

62. Nature

Please indicate the nature of the deliverable using one of the following codes

R = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other

63. Dissemination level

Please indicate the dissemination level using one of the following codes:

- **PU** = Public
- **PP** = Restricted to other programme participants (including the Commission Services)
- **RE** = Restricted to a group specified by the consortium (including the Commission Services)
- **CO** = Confidential, only for members of the consortium (including the Commission Services)

- **Restreint UE** = Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments
- **Confidentiel UE** = Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments
- **Secret UE** = Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments

64. Delivery date for Deliverable

Month in which the deliverables will be available. Month 1 marking the start date of the project, and all delivery dates being relative to this start date

65. Review number

Review number: RV1, RV2, ..., RVn

66. Tentative timing of reviews

Month after which the review will take place. Month 1 marking the start date of the project, and all delivery dates being relative to this start date.

67. Person-months per Deliverable

The total number of person-month allocated to each deliverable.



PART B

COLLABORATIVE PROJECT

B1. CONCEPT AND OBJECTIVES, PROGRESS BEYOND STATE-OF-THE-ART, S/T METHODOLOGY AND WORK PLAN	2
B1.1 CONCEPT AND PROJECT OBJECTIVES	2
<i>B1.1.1 Project Rationale and Vision.....</i>	2
<i>B1.1.2 Project Goal and Objectives</i>	3
<i>B1.1.3 Project Concept.....</i>	6
<i>B1.1.4 RTD Challenges and Innovation</i>	7
<i>B1.1.5 Expected Outcomes</i>	8
B1.2 PROGRESS BEYOND THE STATE OF THE ART	10
<i>B1.2.1 Understanding Rumours and Information Flows in Twitter</i>	10
<i>B1.2.2 Automatic Methods for Detecting Contradictions, Claims, and Misinformation</i>	10
<i>B1.2.3 Spatio-Temporal Grounding</i>	12
<i>B1.2.4 Cross-Media Content Linking</i>	13
<i>B1.2.5 Implicit Information Networks, Trust, and Rumour Spread</i>	14
<i>B1.2.6 Interactive Visual Analytics.....</i>	16
<i>B1.2.7 The PHEME Veracity Intelligence Framework: Scalability and Efficiency.....</i>	17
<i>B1.2.8 Veracity Intelligence for Patient Care.....</i>	17
<i>B1.2.9 Digital Journalism.....</i>	18
B1.3 S/T METHODOLOGY AND ASSOCIATED WORK PLAN	19
<i>B1.3.1 Overall Strategy and General Description.....</i>	19
<i>B1.3.2 Key Performance Indicators.....</i>	21
<i>B1.3.3 Significant Risks and Associated Contingency Plans.....</i>	22
<i>B1.3.2 Timing of Work Packages and Their Components (Gantt Chart)</i>	25
<i>B1.3.3 Graphical Interdependencies between tasks (Pert Chart).....</i>	26
B2. IMPLEMENTATION.....	27
B 2.1 MANAGEMENT STRUCTURE AND PROCEDURES	27
<i>B2.1.2 Management Issues and Procedures</i>	28
B2.2 BENEFICIARIES	30
B 2.3 CONSORTIUM AS A WHOLE.....	41
<i>B2.3.1 Roles in the innovation process.....</i>	42
B 2.4 RESOURCES TO BE COMMITTED.....	44
<i>B2.4.1 Budgetary Analysis.....</i>	44
<i>B2.4.2 Project co-financing.....</i>	46
B3. IMPACT.....	46
B3.1 STRATEGIC IMPACT	46
<i>B3.1.1 Contribution to Expected Impacts as Listed in the Call.....</i>	46
<i>B3.1.2 Impact on the Target User Communities.....</i>	48
<i>B3.1.3 European Dimension.....</i>	49
<i>B3.1.4 Impact through Open Source and Standardisation</i>	49
<i>B3.1.5 Cooperation with and Relationship to European and National Projects.....</i>	50
B3.2 PLAN FOR THE USE AND DISSEMINATION OF FOREGROUND.....	51
<i>B3.2.1 PHEME Dissemination Plans.....</i>	51
<i>B3.2.2 PHEME Exploitation Intentions.....</i>	52
<i>B3.2.3 Management of Knowledge.....</i>	55
<i>B3.2.4 Management of Intellectual Property.....</i>	56
<i>B3.2.5 Open Source Licensing.....</i>	57
B4. ETHICAL ISSUES.....	57
APPENDIX A: LETTERS OF SUPPORT.....	59
APPENDIX B: REFERENCES	64

B1. CONCEPT AND OBJECTIVES, PROGRESS BEYOND STATE-OF-THE-ART, S/T METHODOLOGY AND WORK PLAN

B1.1 Concept and project objectives

B1.1.1 Project Rationale and Vision

From a business and government point of view there is an increasing need to interpret and act upon information from large-volume media, such as Twitter, Facebook, and newswire. However, knowledge gathered from online sources and social media comes with a major caveat – it cannot always be trusted. Rumours, in particular, tend to spread rapidly through social networks, especially in circumstances where their veracity is hard to establish. For instance, during an earthquake in Chile rumours spread through Twitter that a volcano has become active and there was a tsunami warning in Valparaiso (Mendoza *et al.*, 2010). Another example is malicious use of Twitter and other social media during election campaigns to spread disinformation about opposing candidates (Ratkiewicz *et al.*, 2011). Researchers have found that people read untrusted sources for various reasons, the main ones being their interestingness, entertainment value, a friend's online recommendation, or a search engine result (Ennals *et al.*, 2010).

A 2012 report of Pew Internet Research on the future of big data (Anderson and Rainie, 2012) argues that even though by 2020 big data is likely to have transformational effect on our knowledge and understanding of the world, there is also **danger from inaccurate or false information** (called “distribution of harms”).

Social media poses three major computational challenges, dubbed by Gartner the **3Vs of big data: volume, velocity, and variety**. Content analytics methods, in particular, face further difficulties arising from the short, noisy, and strongly contextualised nature of social media. To address the 3Vs of social media, novel language technologies have emerged, e.g. using locality sensitive hashing to detect new stories in media streams (volume), predicting stock market movements from tweet sentiment (velocity), and recommending blogs and news articles based on users' own comments (variety).

PHEME will focus on a fourth crucial, but hitherto largely unstudied, big data challenge: veracity.

This project is about creating the necessary computational apparatus to model, identify, and verify phemes (internet memes with added truthfulness or deception), as they spread across media, languages, and social networks. More specifically, PHEME will investigate models and algorithms for automatic extraction and verification of **four kinds of rumours** and their textual expressions (which we refer to as **phemes**):

- **uncertain information** or **speculation** (e.g. an analyst claiming the Bank of England will raise interest rates at their next meeting),
- **disputed information** or **controversy** (e.g. aluminium may or may not cause Alzheimer's),
- **misinformation** (e.g. misrepresentation and quoting out of context), and
- **disinformation** (e.g. Obama is a Muslim).

Since phemes cannot be understood outside of their social context (e.g. their originator and the affected community/user networks), computational methods need to **bring together three kinds of knowledge**:

- *document-intrinsic* knowledge (e.g. lexical, syntactic, semantic);
- *a priori* knowledge (e.g. world knowledge from Linked Open Data, source trustworthiness); and

- *a posteriori* knowledge from the *social, cross-media, and temporal context* (e.g. who receives information from whom and how, where, and to whom do they pass it on).

In other words, veracity intelligence is an **inherently multi-disciplinary problem**, which can only be addressed successfully by bringing together currently disjoint research on language technologies, web science, social network analysis, and information visualisation. Therefore, to meet the need for automatic veracity intelligence, now is the time to **develop novel, cross-disciplinary social semantic methods for veracity intelligence**, drawing on the strengths of these four disciplines.

The proposed research will set the stage for a new generation of scalable technologies for **discovery, reasoning with, and visualisation of veracity intelligence**, gathered from interconnected media streams, in multiple languages. Dealing with the challenges of discovering and verifying rumours, by bringing together methods from language processing, social network analysis and reasoning against a-priori knowledge (e.g. LOD), is a major contribution that goes well beyond the state-of-the-art.

The research communities, from which the PHEME consortium is drawn, have developed the scientific foundations essential for the new foundational and component-level research envisaged in this project: multilingual LOD-based information extraction, opinion mining and spatio-temporal processing; ontologies, provenance and reasoning systems; Linked Open Data; graph-based information diffusion models; cross-media and multilingual information visualisation tools; big data and text processing with Hadoop and Storm; crowdsourcing; digital journalism; and bioinformatics.

Apart from **healthcare** and **digital journalism** (our two use cases), the commercial partners in PHEME will apply the project results in **other key corporate applications**, including business intelligence, market research, campaign and brand reputation management, customer relationship management, knowledge management, and semantic search. In addition to its high commercial relevance, the project will also **benefit society and citizens** by enabling government organisations (e.g. citizen advice, emergency services) to keep track of rumours and misinformation spreading online. PHEME will create instant feedback loops and shows how digital content is received, understood, and propagated in social networks. By uncovering phemes, it will enable citizens to take informed decisions and act to prevent rumour spread across media.

B1.1.2 Project Goal and Objectives

Research in PHEME will tackle several major bottlenecks in detecting phemes and computing veracity:

- modelling, extracting, and reasoning about **multiple truths** (e.g. disputed scientific hypotheses);
- modelling, extracting and reasoning about the **temporal validity of facts** and the way that affects contradiction detection;
- using **wider interpretation context**, including **cross-media links**, user network graphs, historical user behaviour (e.g. has this user spread disinformation before), information diffusion patterns.
- integrating **uncertain facts and reasoning with large-scale world knowledge**, arising from Linked Open Data (e.g. DBpedia, OpenCyc).

The project will build on and enhance research and technology across four disciplines, into a **computational framework for veracity intelligence**, gathered **across media, languages, and networks**.

In more detail, PHEME will pursue the following scientific and technological objectives:

1. **Develop innovative, multilingual methods for cross-media detection of phemes**, capable of extracting and reasoning about multiple truths (e.g. in controversies) and taking into account *the context in which phemes originated and spread* (e.g. the trustworthiness and influence of their sources, spatio-temporal grounding).

2. **Integrate large-scale, a priori knowledge from Linked Open Data (LOD)** to improve pheme identification methods in specific application domains. For example, the patient care use case will use ONTO's *Linked Life Data*, which is connected to *PubMed* and other authoritative sources.
3. **Model pheme spread dynamics over time, and within and across social networks and media:** PHEME will develop models of rumour and information spread; longitudinal models of trustworthiness and influence of users and information sources; and algorithms that cross-reference and compare facts and rumours across scientific publications, grey literature and social media.
4. **Design novel information visualisation techniques for interactive, geo-temporal pheme and information flow analytics** to support interpretation and decision making: (i) **media maps** to reveal and classify supportive and critical voices and observe the rise and decay of rumours across media and languages; (ii) **pheme and knowledge landscapes** to reveal the context of emerging rumours; (iii) **impact maps** of the spread of phemes in social networks and spheres of influence.
5. **Test and evaluate** the newly developed methods through (i) **quantitative experiments** on gold-standard data, acquired both through traditional domain expert annotation and crowdsourcing; and (ii) **qualitative assessments** in the use cases on *health* and *digital journalism*, involving key stakeholders from two focus groups.

Verifiable measures of the objectives:

- **Foundational research** in Objectives 1 to 4 will be measured by the acceptance of peer-reviewed scientific publications, as well as by uptake of research results by other groups.
- **Component-level research** in Objectives 1 to 4 will be measured by comparing the developed components with the state-of-the-art. This will be done both for performance measures such as speed and efficiency, as well as appropriate measures of the output of the components, such as precision and recall for information extraction. Evaluation through participation in relevant evaluation campaigns, e.g. SemEval tasks on drug-to-drug interactions and Twitter analysis, will be undertaken.
- Objective 5 will be verified through user-centred evaluation, conducted with extensive participation of the end users in each use case, as well as by, in later stages of the project, the impact of the developed technologies on the daily work of the target users. These will be measured by the reduction in the time users need to analyse the incoming media streams; and the increase in the size of media the human analysts in the two case studies are able to monitor on a regular basis.

The two case studies to be addressed in the project have been chosen as domains where users are already engaged heavily in monitoring and analysis of media streams, as well as correlating that to information from authoritative sources (e.g. established news organisations, scientific papers, grey literature). These are also domains where the practices currently employed are largely manual and are therefore becoming increasingly expensive and less efficient to perform, as content volumes continue to grow.

Case Study 1: Veracity Intelligence for Patient Care

The relationship between clinicians and their patients has already been changed by the internet in three waves. First, the provision of pharmaceutical data, diagnostic information and advice from drug companies and from health care providers created a new source for self-directed diagnosis. More recently web 2.0 brought co-creation sites like Wikipedia and patient support forums (e.g. PatientsLikeMe), meaning that a discursive element was added to the didactic material of the first wave. Thirdly, the social media revolution has acted as an accelerant and magnifier to the second wave. It is now possible, under the right circumstances, for groups of patients to go from the appearance of a symptom to world-wide spread of their experience as a meme (or pheme!). A suspected European swine flu outbreak, for example, might be trending on Twitter within hours of its

first suggestion, and long before the veracity of the diagnosis had any chance of proper examination by clinicians.

PHEME's first use case (WP7) will start the process of re-tooling medical information systems to cope with this new context. Two complementary uses of PHEME in the health care domain will be investigated: first as a primary source of data, and second, in combination with electronic patient records (EPRs). In the first use, PHEME will provide rumour intelligence for direct use by clinical and public health practitioners. The ability to spot rumours as they appear (e.g. monitored at a national level for different areas of medical care) could provide daily alerts of problematic cases that are likely to be raised by patients. Timely national media interventions will also be facilitated. For example, cases like the MMR vaccine (which in the UK caused a large-scale controversy) or a speculation for a suspected swine flu outbreak very quickly arrive in medical consultation rooms and the earlier that staff can revise clinical advice and practice, the more effective the patient-doctor interactions will be. In the second complementary use of PHEME, social media analysis will be combined with analysis of the structured data and free text of the EPR, thus linking social media to, and correlating it with, aggregated patient records. This will enable health care practitioners to (i) examine the veracity of social media health topics in the light of clinician-recorded patient encounters and (ii) access information along a social dimension, alongside the usual clinical dimension.

The outcomes of WP7 will be: (1) an analysis of medically-related social media, including veracity; (2) a tagged corpus of medically-related social media; (3) a PHEME-based medical platform; (4) an evaluation of this application. The PHEME medical application (3) will be a web-based platform to enable monitoring social media for mental health related events, and the linkage of these to the electronic patient record, at a population level. For example, geospatial incidence of a new form of substance abuse or a new environmental stressor, may be linked to secondary care statistics derived from the electronic patient record. This will enable better micro-level characterisation of a catchment area, and better response to emerging trends in population mental health.

Case Study 2: Digital Journalism

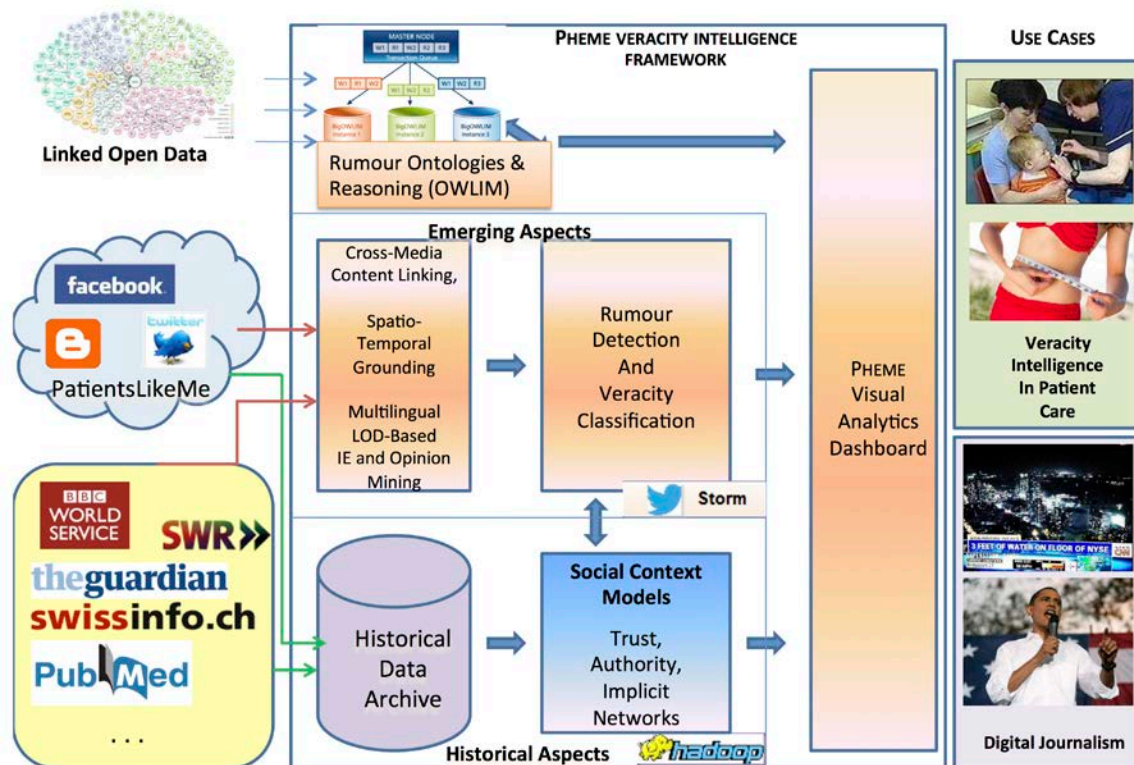
User-Generated Content (UGC), social news, and citizen journalism have changed radically the way news is gathered, analysed, reported, and disseminated by online and traditional news media. An important part of this process has been the convergence between social and broadcast media in the news room (Hänska-Ahy, 2013). The integration of UGC into newsroom routines is key especially in circumstances where there are restrictions in media access (both man-made - e.g. the Iran protests and Arab uprisings - and natural, e.g. disasters like hurricane Sandy) and/or events unfolding at fast pace. Consequently, a newsroom journalist “*is becoming more a facilitator of on and offline news production for media institutions*” (Beckett & Mansell, 2008). However, the unreliable, unverified nature of UGC is a real challenge for journalists, who need to maintain the authoritative status of news media. For instance, the BBC World Service has a dedicated team of journalists who manually verify UGC, coupled with in-house software for multi-platform UGC gathering (Hänska-Ahy, 2013). Implementing such platforms, however, is expensive, while having entirely manual UGC verification and curation is ultimately unsustainable, in particular for smaller news organisations, such as SWI. Multilinguality is yet another challenge, since less than 50% of emerging UGC is in English.

This use case (WP8) will evaluate the new automatic methods for detecting rumours and misinformation; modelling trust and authority in social networks; and tracking pheme diffusion across social and traditional media. PHEME will work closely with journalists from SWI, as well as those from our user group (including the BBC World Service, Südwestrundfunk, and the Guardian) in order to establish their requirements and evaluate how automatic veracity intelligence can facilitate monitoring, verification, and curation of UGC in the news room. Particular emphasis will be placed on **multilinguality**, **cross-media analysis** and **correlation** and **comparison against authoritative sources**. Another key output will be integration of the PHEME content analytics and visualisation components into Ushahidi's open-source SwiftRiver platform (ushahidi.com/products/swiftriver-platform), which is being used, amongst others, by the BBC, the Guardian, Al Jazeera, and the Press Gazette. It supports collaborative filtering and curation of real-time data from multiple media, including Twitter, SMS, email and RSS feeds. The result will be an open-source digital journalism

platform for collaborative, semi-automatic monitoring, analysis, and visualisation of veracity intelligence, extracted from interlinked media.

B1.1.3 Project Concept

Memes are thematic motifs that spread through the internet in ways analogous to genetic traits. *Phemes* add truthfulness and deception to the mix (after ancient Greek PHEME, “an embodiment of fame and notoriety, her favour being notability, her wrath being scandalous rumours” <http://en.wikipedia.org/wiki/Pheme>). The current state-of-the-art in social media analysis can detect memes; this project is focused on modelling, identification, and verification of phemes, as they spread across media, languages, and social networks.



PHEME’s novel **cross-disciplinary methods for veracity intelligence** combine **deep linguistic methods** from text processing; **extra-linguistic, world knowledge** from ontologies (e.g. Linked Open Data); and **graph-based methods** for modelling social context (e.g. implicit information networks, authority, and trust).

The infrastructural underpinning will be a **scalable framework for veracity intelligence** that ensures scalable access and processing of both **large volumes of historical data** (e.g. published news articles, previous tweets) and of a continuous **stream of diverse media, in multiple languages**. The PHEME framework is described in more detail in Section B1.3.1.

PHEME is an **ambitious cross-disciplinary project**. The veracity and variety big data challenges are addressed through novel cross-disciplinary methods, drawing on strong track records in each field: (i) multilingual natural language processing and information extraction (USFD, USAAR); (ii) social informatics (UWAAR); (iii) graph-based methods (USFD, MOD); (iv) LOD, ontologies and reasoning (ONTO); (v) visualisation (MOD, IHUB). The volume and velocity challenges for content analytics in PHEME are addressed through big data processing with Hadoop and Storm (ATOS, ONTO). The patient care use case is underpinned by KCL’s School of Medicine, who are active users of biomedical text mining for patient record analysis. The digital journalism use case involves two complementary industrial partners. IHUB develop two widely used crowdsourcing and collaborative information filtering platforms (Ushahidi for news crowdsourcing and SwiftRiver for information gathering) and will build the veracity digital journalism dashboard. SWI are a news provider covering content in nine languages. SWI will be the end-users giving us the necessary access to newsroom

journalists; journalist verification practices and multilingual news content and associated comments and social media.

In order to ensure that the veracity intelligence algorithms will not fall short on the technologically challenging volume and velocity aspects, as well as being directly applicable to real-world applications, PHEME has been designed with strong industrial participation from the onset. This is also in accordance with the aims of this call, which seeks ambitious scientific research on big data, coupled with high commercial involvement to ensure adoption and transfer into new language technology products.

The focus of **cross-media research** in PHEME is on combining **authoritative sources** (e.g. newspaper articles, transcriptions of news podcasts, scientific papers); **user-generated content** (e.g. forums, blogs); and **social networks** (e.g. Twitter, Facebook, YouTube). In the case of social networks, PHEME will analyse not just the shared content but also the graph structure (who is connected to whom), the user profiles, and the information exchange networks (e.g. who commented on which post, which user name is mentioned by which other users).

PHEME will also **go beyond text** and also experiment with processing **podcasts from authoritative sources**, which have been transcribed automatically first by the podcast owners. More specifically, SWI owns podcasts in English and German. We have also budgeted for USAAR will acquire the APA corpus. This German corpus from the Austrian Press Agency (APA) contains terabytes of printed press data and manually transcribed broadcast and TV programmes, enriched with metadata (e.g. about participants and turns in an interview). USFD has a similar (albeit smaller) English corpus of transcribed BBC news programs from the PrestoSpace project.

New authoritative content and online media will be acquired during the project from a variety of sources, including Twitter, political and medical blogs, newspapers, SWI's own published news and podcasts, patient forums, and scientific articles. The consortium already has **significant volume of historical data**, both **authoritative content** (e.g. 230 000 news text, podcasts, and forums published by SWI in 9 languages; the APA and BBC corpora mentioned above; USFD's METER text reuse corpus of Press Association news and articles on the same events (265 events in total) from The Sun, Daily Mirror, Daily Star, Daily Mail, Daily Express, The Times, The Daily Telegraph, The Guardian and The Independent (Clough et al, 2002); KCL's anonymised CRIS patient data and **historical social media content** (7.2 TB compressed Twitter data, starting from May 2009 (on average over 2 million tweets/day from a Twitter gardenhose account at USFD); a YouTube video and transcriptions dataset (Morency et al, 2011); the KONECT network connections datasets (Kunegis, 2013)). For the healthcare use case, we will use also a corpus of manually tagged adverse drug events, which was jointly developed by USFD, Fraunhofer Institute, B-IT, and Merck (Gurulingappa et al; 2012). Other relevant healthcare corpora come from the i2b2 initiative (<https://www.i2b2.org>), and the University of Pittsburgh NLP Repository (<http://www.dbmi.pitt.edu/nlpfront>).

Three languages addressed: PHEME will cover three European languages (EN, DE, BG), in two of which data is freely available, both as authoritative sources (e.g. broadcast news) and social media (e.g. Twitter, blogs, podcasts) and linguistic resources: English and German. In addition, some experiments will be carried out on a less widely studied Slavonic language, Bulgarian, to demonstrate the extensibility of the system. The chosen languages cover two of the major language groups within Europe (Germanic and Slavic), ensuring that expansion of the system to other EU languages at a later stage is not overly burdensome.

B1.1.4 RTD Challenges and Innovation

In order to reach its ambitious scientific and technological objectives, foundational and component-level research is required in the PHEME project. This includes the following **novel contributions**:

- **Modelling, extraction, verification and reasoning about rumours**, including measuring authenticity and credibility of phemes. PHEME will create ontological models of rumours, veracity, and temporal validity of facts, as well as new methods for reasoning about these. This will be coupled with automatic methods to extract phemes across media and languages. Disputed information will be detected through entailment and contradiction methods. New

methods for assessing information credibility will allow PHEME to identify misinformation, whereas longitudinal models of pheme diffusion patterns, source authority, and trust will aid disinformation detection.

- **Providing richer interpretation context through cross-media analysis:** PHEME will analyse and interlink textual content, social interaction graphs, and other media, including automatically transcribed podcasts and TV programmes. User-generated content (e.g. tweets, blogs, patient forum posts) will be aligned automatically to contemporaneous authoritative content (e.g. news, scientific articles), linking across media and languages.
- **Veracity assessment based on multiple evidence:** PHEME will develop new veracity intelligence techniques, which combine the linguistic, syntactic, and semantic information within the document with world knowledge from Linked Open Data resources. Social network evidence will also be included, e.g. automatically detected implicit information networks, user trustworthiness models, rumour network propagation patterns.
- **Addressing the temporal validity of information** and historical content, to assess contradictions, including longitudinal models of users, networks, trust, and influence. PHEME will develop techniques for managing multiple spatio-temporally overlapping data at the appropriate temporal granularity level and accounting for this when checking for spatio-temporal consistency (e.g. both Cameron and Brown were prime ministers of the UK in 2010).
- **Using Linked Open Data** as large-scale source of **world knowledge for veracity verification**, necessary for dealing with synonymy, meronymy, hypernymy, and functional relations, in order to resolve seemingly contradictory statements as actually consistent.
- Developing **interactive visual analytics methods**, to support search and browsing of the extracted veracity intelligence. The pheme search and browsing dashboard will span media, social networks, geographic locations, and time, within a single, coordinated interface.

The RTD activities in the PHEME project will build upon existing language tools and machine learning methods, especially those that are being developed in related EU projects in FP7 (e.g. Excitement, TRENDMINER, X-Like, Meta.Net, MONNET, EuroSentiment, FIRST, BIG).

B1.1.5 Expected Outcomes

The tangible outcomes of PHEME will be of several kinds:

- On the level of **models and approaches**, the PHEME project will deliver **multi-disciplinary advances, towards creating actionable knowledge**. These will include:
 - Novel models and approaches for *interpretation, diffusion and visualization of veracity intelligence, acquired automatically across media, languages, and social networks*. This will lead to better understanding of phemes in social media: where they originate, how they spread, who and why spreads them, ultimately resulting in cleaner big data sources.
 - New evaluation datasets
 - High quality journal and conference publications
- On the **methods and technology** level, the outcome of PHEME will be open-source algorithms for multilingual detection of phemes over social networks, stream media, and authoritative sources. The discovery, reasoning with, and visualisation of veracity intelligence across different media will be a key novel contribution.
- On the **tools and applications** level, the project will deliver:
 - **An integrated, scalable framework** for extraction, aggregation, and visualisation of veracity intelligence, including also automatic cross-referencing against authoritative sources. The new algorithms and technology will be made available as web services, integrated into a scalable veracity intelligence framework.

- New technology necessary for the PHEME industrial partners to build smarter products, e.g. ONTO's semantic publishing, search, and knowledge management products; WebLyzard's media monitoring platform; IHUB's social intelligence products (SwiftRiver and Ushahidi).
- The case studies will provide two demonstrated deployments in medicine and digital journalism. PHEME will thus enhance the productivity of SWI's editorial team and help KCL to improve the quality of healthcare advice provided to their patients.

The two case studies will be used to benefit all project deliverables. The software components and theoretical models and approaches will be refined in the light of the experience of these case studies, using case-study specific quantitative data sources alongside public web data. The case studies will also provide key input to dissemination and help inform the exploitation strategy.

Consequently, there are five major strands of activity, within each of which there are measurable or verifiable outputs to be delivered by each major milestone (see section 1.3.4).

B1.2 Progress beyond the state of the art

This section positions PHEME vis-à-vis latest techniques and on-going research efforts, and defines PHEME's novel contributions.

B1.2.1 Understanding Rumours and Information Flows in Twitter

STATE OF THE ART: Procter et al. (2013) analysed false rumours in a corpus of tweets collected during the August 2011 riots in England, using established qualitative methods from social science. The methodology was based on the classic two-step flow model of communication, highlighting how information flows from 'opinion leaders' to others (Katz and Lazarsfeld, 1955; Wu et al., 2011). First Procter et al. grouped source tweets and retweets into 'information flows' (Lotan et al., 2011), then ranked these by flow size, as a proxy of significance. To understand how Twitter was being used, they developed a code frame (Krippendorff, 2004) to categorise information flows (e.g. a report of an event, a comment about a report, a request for information, etc.) and used the groupings to explore how a given rumour spread through Twitter.

A common finding was that the mainstream media is seen to lag behind crowd-sourced ('citizen journalism') reports appearing in social media. This emphasises how collaborative efforts by large numbers of 'producers' (Bruns, 2006) can provide competing and, at times, better coverage of events than mainstream media. The use of links to other media, e.g. phone images, blogs and online newspaper sites as corroborating evidence, is another common feature in all seven of the rumour case studies. However, they show that this evidence cannot always be taken at face value, leading some to claim that images, for example, had been had been faked ('photoshopped') to substantiate a false rumour.

Overall, the findings suggest that while Twitter is a fertile medium for launching rumours, it also provides robust mechanisms for self-correction (Mendoza et al., 2010). Also, the proportion of information flows supporting or denying a rumour in these case studies was broadly consistent with the findings of Mendoza et al. (2010) who noted that users deal with 'true' and 'false' rumours differently: the former are affirmed more than 90% of the time, whereas the latter are challenged (i.e. questioned or denied) 50% of the time.

PHEME'S NOVEL CONTRIBUTION: Procter et al.'s methodology did not rely on computational methods for identifying rumours, which is where PHEME's novelty lies. In PHEME, Procter (UWAR) will extend this social informatics analysis of rumours, by examining how phemes spread across media. Moreover, the types of rumours studied will grow to include speculations, controversies, and misinformation. This qualitative analysis will help formalise rumours and phemes, define the associated linguistic annotation schemas, and provide human insight into the mechanisms of rumour spread across media, languages, and cultures.

B1.2.2 Automatic Methods for Detecting Contradictions, Claims, and Misinformation

STATE OF THE ART: The task of automatic **Contradiction Detection** (CD) is relatively unstudied in language processing, in comparison to other tasks such as information and sentiment extraction. It has been formulated primarily as a kind of textual entailment, where a pair of statements asserts contradictory information (Voorhees, 2008). (Marneffe et al, 2008) have established a typology of contradictions, which distinguishes easier to detect, overt contradictions (antonymy, negation, and data/number mismatch) from harder-to-detect, implied contradictions arising from factive or modal words, structural and lexical contrasts, and world knowledge. Automatic contradiction detection approaches have focused mostly on detecting contradictions based on negation, antonymy, and numeric mismatches, e.g. (Marneffe et al, 2008; Harabagiu et al, 2006; Kawahara et al, 2010), using supervised machine learning methods. Linguistic features are typically derived from named entity types, syntactic and semantic parsing, temporal information, antonyms, events, paraphrases, and string overlap. Topic similarity and event coreference also play an important role, since genuine contradictions only arise in cases where two texts discuss the same event or entity.

(Ritter et al, 2008) studied the harder task of detecting contradictions where world knowledge is required (e.g. the assertion that Mozart is born in Austria does not contradict that he is born in Salzburg). The method uses WordNet as a source of synonymy and meronymy knowledge, but found problems with its low coverage. Other relevant knowledge is hypernyms (renal failure is a kind of kidney disease), so there is no contradiction between documents using one term vs. the other. Ritter et al also established that in most cases seemingly contradictory statements are actually entailments, if world knowledge is present (34% of errors are due to missing knowledge and 49% of errors are due to location ambiguity). The problem is that current CD methods do not draw on sufficiently large world knowledge sources, which leads to low accuracy results.

Controversial information may appear in both “traditional” and social media (e.g. whether aluminium causes Alzheimer’s disease). Fact-checking websites (e.g. factcheck.org and politifact.com) are useful for checking known disputed claims manually. However, the automatic detection of such disputed information has not been studied extensively in NLP. A first step in that direction was made in DisputeFinder (Ennals & al., 2010) - a now obsolete Firefox plugin which used a database of known disputed claims to detect and highlight mentions of these claims in web pages. The claim detection algorithm had limited accuracy, due to being based purely on local lexical matching. PHEME’s more sophisticated algorithms for contradiction detection will be coupled with features on who states the claim and their trustworthiness. We will also investigate whether claims could be identified using Hearst-like patterns, e.g., “<Entity> claims that X”.

Misinformation and disinformation are two kinds of rumours, spreading through social media and online networks. As argued by (Qazvinian et al, 2011), rumour classification is a different class of problem from opinion mining. Studies of rumours and urban legends in psychology (Guerin, 2006) have motivated their popularity in terms of conversational and social properties, i.e. that people spread these because they make appealing stories, are hard to verify, and at the same time, improve social status. Online social networks have become an extremely fertile ground for rumours and misinformation (Mendoza et al, 2010), especially during natural and man-made disasters, when shocking, unverified statements can even sometimes be reported by mainstream media¹. (Ratkiewicz et al, 2011) use the AdaBoost machine learning algorithm for detecting political abuse in Twitter, based on the topology of the social network (e.g. mean degree, min strength) and basic mood-based sentiment features. (Qazvinian et al, 2011) go beyond network-based features and include lexical and part-of-speech information, in order to classify tweets as rumour-related or not, given an already known rumour (e.g. Obama being Muslim). They also identify whether the person posting the tweet endorses the rumour or not. The main limitation of this work, however, lies in its reliance of already identified rumours from external sources and failure to consider the wider context.

In contrast, the PHEME use cases (digital journalism and healthcare) require methods for automatic identification and verification of newly emerging rumours and their spread as phemes across media. To this end, we will use semantic-based features, as well as automatically derived knowledge about the poster and the information diffusion patterns from the social networks (see section 1.2.4 below). Moreover, (Qazvinian et al, 2011) only focused on Twitter, whereas PHEME will deliver improved performance by considering evidence from other media, e.g. via embedded URLs, urban legend websites, and news reporting rumours.

Development and Evaluation Corpora: Several corpora underpin state-of-the-art research in this area. The datasets from the Recognising Textual Entailment (RTE) challenges² contain text pairs annotated for entailment, contradiction, or unknown relation (Voorhees, 2008). These datasets are artificially created to be balanced between positive and negative examples, which does not reflect realistic distribution in naturally occurring web content (Ritter et al, 2008). Two smaller corpora of naturally occurring contradictions have been created (from newswire and Wikipedia): 131 sentence pairs (Marneffe et al, 2008) and 8,844 sentence pairs (Ritter et al, 2008). With respect to rumours in Twitter, the biggest available dataset (Qazvinian et al, 2011) includes 10,000 manually annotated

¹ <http://gigaom.com/2012/10/30/hurricane-sandy-and-twitter-as-a-self-cleaning-oven-for-news/>

² <http://www.nist.gov/tac/data/index.html>

tweets with respect to five pre-identified rumours, whereas the Truthy dataset (Ratkiewitz et al, 2011) has 61 false claims and 305 legitimate ones.

Recent evaluation exercises of textual entailment (e.g. RTE-7) directly support parts of PHEME's core functionality, while the project also reaches beyond the state of the art in this field. These exercises have built upon the base task of finding supporting sentences or phrases when given a hypothesis or assertion. General goals of research in textual entailment include determining whether a conclusion is appropriate given a set of documents, and deciding how sufficient an evidence base is for any given claim. This field, and recent research generated in the course of these evaluations, help PHEME in two key tasks. Firstly, textual entailment supports the detection of new claims (e.g. those partially unsubstantiated by current evidence but not contradicted). Secondly, it helps determine whether a claim is supported given a set of documents. This latter task is challenging, with the best system managing an F-measure of only 19%. These evaluation exercises and their datasets provide an excellent basis for evaluating PHEME's novel contributions.

PHEME's novel contributions: Tapping into the wider context (including the social graphs):

A key novel aspect of work in PHEME will be in developing rumour detection algorithms, specifically designed for analysing shorter and interlinked social media content. The latter poses challenges due to shortage of contextual information. To overcome this problem, PHEME will tap into the currently unexplored past user content and metadata. We will examine the user's past messages for any previously identified rumours and phemes, which they propagated. Links posted by the user, and the content of their user profile, also provide additional evidence. Outside of the creator's content, we will also experiment with adding their friend's content and using that as additional context. It would also be possible to move to other sites, using a multi-layered social network model (Magnani, 2011), to extract a user's and their connections' messages.

Moreover, PHEME will integrate evidence from longitudinal models of users, influence, and trust.

Exploiting Large-Scale World Knowledge from Linked Open Data: The major bottleneck in current methods is in their limited use of world knowledge. PHEME's novelty is in tapping into Linked Open Data resources (e.g. GeoNames, YAGO), as sources of large-scale world knowledge, necessary for dealing with synonymy, meronymy, hypernymy, and functional relations. For example, the fact that Salzburg is in Austria is present in GeoNames and therefore a seeming contradiction between which one is Mozart's birthplace, can be eliminated. In particular, OWLIM will be adapted as a highly scalable semantic repository for storing rumours and phemes, and supporting reasoning about functional properties, temporal validity of facts, and rumours. Also, FactForge will be used as an integrated system of gathering facts from the web semantic databases (including GeoNames, FreeBase, DBpedia, WordNet, and YAGO). In addition, functional properties will be modelled explicitly, to enable more reliable contradiction detection.

New Training and Evaluation Corpora: PHEME will collect and annotate new cross-media corpora of naturally occurring phemes in news, user-generated content, and social networks. Unlike existing corpora, which only provide short texts and annotations, we will also include relevant social network data (e.g. the followers and followees of a given Twitter user, and the history of interactions between them) and historical content (e.g. the user's previous tweets). Such richer corpora will provide the necessary posteriori knowledge and social context, to aid the semantic interpretation methods. Corresponding content from mainstream media, coupled with corrections, clarifications, and retractions, will also be included. Lastly, **multilinguality** is a key novel contribution of PHEME - all currently available datasets and algorithms cover only English, whereas PHEME will address also German and Bulgarian.

B1.2.3 Spatio-Temporal Grounding

STATE OF THE ART: Unlike traditional news, a notable proportion of social media content posted online is explicitly geotagged (Sadilek et al, 2012), and studies suggest that it is possible to infer the geo-locations of about half of the remaining such content (Mahmud et al, 2012). Social media also has at least a creation time as temporal context. In a nutshell, social media content contains a wealth

of explicit and implicit spatio-temporal (ST) metadata, which is not currently exploited by the NLP methods discussed above.

Given the constraint that a single entity can only be in any one place at a time, these forms of ST information give a means of determining the factuality of statements. There is a body of initial work into determining the spatio-temporal bounds of assertions (Ji & al., 2011), though this mostly relies on explicit information and is not a thoroughly-investigated field.

Temporally, current systems are capable of detecting the publication date of documents (Chambers, 2012) and of grounding some of the time expressions contained therein (Strötgen & Gertz, 2010). Detecting events and assertions and temporally ordering these with regard to times is critical to ST grounding of facts and rumours; the state of the art in event detection is good (Kolya & al, 2012), but ordering events and times relative to each other or across documents remains an active area of novel research with some progress to be made. Fortunately, linking events to times – the most important type of temporal association for PHEME – is also the subtask at which automated systems have the most success (Verhagen & al, 2010).

Spatially, the challenge of grounding the locations in document content is critical to accurate spatial bounding. The state of the art is somewhat less mature than that of temporal context; while many tools can identify a range of named entities, recognition of new spatial information (especially when general nouns are used in a spatial sense (e.g. room) is a subject of active research (Gaizauskas & al, 2012).

PHEME’S NOVEL CONTRIBUTION: The temporal delimitation of any assertion is of great importance, because the assertion is true only inside these bounds. Specifically, it is possible to extract two truths that seem to contradict (e.g. “USA president is G W Bush” and “USA president is B Obama”) but are in fact both accurate when the appropriate temporal information is added. In other words, there is something like temporal validity of facts, which needs to be taken into account when detecting contradictions.

USFD is a leader in both time expression resource creation (Derczynski & al, 2012) and interpretation for English. PHEME will adapt our tools to social media data, through the creation of new training data in this genre. We will also cover new target languages with minimal effort, through lightly-curated annotation porting, taking advantage of the language-independent nature of grounded spatial and temporal data.

Another important benefit of storing and analysing “traditional” and social media content over time is that these archives enable longitudinal analysis. For instance, longitudinal analyses on the online social graphs can reveal the evolution of social relationships and thus build models of trustworthiness and authority (Section 1.2.5). It is also possible to start building user profiles over time, including previously spread rumours and, in general, what users talked about in the past. Focused on specific events, longitudinal analysis reveals discourse around events, arising from both social and traditional media. Similarly, in journalism and brand and reputation management applications, there is also demand for retrospective analyses of media content after a significant incident (e.g. to establish whether social media was used to entice more riots).

B1.2.4 Cross-Media Content Linking

STATE OF THE ART: The short nature of social media, coupled with their frequent grounding in real world events, means that content cannot be understood without reference to external context (e.g. news). While some posts contain URLs, the majority do not, which motivates methods for cross-media content linking.

(Abel et al 2011) link tweets to current news stories in order to improve the accuracy of semantic annotation of tweets. Several linkage strategies are explored: TF-IDF similarity between tweet and news article, hashtags, and named entity-based similarity, with the entity-based one performing the best. The approach is similar to USFD’s keyphrase-based linking method for aligning news video segments with news web pages (Dowman et al, 2005). MOD has also developed algorithms for aggregating social media content on climate change from Twitter and Facebook with online news (Hubmann-Haidvogel et al, 2012). These will be re-used in PHEME in the journalism use case, when

aligning tweets to audio and video materials from authoritative news sources (transcribed automatically first).

An in-depth study comparing Twitter and New York Times news (Zhao et al, 2011) has identified three types of topics: event-oriented, entity-oriented, and long-standing topics. Topics are also classified into categories, based on their subject area. Nine of the categories are those used by NYT (e.g. arts, world, business) plus two Twitter specific ones (Family&Life and Twitter). Family&Life is the predominant category on Twitter (called 'me now' by (Naaman et al, 2010)), both in terms of number of tweets and number of users. Automatic topic-based comparison showed that tweets abound with entity-oriented topics, which are much less covered by traditional news media.

Going beyond interlinking news and tweets, future research on cross-media linking is required. For instance, some users push their tweets into their Facebook profiles, where they attract comments, separate from any tweet replies and retweets. Similarly, comments within a blog page could be aggregated with tweets discussing the same topic, in order to get a more complete overall view. In addition, all current methods are monolingual, whereas PHEME will link content also **across languages**.

PHEME's novel contributions: PHEME will concentrate on how information and phemes spread from sources such as Twitter, scientific articles, policy documents and press releases, into newspapers stories and social media content. The methods will firstly cluster documents according to the statements made within, across different media and languages, and present these. Secondly, related documents will be identified within these clusters, due to one citing the other, or reusing some of the content. This automatically inferred knowledge will underpin the interactive visual analytics methods (see Section 1.2.5 next), which will show the flow between different texts on the web, in order to highlight areas of consensus and conflict and the most active areas of discussion; and, to help users investigate the origins of rumours and information online.

The challenge is to build computational models of cross-media content merging, analysis, and visualisation and embed these into algorithms capable of dealing with the dynamic, contradictory and interlinked nature of online media, social networks, and authoritative content. In particular, PHEME will develop algorithms for cross-media content clustering, modelling contradictions between different sources, and inferring change in user interactions, reputation and attitudes over time.

B1.2.5 Implicit Information Networks, Trust, and Rumour Spread

Information diffusion plays a crucial role in a range of phenomena, including the spread of phemes. A comprehensive body of research investigates diffusion mechanisms (Romero et al. 2011), strategies for identifying influential users (Kimura et al. 2010; Tang et al. 2009), maximizing the impact of stimuli (Belák et al. 2012), and the role of external and internal factors on information diffusion (Romero et al. 2011).

A large body of research deals with information flow in explicit networks where the connections between nodes are defined by a social network such as Facebook, by follower relations (Twitter), or by reply-to connections in Twitter (Romero et al. 2011) or Flickr (Cha et al. 2009). Therefore, these studies apply methods from social network analysis, and even address issues such as tracking epidemics (Lamos 2010).

Tracking information flow in implicit networks is a more challenging task because it involves: (i) identifying identical information units; (ii) determining the point of time at which this information has been published; (iii) tracking the flow within this network, and (iv) inferring the implicit diffusion network.

Identifying similar contagions is a complex task since standard hash functions such as MD5 hashes (Rivest 1992) or SHAsums (Donald & Jones 2001) that are used to detect identical text snippets are no longer applicable, since text snippets mutate during the diffusion process. Current research, therefore, uses locality-sensitive hashing algorithms to address this problem (Paulevé et al. 2010). Such techniques have already been successfully applied for content and package level spam-filtering (Marsono 2011; Pérez-Díaz et al. 2012) and the identification of source code duplicates (Chang & Mockus 2008). (Gomez-Rodriguez et al. 2012) used MemeTracker to identify 343 million memes in

172 million news articles and track them over a year. (Bendersky & Croft 2009) discuss algorithms and metrics for determining text reuse. Another approach, using sketches of shingles of tokens, has been used to detect near-duplicate web pages (Broder, 2000). Research in PHEME will also be using USFD's METER corpus of text reuse (Clough *et al*, 2002).

Implicit networks created by dialogue can also be found over explicit social networks such as Twitter. For example, those contributing to a hashtag conversation are not bound by explicit links such as follower or friend relations, but instead cause information to diffuse among those interested in that topic (Rossi and Magnani, 2012). Monitoring relationship forming and the behaviour of phemes within these ad-hoc networks provides a readily available source of rumours (and facts) and people's interaction with them.

Although approaches demonstrate the potential and usefulness of this line of research, they are still limited in terms of scope and domain. For instance, most approaches focus either on one particular social media site such as Facebook or Twitter (Myers *et al*. 2012; Romero *et al*. 2011), on selected sets of forums (Belák *et al*. 2012), blogs (Lim *et al*. 2009) or USENET articles (Mary McGlohon 2009) and, therefore, only cover a tiny fraction of the potentially involved stakeholders. Two recent integrated approaches (Gomez-Rodriguez *et al*. 2012; Miller *et al*. (2011) cover blogs and news articles, but still suffer from limitations. In particular, they fail to take into account textual content when estimating the influence probabilities and consider the dynamic nature of connection strengths and user influence (i.e. how these change over time).

Specifically with respect to modelling the spread of misinformation, (Nguyen *et al* 2012a) identify the most suspected misinformation sources in an online social network and, in follow-up research (Nguyen *et al*, 2012b), simulate how misinformation spread could be contained. However, even though they use real user network connection data (Facebook, ePinions), the actual misinformation data is simulated. Consequently, even though highly influential users were identified by the methods, it remains unclear in practice, whether the algorithms will be effective on real cross-media rumour datasets.

Trust and credibility: Since PHEME deals with a wide *variety of content*, where many potentially conflicting assertions are made, it is important to model trust as a mechanism for deciding which of a set of sources is the reliable one. Network models for trust are already used in information retrieval (Gyongyi *et al*, 2004) and have been adapted for Twitter (TrstRank, 2010). These typically take a naïve PageRank style score and propagate trust between nodes in the network automatically. Provenance is an important element of trust modelling, where PHEME will reuse results arising from the TRENDMINER project.

To manage trust based on content requires some authoritative or near-authoritative sources. (Levin, 1998) calculate global reputation for each actor in a network, and also allow the listing of certain nodes as 'bad', thus cutting out unreliable parts of the network. With regard to social networks, trust is indirectly represented by the influence and attention received; if a user generates or propagates unreliable information, they are likely to receive less attention or not have their messages re-propagated (Quercia *et al*, 2010; Nel *et al*, 2010).

What such trust models currently lack is the ability to automatically **verify or refute the content** issued by actors in the network and accordingly bias the trustability of nodes and messages.

PHEME'S NOVEL CONTRIBUTION: The project will model, identify and track over time implicit information exchange networks, i.e. the users (people, media corporations, enterprises) and the phemes they exchange through a variety of media, including news, (micro)blogs, and social networks.

PHEME will augment the state-of-the-art techniques, to go beyond network-specific features and include linguistic, semantic, and spatio-temporal features. Another currently unexplored issue is taking into account historical data on user behaviour, i.e. modelling user reputation based on previously spread rumours or disinformation. Corrections, clarifications, and retractions made by mainstream media will also be used as evidence source, where available.

By capturing longitudinal changes PHEME will distinguish spikes (externally induced sharp rises in activity) from chatter (internally driven, sustained discussions), since the frequency and shape of

spikes is a powerful indicator of information diffusion (Gruhl et al. 2005). Occasionally, spikes result from chatter through a process of resonance, where insignificant exogenous events trigger massive reactions. Such sensitive dependence on initial conditions occurs when large sets of individual interactions generate large-scale, collective behaviour. PHEME will reveal the structure and determinants of these paths to guide organizations in their efforts to gather actionable knowledge from interlinked online media and networks. This represents a significant contribution, because it requires an integrated analytical framework to capture semantic, spatial and temporal effects in content production and content consumption simultaneously.

With respect to **trust**, PHEME will deliver methods for corroborating or refuting claims, based upon highly-trusted sources, leading to an estimation of a source's trust and reputation. This gives a means of capturing linguistic and network behavioural datasets (Romero et al, 2010; Weenig et al, 2001) of users who trust (or distrust) sources. Further, as trust is reflected in audience engagement and the language used, using these datasets will enable PHEME to develop novel algorithms, based on research from epidemiology, that estimate how much a given actor can be trusted and how influential they are.

B1.2.6 Interactive Visual Analytics

The main challenge in browsing and visualisation of interlinked media and social network content is in providing a suitably aggregated, high-level overview. Timestamp-based list interfaces that show the entire, continuously updating stream (e.g. the Twitter timeline-based web interface) are often impractical, especially for analysing high-volume, bursty events. For instance, during the royal wedding in 2011, tweets exceeded 1 million. Similarly, monitoring long running events across media, places, and time is equally complex.

One of the simplest and most widely used visualisations is word clouds. These generally use single word terms, which can be somewhat difficult to interpret out of context. Word clouds have been used to assist users in browsing social media streams, e.g. blog content (Bansal, 2007) and tweets (Nagarajan *et al*, 2009; Shamma, 2010). The main drawback of cloud-based visualisations is their static nature. Therefore, they are often combined with timelines showing keyword/topic frequencies over time (e.g. Hubmann-Haidvogel, 2012; Adams, 2011), as well as methods for discovery of unusual popularity bursts (Bansal, 2007).

In addition, some visualisations try to capture the semantic relatedness between topics in the media. For instance, BlogScope (Bansal, 2007) calculates keyword correlations, by approximating mutual information for a pair of keywords using a random sample of documents. Another example is the information landscape visualisation, which conveys topic similarity through spatial proximity (Hubmann-Haidvogel, 2012).

Lastly, some visualisations have tried to convey the social context. For instance, the PeopleSpiral visualisation (Dork, 2010) plots Twitter users who have contributed to a topic (e.g. posted using a given hashtag) on a spiral, starting with the most active and 'original' users first. User originality is measured as the ratio between the number of tweets authored by the user versus re-tweets made.

PHEME'S NOVEL CONTRIBUTION: The project will build visual analytics tools for the collected veracity intelligence, including visualisations of geospatially and semantically referenced information, across news, media and social networks. Exploring the storytelling potential of big data visualization (Segel and Heer, 2010), the interactive components of PHEME will increase the understanding of the complementary relationship between the explorative and communicative dimensions (Hullman and Diakopoulos, 2011).

The PHEME visual analytics dashboard will allow users to search and browse the automatically extracted veracity intelligence, spanning content, social networks, geographic locations, and time, within a coordinated dashboard consisting of multiple, linked views (Hubmann-Haidvogel et al., 2009). The positive effects of such interfaces on user efficiency and productivity are already proven (Chimera and Shneiderman, 1994). PHEME will extend MOD's open-source visualisation framework (Scharl et al., 2013) with new visualisations of threaded dialogs, pHEME diffusion patterns, and maps of spheres of influence and trust in social networks.

B1.2.7 The PHEME Veracity Intelligence Framework: Scalability and Efficiency

The PHEME veracity intelligence framework needs to combine software components related to different tasks (e.g. linguistic pre-processing, veracity analysis, rumour detection) in a flexible fashion. Based on ATOS's experience from FIRST and ONTO's from TRENDMINER, a Service Oriented Architecture (SOA) approach will be adopted. Since PHEME will manage large volumes of content, scalability and on-demand compute power will be achieved through cloud computing.

Traditionally, the Cloud Computing approach consists of three architectural layers: (i) Software-as-a-Service (SaaS): the software structure (components, interfaces, workflows), based on SOA principles; (ii) Platform-as-a-Service (PaaS): the environment configuration required for the execution of SaaS; (iii) Infrastructure-as-a-Service (IaaS): describes all the physical resources and their characteristics. In the PHEME framework (WP6), the concept will also need to be extended to Database-as-a-Service or Storage-as-a-Service .

PHEME will also rely on MapReduce to ensure scalable text processing on large historical data. MapReduce was originally introduced to support large-scale indexing tasks. Of particular relevance to PHEME is the Hadoop Distributed File System (HDFS) - a simple and efficient model for incremental indexing in digital document repositories. Hadoop is already widely used within research, to index sizeable collections such as the Terabyte TREC or the MAREC collection.

Finally, PHEME will build on the Storm framework and its adoption within the TRENDMINER project, to provide support for real-time content processing. Storm provides a set of general primitives for distributed real-time computation, e.g. processing messages and updating databases in real-time. Storm can be used for continuous computation (executing a continuous query on data streams and streaming out the results) and/or distributed RPC (running expensive computations in parallel on the fly).

PHEME'S NOVEL CONTRIBUTION: The PHEME framework will be designed as follows.

- at the software level a SOA architecture to compose and orchestrate the diverse services, e.g. multilingual syntactic analysis; spatio-temporal tools; veracity reasoning; and rumour detection,
- at the application level with an empirical test-bed platform to configure and test service execution,
- at the infrastructure level, enabling large-scale analysis of diverse content.

The PHEME framework will build on tools and expertise in real-time stream media processing from FIRST and TRENDMINER. Changes however are needed to accommodate historical data. Following (Marz, 2012), we will decompose the veracity analytics problem into three layers: the batch layer, the serving layer, and the speed layer. Incoming data will be stored both in a historical raw data storage and diverted for real-time computation. The historical raw data is then processed in the Batch Layer (i.e. using MapReduce) and the results of the process indexed or exposed in the Serving Layer (e.g. using SOLR indexes). The incoming content stream is analysed in the Speed Layer (e.g. using Storm) in parallel, providing a real-time view. Both views can be used to compose intelligent queries that take into account both historical and/or real-time data.

B1.2.8 Veracity Intelligence for Patient Care

Public health professionals have started to undertake social media surveillance for various purposes, e.g. real-time tracking of flu epidemics (Lampos et al, 2010), syndromic classification of Twitter messages (Collier and Doan, 2011), or examining attitudes towards vaccination (Salathe, 2011). In this context, veracity, rumours, and spam are a major challenge. For example, PatientsLikeMe.com contains a wealth of discursive information and personal patient data but veracity and the influence of rumour is a major problem (not least because pharmaceutical lobby organisations are assumed to be active in such social media). Given the volume and velocity of content, however, there is a strong need for automated algorithms for rumour detection, to help public health researchers with aggregating and monitoring patient forums, Twitter, etc.

PHEME will complement and extend existing collaboration between KCL and the pharmaceutical industry which have included a pre-competitive consortium with Pfizer, J&J and Lundbeck to support the development of USFD's GATE-based biomedical text mining applications to ascertain social care in dementia for economic analysis, as well as the development of other GATE tools, funded by Roche, to ascertain negative symptomatology in schizophrenia from free text fields.

PHEME's spatio-temporal focus will complement current developments in the KCL/SLAM BRC Clinical and Population Informatics theme to harness geospatial expertise in order to characterise better SLAM's multicultural and socially diverse catchment area - not only to link with secondary care statistics provided by their CRIS electronic patient register, but also to make use of micro-level characterisation through a large BRC-funded survey of mental health and environmental stressors: the SELCOH project (Hatch et al, 2011).

The BRC have recently been awarded NIHR funding to develop the CRIS application for other sites and create a substantially larger academic network as a result. Data will thus be imported over the next 12 months from health services covering Oxford and Cambridge as well as north and west London services. This will result not only in substantially greater sample sizes for analyses, but also in much greater sample heterogeneity. Lastly, covered by current NIHR BRC funding, a shared electronic record is currently being piloted, having been developed in collaboration with Microsoft. This will bring in valuable patient-generated data of potential relevance in the later stages of PHEME.

PHEME'S NOVEL CONTRIBUTION: PHEME will show how the new veracity intelligence methods can be applied to a health-related use case, and how social media analysis can be integrated with public health monitoring and with analysis of the electronic patient record (EPR). Specific topics that will be considered as use cases for demonstration are:

- The impact of public health concerns and health-related rumours, including issues of medications, trace elements and food additives (e.g. Alzheimer's, autism, Attention Deficit Hyperactivity Disorder)
- The emergence and use of new (and old) illicit and legal drugs
- The correlation of phemes around dieting with eating disorders reported in the EPR.

In addition, PHEME will strengthen existing partner links with the IMI-EUPATI project, in order to inform requirements for data collection, and in order to bring benefits to public awareness efforts such as those carried out in EUPATI. The UK Health Protection Agency (HPA) is also an associated member of PHEME.

B1.2.9 Digital Journalism

Journalists have long used information and communication technologies in their work. Computer-assisted reporting was first used in the 1950s to support journalists covering election results. Investigative journalists in particular have often used public datasets combined with data collected specifically as part of their investigation to provide the evidence and analysis to support their stories. Two things have changed.

First, new forms of data have become available, often born-digital as well as the sheer availability of data. In addition, the technical capacity for processing datasets has increased significantly, not least through the development of on-demand resource provision through computational grids or cloud computing. Journalists have entered a world where data is abundant – a situation described as the “data deluge” (Hey *et al* 2009) – but where their traditional methods for data analysis fail to translate to this new data and scale to the volume encountered. Where previously the challenge for the journalist was to find data to underpin their analyses, today the massive availability of different forms of data is challenging their capabilities and capacity to analyse them and, in particular, to assess the veracity of online information and the credibility of its source.

Second, social media datasets are raw, real-time streams of “messy” content, making them difficult to process without significant human effort and expertise to make them ‘fit for purpose’.

Journalists are currently using a plethora of applications, in order to meet their diverse needs, e.g. Tweetdeck for monitoring the social web; Storify - for news aggregation; crowdsourcing tools like

Ushahidi, and online content filtering sites like Storyful. The focus of all these tools is on getting the right content to journalists, but not on helping them with *interpretation and verification of the authenticity and credibility* of that content. Methods and tools vary according to the nature of the journalistic task, however. For example, observations of the Guardian newsroom (Procter et al., unpublished) revealed that journalists prefer simple Twitter clients rather than more sophisticated tools such as Tweetdeck in activities such as live blogging. For reliability's sake, journalists preferred to rely on sources that their experience suggested they could trust. This solves the problem of reliability but limits their capacity to exploit social media to its full potential.

PHEME'S NOVEL CONTRIBUTION: This use case will prototype an open-source digital journalism tool, to support the cross-linking, verification, analysis, and visualisation of veracity intelligence, operating across media and languages. Ushahidi's widely used SwiftRiver open-source platform for collaborative filtering and curation of real-time data from multiple channels, will feed content into the PHEME algorithms for automatic detection of rumours and phemes. Another novel feature will be the methods for discovering and visualising implicit information exchange networks (originators, receivers, diffusers), pheme spread patterns, and trust.

Spatio-temporal knowledge plays also an important role in this use case. A key challenge is to identify the regionality of events (e.g., neighbourhood, city, or country level) (Xu et al, 2012). Regionality is important because different events are relevant at varying scales, which impacts their newsworthiness and interestingness to digital journalists and users interested in local content.

We have a strong user group currently comprising: SWI, the BBC, BBC World Service, SWR, the Guardian, and the Open Society Foundations Media Programme (see Appendix A for details).

B1.3 S/T Methodology and associated work plan

B1.3.1 Overall Strategy and General Description

In order to achieve its ambitious goal, the project comprises a number of research and technical development activities organised as follows.

WP2 will develop **novel ontologies** for representing multiple truths, rumours, and the temporal validity of facts, based on a qualitative social science analysis. WP2 will select and adapt multilingual tools for deriving **document-intrinsic** lexical, syntactic, and semantic features, coupled with **spatio-temporal grounding**.

WP3 will focus on methods for **contextual interpretation**, to derive **posteriori knowledge** from the social networks and cross-media links: identifying new stories and conversations; information diffusion networks, and longitudinal models of trustworthiness and influence of users and information sources.

WP4 will develop PHEME's **innovative, multilingual methods for cross-media detection of rumours**. It will make use the document-intrinsic (WP2) and posteriori (WP3) knowledge, as well as integrate uncertain facts and reasoning with **a priori, large-scale world knowledge** from LOD resources.

WP5 will create a **visual analytics dashboard** for interactive, geo-temporal rumour and impact analysis.

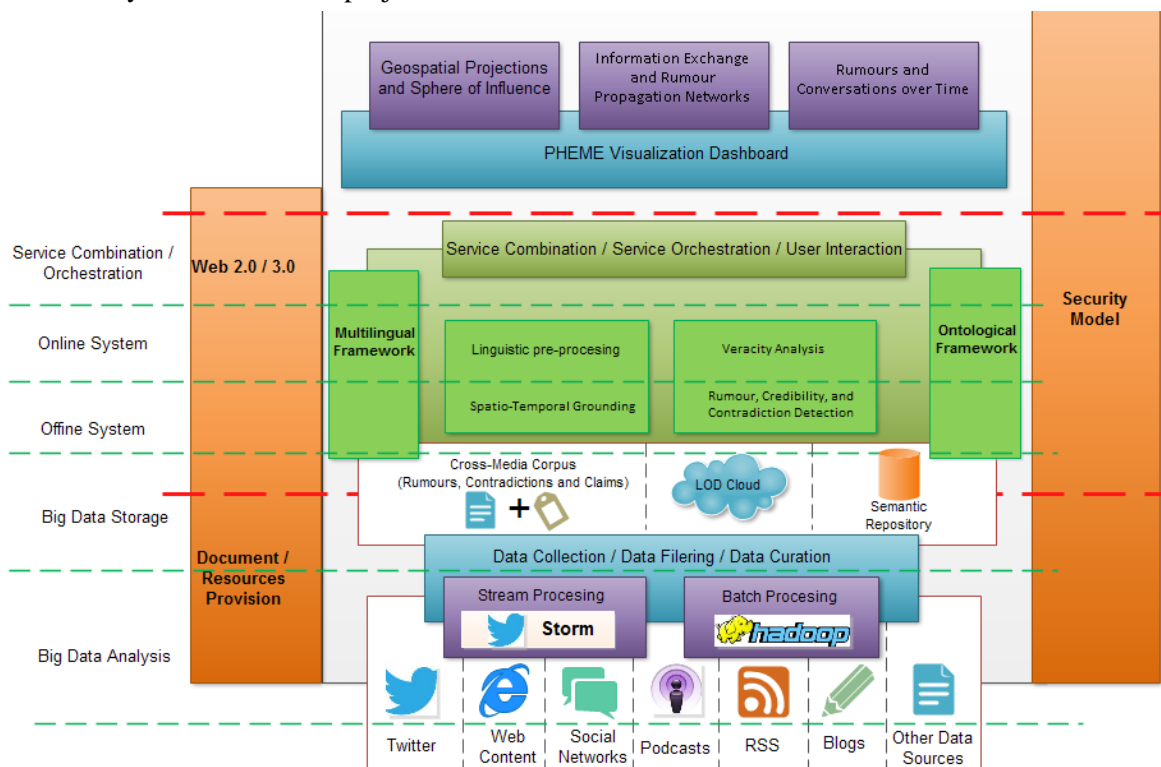
WP6 will **implement the PHEME veracity framework**, which will integrate the veracity intelligence tools and services developed in WP2, WP3, WP4, and WP5. Emphasis will be on storing and processing large volumes of historical data, coupled with real-time analysis of incoming new content. A high-level architectural diagram of the PHEME framework appears below. WP6 will also **evaluate** the scalability and accuracy of all methods.

The achievement of the project objectives will be validated in **two case studies**. WP7 will customise and integrate the PHEME veracity intelligence tools into a prototype for monitoring medical rumours and cross-relating them against electronic patient records. WP8 will create a digital journalism showcase, helping newsroom journalists to find emerging stories and assess their veracity.

Performance evaluation and self-assessment in PHEME is a **horizontal activity**. T6.4 will evaluate the accuracy and scalability of the WP2, WP3, and WP4 algorithms. T5.4 will carry out usability evaluations of the visual analytics dashboard. WP6 and WP7 will evaluate the applicability of the PHEME technology by engaging end-users organised in user groups. While the key performance indicators in WP2, WP3, and WP4 focus primarily on quantitative measures, the case studies use qualitative performance measures.

Dissemination and exploitation activities (WP9) and project management (WP1) will also be horizontal activities, running for the entire duration of the project. In order to support take-up, most PHEME tools will be made **open source** and dedicated support and **training activities** will be established.

Next we present a high-level overview of the PHEME veracity intelligence framework, which will be one of the key outcomes of the project.



The PHEME Veracity Intelligence Framework: High Level Diagram

The basic architecture of the PHEME veracity intelligence framework is based on common SOA principles. The fundamental part is the logical view that should allow a modular and highly generic structure. For this, a three-tier approach has been chosen. This means that the system is decomposed into three different layers which correspond to the following functional blocks: the application, the services, and the persistence of the system.

The **Application Layer** is the application provided to the end users. According to the different use cases defined in the project (WP7 and WP8), different applications can be built to provide adapted user interfaces according to the specific requirements. The Application Layer should deal with the configuration of the user interface, and the management of the user interaction to dispatch the events towards the internal system (Service Layer). A very generic approach will be to allow the easy implementation of new applications adapted for specific business processes.

The **Service Layer** is the core of the system. It contains all the main services provided by the system. Those services are called Core Services and as they could be numerous and very different, they are grouped by Service Categories. Those services are atomic functions that could be called whenever by the system. They will be specified by means of Service Component Architecture (SCA) specification and deployed through the runtime Apache Tuscany.

The last layer is dedicated to system and data persistency, which is why it is called the **Persistence Layer**. It has in charge the mechanisms and models to store information. For each kind of information, a repository is required to store the data. Each repository should provide a basic API to describe its own CRUD (Create, Read, Update and Delete) functionalities to permit easy access to the data. After the first results of integration, a generic approach would be studied to define a standard API for services connexion to the PHEME platform. In PHEME, the most important repository is related to the storage of knowledge. As the formalism to represent the knowledge in PHEME is based on RDF, RDF repositories will have an important place in the architecture.

The various multilingual content analytics services from WP2, WP3 and WP4 will be included in this Layer. The integration will be designed and tested for on-demand scalability, using a distributed computing based on MapReduce and the Hadoop framework for batch analysis of historical data. The Storm framework will be used for real-time analysis of incoming content, which is complementary to Hadoop based batch processing. Finally, the outputs from the analysis content services will be stored in an integrated repository accessible for the services defined in the Core Layer by means of a standard API.

B1.3.2 Key Performance Indicators

Below we define the key performance indicators, which are specific to PHEME, as well as quantified targets, against which the state of achievement of the results will be measured yearly. The indicators are not limited to measuring simple technical aspects but correspond to the concrete expected results indicated in earlier sections of this document.

In terms of **pure scientific innovation** there are several measurable goals that will be measured in terms of publications (taking into account the quality of journals/conferences):

- Number of scientific publications of PHEME;
- Number of citations that the publications receive (measured with Google scholar) (but please note that these take time to appear, at least 1 calendar year post-publication);
- Number of impact points in journal publications resulting from PHEME

In terms of the **algorithms to be developed** (WPs 2, 3, 4, 5) the following factors will be measured:

- Data volume used in terms of veracity analysis and knowledge sources;
- Precision, recall, f-measure and other quantitative evaluation metrics, as detailed in the respective task descriptions in WP2, WP3, and WP4. State-of-the-art algorithms will be used as baselines. We will aim to achieve statistically significant improvement in performance.
- For Task 3.4, the longitudinal models will be evaluated in terms of their predictive performance, i.e. whether a beginning rumour will “go viral” by obtaining broad coverage of the social network, or else die out.
- Performance on the PHEME use case datasets, which will be measured in two iterations (from M14 and from M26), as detailed in T6.4 on accuracy and scalability evaluation.
- The PHEME visual analytics dashboard will be evaluated for usability in several ways (e.g. heuristic and summative evaluations with users from WP7 and WP8), as detailed in Task 5.5.
- The PHEME integrated framework will be evaluated using relevant architecture and integration metrics, including complexity, criticality, reliability, message rates, network load, response time, CPU usage, memory usage, and effective throughput.
- The use cases will measure using user tests performed with real users and the prototypes, thus reporting qualitative evaluation outcomes (Tasks 7.4 and 8.4);

Dissemination and impact will be measured in multiple ways, including access statistics for the PHEME web site and the download sections of the software tools; number of published papers; number of organised events (both technical and industry-oriented); presentations given and stakeholders contacts; and social network presence.

In terms of **evaluating the exploitation impact** of the project on the commercial and other participants, this is typically somewhat hard to quantify especially during the project itself. Nevertheless, we hope to achieve a significant increase in customers due to the software underlying PHEME services. There is a substantial market potential for veracity intelligence tools, and we believe that the SMEs in PHEME and the associated partners and spin-outs will increase their customer base by a large measure as a result.

B1.3.3 Significant Risks and Associated Contingency Plans

A comprehensive state-of-the-art Risk Assessment and Management Plan will be implemented within the first six months of the project and will address different kinds of risk (external, internal, strategic, operational, other).

In more detail, since the PHEME project mainly deals with software development, carried out by several partners, it is exposed to risks in the following key areas:

- **Technology and Standards** (building on the most promising technology, using and providing necessary standards) – PHEME partners are involved in the relevant standardisation bodies and are at the leading edge of technology in their areas.
- **People** – PHEME has world-leading partners in all areas of expertise needed in the project. Many of them also have an established history of previous collaborations. Finally, the consortium contains the right skill set covering the complete chain of technology providers, integrators, and end-users, to guarantee success.
- **Organisational** – the management structure of the project has been designed carefully to be appropriate for the size of the consortium and the type of the project. The coordinator together with the EC will monitor how management functions and changes can be made swiftly.
- **Architecture and Tools** – PHEME builds on state-of-the-art architectures, standards and tools, developed in successful national and international initiatives. The SMEs and corporations who will be developing the novel PHEME socio-semantic intelligence tools have a strong track record in building successful innovative solutions and in implementing big data analytics projects.
- **Requirements** – the requirements of the project have been chosen carefully to correspond with real market needs. These will be reviewed again at the start of the project by the business case partners (ATOS, ONTO, IHUB, SWI) in order to produce a focused set of real-world use cases. The consortium will also develop a clear exploitation plan addressing diverse business needs.
- **Effort and budget planning** – the project management team will monitor regularly the resources spent on each task and work-package. The plan for the next six months will also be reviewed and adjustments made if necessary.

A careful partner selection, an overall open architecture design and an in-depth knowledge of the market were key considerations during project inception, in order to minimize the risk of failure.

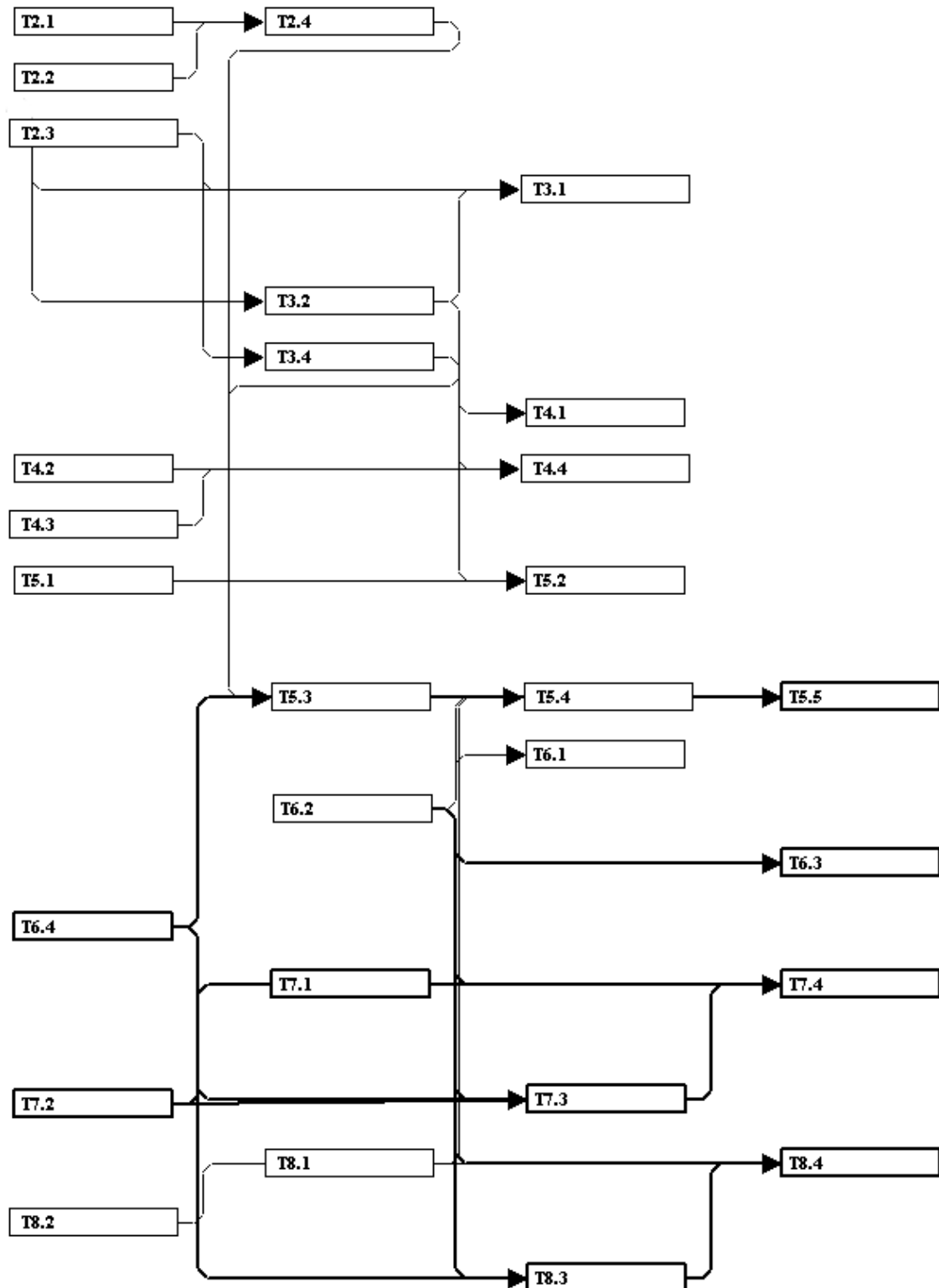
An initial summary of currently envisaged contingency strategies appears below:

Description of possible risk	Impact	Probability of occurrence	Remedial Actions
State of the art environment changes, project loses relevance	Moderate	Low	Technology watch activities planned, as well as close cooperation with other ongoing EC and national projects, to track new technology developments. Changes to work plan made if required.
Organisational financial	High	Low	The Project Coordinator will monitor partners for financial problems and will report to the

problems			Management Group and the EC project officer
Staffing & Recruitment Problems	Low	Low	In most areas of vital importance to the project (WP2, 3, 4, 6, 7s) partners intend to use existing staff. Also, PHEME has more than one partner with expertise in the key areas, which can be seen as risk balancing to avoid delays.
Key staff are ill during critical project time	Low	Medium	Partners have more than one person involved in critical elements of PHEME. Important expertise is available from more than one partner.
Partner leaves the project	Moderate	Low	The consortium is highly motivated and committed to the project. In the unlikely event that a partner leaves, the remaining partners have the skills and capacity to compensate and re-allocate the work load.
Development time is underestimated	Moderate	Medium	Project milestones will help monitor and detect problems early, and take corrective action. The use cases and technology development will be done in parallel, but the former can be re-scoped and re-timed to mitigate against software delays. The technical providers have proven development experience and timely delivery. All key partners were chosen to have track record with EU projects.
Software components fail or attain limited functionality	Low	Medium	The partners are all leaders in the fields covered by the project and have proven existing technology on which service development will be based. In addition, consortium members with complementary technologies have been allocated to the key workpackages WP2-5, thus ensuring technological robustness.
Potential users fail to understand the usability	Moderate	Medium	PHEME has a market driven exploitation and deployment strategy, informed by ongoing market and technology watch activities. The user groups, KCL, IHUB, and SWI as business case partners, will ensure the projects remains inline with user requirements.
Unfortunate choice of standards	Moderate	Low	PHEME tools will use the Software-as-a-Service approach, to mitigate against this risk, as new data input/output standards can easily be added. USFD, USAAR, and ONTO are already involved in the relevant standardisation bodies and will alert the consortium early.
Evaluation is hindered by lack of data	Moderate	Low	The creation of gold-standards for evaluation is an essential part of PHEME. The business case partners already have sufficient data to trial the scalability of the big data analytics services.

Accuracy of the automatic services is not sufficient	Moderate	Low	The evaluation corpora will help us monitor accuracy continuously and identify problems as soon as they arise. Alternative extraction algorithms will be pursued, combined with usage of additional domain-specific resources and engagement of human annotators to produce further training data for the algorithms. A semi-automatic extraction paradigm can also be pursued.
The big data analytics services cannot scale up	High	Medium	ATOS has a proven track record in big data analytics, using Hadoop. WP5 will investigate parallelisation of the PHEME algorithms, to achieve this. Also, through the SwiftRiver platform, applications will filter out irrelevant content, this enabling efficient analysis while maintaining high recall.

B1.3.3 Graphical Interdependencies between tasks (Pert Chart)

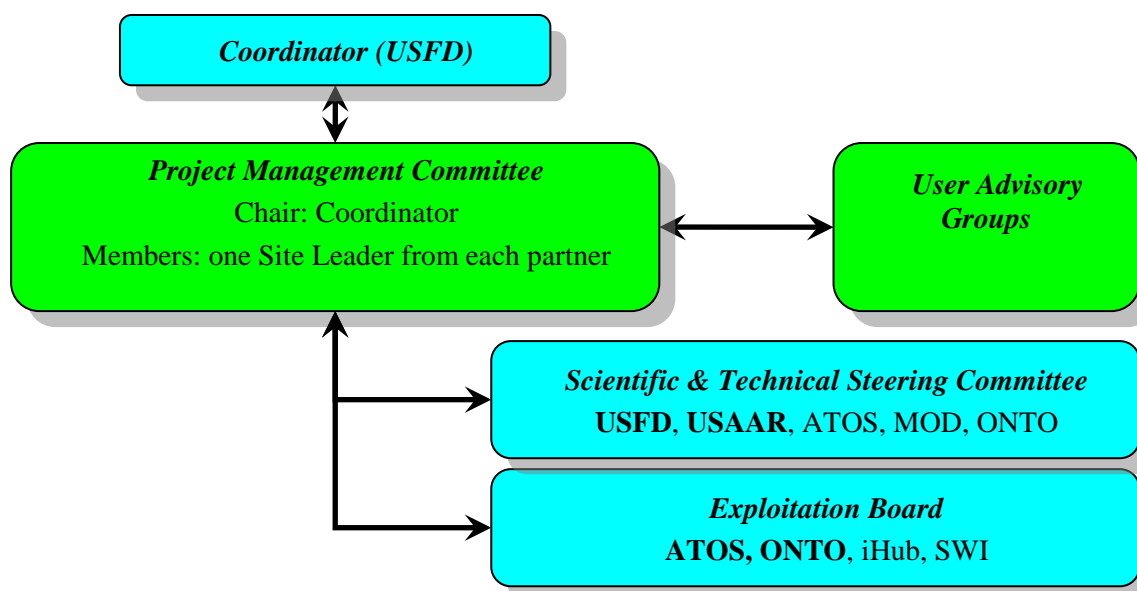


B2. IMPLEMENTATION

B 2.1 Management structure and procedures

The management structure is designed to ensure a clear assignment of responsibility, and effective communication among the partners, so as to achieve successful and timely completion of the tasks within the project. It is intended to be both lightweight and responsive.

The cross-disciplinary nature of this proposal, coupled with the strong industrial involvement and focus on big data analytics, necessitates the involvement of nine project partners, all with clearly defined, complementary roles. Therefore, the necessary management structures, to match a project of this size, have been put in place by USFD (see figure below). To ensure smooth cooperation, USAAR will assist with the scientific leadership, whereas ATOS will lead the industrial integration and exploitation activities. The Partners have worked in the past before, in national and/or European projects, which together with USFD's established consortium management track record will ensure smooth project progress and delivery against all objectives and key performance indicators.



The Coordinator: The University of Sheffield (USFD) will act as co-ordinator, with overall responsibility for the management and administration of the project.

The coordinator, Dr Kalina Bontcheva, and the USFD team as a whole have extensive experience in EU project management, including the management, administration and legal compliance responsibilities of successive Framework programmes. At the organisational level, this is supported by an administrative office that provides assistance in the management and coordination. A professional project administrator will be assigned to handle the financial aspects of the project, and other administrative support as is necessary.

USFD will be responsible for the preparation of budgets, expenses, task allocations and compliance to the overall plan. It will also be responsible for aspects such as obtaining audit certificates from project partners, and for the distribution of payments from the commission to the partners. USFD operates regular checks on the financial performance of all European projects, and will collect the summary statements of expenditure from all the project partners. By this means, financial planning and forecasting for the project will be implemented effectively and transparently.

The coordinator will be responsible for monitoring the activities within the project, for its overall technical consistency and quality, for communication with the Commission and other relevant parties.

The Project Management Committee (PMC): Each partner will assign a **site leader**, responsible to the coordinator for the contributions from that site. The coordinator and the site leaders will together form the **Project Management Committee (PMC)**, which will meet three times a year in conjunction with the PHEME project meetings.

The PMC will be chaired by the coordinator. It is responsible for the overall success of the project. In the event of disagreements, these will be resolved by a majority vote, with each partner casting one vote; the coordinator will have the casting vote if necessary.

Since all partners are involved in the PMC, it will be directly informed about all the activities of the project and its decisions will be able to be implemented speedily.

The Project Management Committee will review the overall progress of the project, acting as the common forum for discussion and a body for high-level decisions such as: approval of budgets and work plans; approval of major changes in the work of the project; changes in the consortium; proposed changes to the contract or the Consortium Agreement; suspension or termination of all or part of the project or of the contract; and actions to be taken in the case of misconduct of a partner.

Work Package Leaders: Each work package will be assigned a **work package leader**, responsible for continuously monitoring and ensuring progress with the tasks in the work package, coordinating efforts between the partners involved, and managing the timely generation of deliverables.

The Scientific and Technical Steering Committee (STSC) will be chaired by USFD, who will be assisted by Thierry Declerck (USAAR) as scientific research leader. It will ensure the timely progress of the project and the high quality of the results. It reports to the Project Management Committee and will meet at least three times a year. Between meetings it will be permanently in contact, through skype and e-mail. Its responsibilities are detailed in the consortium agreement.

The Exploitation Board (EB) will comprise all the industrial partners in the consortium (ATOS, ONTO, SWI, and IHUB) and will be led by ATOS, who have extensive track record in successful market watch and exploitation leadership in FP6 and FP7 projects. The Exploitation Board will be responsible for ensuring that the work of PHEME is prepared for future exploitation. It will oversee the production of market analysis reports. On the basis of these it will ensure that the market is fully addressed by the combination of the industrial partners. It will oversee the dissemination of best practice throughout Community countries. To do this it will fully exploit the case studies. The Exploitation Board will meet at least quarterly and it is likely that these meetings will be co-located with the Project Management Committee meetings.

The Consortium Agreement will lay down the events which can trigger a change in the composition or the chair of the PMC, STSC, or the Exploitation Board. The Consortium Agreement will also define the procedures to be followed in such cases.

B2.1.2 Management Issues and Procedures

Consortium Agreement: Prior to the start of the project, the Project Coordinator will ensure that all partners have signed the project "Consortium Agreement". This document will set out all the internal rules of the project and will be signed by all partners prior to the signature of the Grant Agreement. The consortium agreement will adopt a set of rules for ensuring the quality of each partner's work and will be the base for a further Exploitation Agreement document and a subsequent Exploitation Plan.

The obligations and rights of the participants will be regulated according to the standard rules contained in the contract signed with the Commission and detailed in the Consortium Agreement. This Consortium Agreement will make explicit reference to important administrative points such as: decision procedures within the project; risk management strategies; legal aspects regarding software to be used/produced in the project; trademarks, patents, etc.; the right of each partner in the exploitation of results; equal opportunities and gender equality policies.

Quality Management: Within PHEME, a Quality Assurance System will be established in the earliest stages of the project, with a Quality Assurance Plan to be delivered at Month 3, to be used

internally by the Consortium. It will describe the guidelines adopted by the project on preparation and validation of deliverables, internal peer reviewing, periodic reporting, preparation of financial statements, as well as risk management.

Particular emphasis will be placed on quality control of deliverables. Each deliverable will be reviewed externally by a partner who did not take part in writing the deliverable, if possible external to the WP in which the deliverable is produced. In this way problems can be detected early, reported to the technical and scientific committee and acted upon quickly. Internal submission deadline for deliverables will thus be three weeks ahead of the official date to allow time for the quality procedure.

Reporting and Communication: Each partner submits to USFD a report every 6 months detailing progress and effort expenditure. Work package leaders also produce a report on WP progress. These reports cover: activities during the reporting period; major achievements; dissemination (publications, conferences and workshops); project and work package meetings; deviations from the plan, risks and issues affecting timely delivery, cost or quality; next steps and foreseen actions. These reports are for the internal use of the project and the EC's project officer. Yearly, each partner prepares a financial statement, which is collated and presented to the EC by USFD.

A project plenary meeting will be held two to three times a year. WP leaders may organize work package meetings for a single or group of work packages as the need arises. The PMC and the Exploitation Director may arrange further face-to-face and telephone conferences as and when needed.

Several tools and services will be made available to the PHEME partners not only to assist them in their work, but also to stimulate their cooperation. First, a web based reporting, via Google Drive will be used by USFD to produce standard reports. This will simplify the collection of information across different teams and work packages in an easy manner to limit the administrative charges of partners. Secondly, dedicated mailing lists will be created and archived. Thirdly, the project will rely extensively on on-demand audio-conferencing (every 3 weeks) to discuss progress and identify issues.

Conflict Resolution: The general principle applied to the resolution of disputes is that steps to resolve any dispute should be taken at the most appropriate level, and referred to the next highest level in the organizational structure should satisfactory resolution not be achieved. Ultimately, any unresolved dispute is referred to the Management Group who, in the event that an amicable agreement cannot be reached within that body, will have the right to appoint external independent arbitrators.


Risk Management: The PHEME project management team will perform continuous evaluation throughout the project, identifying any possible problems/risks at an early stage so that solutions can be elaborated in time. A systematic approach will be adopted for monitoring resource spending against project budget and achievements against schedule and critical success factors. A first list of risks together with their contingency plans has already been reported above.

Meetings: The beneficiaries will ensure adequate representation at the following meetings:


Type of meeting	Purpose	Participants	Venue
Project kick-off meeting	To launch the project and refine plans and arrangements for the initial implementation phase.	Consortium members and Project Officer	Luxembourg or suitable project site, to be decided in agreement with the Project Officer

Type of meeting	Purpose	Participants	Venue
Progress meeting	To review progress and discuss any significant problems and deviations.	Coordinator and Project Officer	Luxembourg or suitable project site, to be decided in agreement with the Project Officer; can be handled by video conference
Review meeting	To evaluate intermediate and final results. To assess quality, impact and effectiveness of project work.	Coordinator and relevant work package leaders, Project Officer, Peer Reviewers	Luxembourg or suitable project site, to be decided in agreement with the Project Officer
Concertation meeting respectively Programme conference and exhibition	To actively participate in discussions and demonstrations organised by the ICT programme. To present work in progress and demonstrate intermediate results. To identify and discuss areas of common interest. To plan joint investigations and dissemination activities.	Coordinators of consortia and/or work package leaders, plus external experts, suppliers and users where appropriate	To be defined


B2.2 Beneficiaries

1	The University of Sheffield	
USFD / UK / RES	<p>Organisation Profile</p> <p>The Natural Language Processing (NLP) group at the University of Sheffield is one of the largest and most successful research groups in language and information in the EU. The group is based in the Department of Computer Science, and includes world-class teams in the areas of speech, language, knowledge and information processing, biotechnology, and machine learning for medical informatics. The Department of Computer Science was founded in 1982 and since then has established national and international renown for its teaching and research. It was awarded a top Grade 5 in the most recent nationwide Research Assessment Exercise. Companies that support its work include Daimler-Chrysler, GlaxoSmithKline, Motorola and Nokia.</p>	
	<p>The Natural Language Processing Group has focused on robust engineering of open source NLP software and on quantitative evaluation and repeatability. The group has extensive experience in the fields of NLP infrastructures (GATE), information extraction, machine learning methods, dialogue systems, question answering, terminology extraction, NLP methods for Knowledge Management and the Semantic Web. USFD has a world-leading research record on human language technologies, developed within national and international research projects in these areas. Our participation in Arcomem (opinion mining from social media), TRENDMINER (mining and summarising trends in media streams), AnnoMarket (text mining services marketplace), and LARKC (large-scale reasoning and web search) will form a solid base to build on here.</p>	


	<p>Role in the project: USFD will coordinate the consortium, where the team has an excellent track record. USFD will also lead WP3 on contextual interpretation, where Cohn and Lampos will develop story and conversation detection algorithms and longitudinal models of users, trustworthiness, and influence, based on Lampos’s pioneering research on detecting epidemics on Twitter. USFD will also adapt EN linguistic tools (T2.3) and build novel multilingual spatio-temporal annotation tools (T2.4), based on Derczynski’s work on English and TempEval. USFD will also develop the misinformation and disinformation detection algorithms in WP4.</p>
	<p>CVs of key personnel</p> <p>Dr. Kalina Bontcheva is a senior research scientist and the holder of a prestigious EPSRC career acceleration fellowship, working on text summarisation of social media. Her main interests are information extraction, opinion mining, natural language generation, text summarization, and software infrastructures for NLP. She has been a leading developer of GATE since 1999. She coordinated the EC-funded TAO STREP project on transitioning applications to ontologies, as well as leading the Sheffield teams in TRENDMINER, MUSING, SEKT, and MI-AKT projects.</p> <p>Dr. Trevor Cohn is working on the interface between machine learning and NLP. Dr. Cohn holds a PhD on the topic of supervised machine learning approaches for modelling structured prediction problems in natural language. He worked for 3 years at the University of Edinburgh on an EPSRC project on text summarisation and paraphrasing, while also collaborating on second EPSRC on machine translation, which he authored while still a PhD student. Cohn’s primary research interest is in developing machine learning approaches for prediction problems in NLP, machine translation, parsing, and summarisation. He has published extensively in high profile journals and conferences in both the fields of machine learning and natural language processing.</p> <p>Dr. Leon Derczynski is a research associate, who finished a PhD in Temporal Information Extraction at USFD in 2012 under an enhanced EPSRC doctoral training grant. His main interests are information extraction, spatio-temporal semantics and handling noisy linguistic data. Leon Derczynski contributes to ISO consortiums for spatio-temporal annotation (TC37/SC4) and co-organises TempEval, the established evaluation challenge steering the state-of-the-art in temporal information extraction technology.</p> <p>Dr. Vasileos Lampos has a PhD from Bristol University on the topic of detecting flu epidemics from Twitter messages (Lampos <i>et al</i>, 2010). He is currently working on the TRENDMINER project on regressions models for discovering trends and regional and demographic variations.</p>

2	<p style="text-align: center;">University of Saarland</p>	 <p>UNIVERSITÄT DES SAARLANDES</p>
USAAR / DE / DEC	<p>Organisation Profile</p> <p>University of Saarland (USAAR) is represented by the faculty of computational linguistics and phonetics (COLI). Internationally renowned, the faculty and its staff have made Saarbrücken one of the leading centres for language science study and research world-wide. Focus is on computational linguistics, psycholinguistics, phonetics, and spoken language systems.</p> <p>Complementary to these, there are four independent research groups (Cognitive Models of Human Language Processing, Computational Modelling of Discourse and Semantics, Machine Learning for Natural Language Processing and Multimodal Speech Processing) focusing on further aspects of NLP.</p> <p>The department offers international degree programs at all levels, with state-of-the-art courses and ample opportunities to work with leading researchers.</p>	


	<p>The department's internationally leading research projects are supported by a range of national and European funding agencies. It is part of the Cluster of Excellence "Multimodal Computing and Interaction" at Saarland University. USAAR also has a long-standing collaboration with the German Research Center for Artificial Intelligence (DFKI).</p> <p>Project management at COLI is conducted by a professional office; http://eurice.eu/about/ European research and project office GmbH.</p>
	<p>Role in the project: Leader of WP2 working on multilingual language processing tools, with focus on German, and content geolocation. Pre-existing linguistic components for German and modelling temporal information. T4.2 on detecting disputed information through entailment and contradictions. Contribution to the case studies on healthcare and digital journalism.</p>
	<p>CVs of key personnel</p> <p>Thierry Declerck is a senior research scientist, leading for USAAR two European Projects (Esperanto and INTERA), dealing with the relation between language technologies and Semantic Web and with infrastructures for Language Resources. As a member of DFKI, he is currently co-ordinating the FP7 project TRENDMINER, and in PHEME he will ensure that synergies will be created, ensuring complementarities and avoiding duplication of work. His main topics of interest are Ontology-based information extraction, opinion mining, temporal IE, and HLT for eHumanities. Relevant for PHEME is his work on ontology-based NLP for folktales, stories in which non-factual entities and events can occur.</p> <p>Dr Vera Demberg. Since 2010 she is heading the Junior Research Group "Cognitive Models of Human Language Processing and their Application to Dialogue Systems". The research interests of this group are Cognitive Models of Human Language Understanding, Multimodality in Language Processing, Parsing and Grammar Formalisms and Experimental Psycholinguistics. Vera Demberg is currently supervising closely the PhD work of Fatemeh Torabi Asr (Discourse coherence and discourse relations) and the Postdoc work of Asad Sayeed (information retrieval, syntactic theory, sentiment analysis).</p> <p>Prof Dr Stephan Busemann. In 2011 Stephan Busemann became Honorary Professor of Computational Linguistics at the University of the Saarland, and he is at the same time Associate Head of DFKI's Language Technology Lab, where he was until recently responsible for the coordination of the European EUROMATRIXPLUS and ACCURAT projects. At USAAR he was also involved in the DeepThought project.</p>

3	MODUL University Vienna	
MOD / AT / RES	<p>Organisation Profile</p> <p>The Department of New Media Technology at MODUL University Vienna (MOD) conducts cross-disciplinary research on the integration of semantic and geospatial visualisation technologies, human-computer interaction and visual analytics, with a special focus on large-scale media monitoring and Web intelligence applications. The research group develops Web intelligence applications for government agencies and corporate partners in both Europe and the United States, for example, and showcases its applied research results through award-winning Web and social media applications including the <i>Media Watch on Climate Change</i> and related crowdsourcing applications (http://www.ecoresearch.net/triple-c). MOD's group of researchers has extensive experience in managing large-scale research projects and couples technological expertise with a strong background in the social and economic sciences.</p>	
	<p>Role in the project: Leading partner for WP5 on interactive visual analytics for veracity intelligence. T3.3 on detecting implicit information diffusion networks across media, based on prior research (Sharl <i>et al</i>, 2007). Contribution to the case study in patient care.</p>	
	<p>CVs of key personnel</p> <p>Prof. Arno Scharl is Head of the Department of New Media Technology at MODUL University Vienna. Prior to this appointment, he held professorships at the University of</p>	


	<p>Western Australia and Graz University of Technology, and was a Visiting Fellow at the University of California at Berkeley. Prof Scharl completed his doctoral research and habilitation at the Vienna University of Economics and Business. He has coordinated several award-winning semantic systems projects including IDIOM, RAVEN, and Triple-C (Climate Change Collaboratory), authored more than 130 refereed publications, and edited two books on The Geospatial Web and Environmental Online Communication, respectively. His current research interests focus on media monitoring and Web intelligence, information diffusion, and the integration of semantic and geospatial Web technology.</p> <p>Dr. Marta Sabou is an Assistant Professor at the Department of New Media Technology. Prior to this, she was a researcher at the Knowledge Media Institute, Open University and holds a PhD from Vrije Universiteit Amsterdam. She won the IEEE Intelligent System's Ten to Watch Award (2006). She has worked in European projects on semantic technologies: WonderWeb, KnowledgeWeb, NeOn (task lead), OpenKnowledge and SmartProducts (WP leader on ontology-based user profiling – an expertise that will support the user profiling work in this project). Her work has been published in high-impact conferences (WWW, Int. Semantic Web Conferences), the Journal of Web Semantics, and IEEE Intelligent Systems.</p> <p>Dipl.-Ing. Stefan Gindl is a Researcher and Lecturer at the Department of New Media Technology, and a founding member of the Interest Group on German Sentiment Analysis (IGGSA). His research focuses on sentiment analysis with an emphasis on context-awareness and incorporating common sense into existing approaches. During his previous studies of Medical Informatics at the Vienna University of Technology, he investigated negation detection in the medical domain. Specifically, he developed and implemented algorithms suitable for the detection of negated phrases in clinical practice guidelines.</p>
--	---

4	Ontotext AD	
ONTO / BG / SME	<p>Organisation Profile</p> <p>Ontotext is a semantic technology lab of Sirma Group. It was founded in year 2000 and at present employs about 30 researchers and engineers. Ontotext is focused on research and core technology development for knowledge discovery, management, and engineering, Semantic Web and Web Services. Ontotext's technology delivers real-world applications in Web Services and Enterprise Application Integration, Knowledge Management and Text-mining, Business Intelligence, Life Science, and Media Research. Ontotext is developer of several outstanding Semantic Web tools, including: the KIM semantic annotation platform and OWLIM – the fastest and most scalable OWL semantic repository. Ontotext is also a major contributor to few of the most popular open-source projects in the area: GATE (language engineering platform) and Sesame (RDF semantic repository).</p> <p>SIRMA (www.sirma.com), established in 1992, is a group of diverse, privately-owned software businesses with major offices in Bulgaria (Sofia, Plovdiv, Varna, Rousse), Canada (Montreal). It includes more than 10 companies and business units. Sirma is one of the oldest and biggest software houses in Bulgaria, at present top-3 software producer with around 200 employees. In 2008 Ontotext secured external funding – NEVEQ (a VC fund) acquired a minority share in a deal for 2.5M€ Based on public information, this is the biggest investment in semantic technologies in Europe for the year. In order to accommodate the investment, Ontotext is in process of establishment as a separate company.</p> <p>OntoText has participated in a number of EU projects, among which the following: TRENDMINER which aims to deliver innovative, portable open-source real-time methods for cross-lingual mining and summarisation of large-scale stream media; AnnoMarket which develops a text annotation marketplace; RENDER which aims at developing methods, techniques, software and data sets that will leverage diversity as a crucial source of innovation and creativity; CUBIST which aims at supporting federation of data from a variety of unstructured and structured sources; having a BI enabled triple store as an Information</p>	

	Warehouse; using semantic information to improve BI best practices; enabling a user to perform BI operations over semantic data; MOLTO (Multilingual Online Translation) which aims at delivering pluggable open-source libraries enabling standard translation tools and workflows.
	Role in the project: Leading partner for T4.1, working on reasoning about rumours, using LOD world-knowledge. Leading WP2 leader, working on ontological modelling of rumours (T2.2) and language processing tools for Bulgarian. Involvement in integration and the patient care use case.
	<p>CVs of key personnel</p> <p>Georgi Georgiev is a PhD graduate in Molecular Biology and Bio-Physics at the Sofia University. His current research interests include knowledge representation, information extraction and mining of text in various domains. Since 2007 Georgiev is an information extraction architect and leads the text analysis group at Ontotext AD. Georgi has more than 30 scientific reports and publications, participates in various organization committees on scientific events and is reviewer at text analysis and mining conferences. He is member of the Federation of the European Biochemical Societies and Bulgarian Science Union.</p> <p>Dr. Petya Osenova has a PhD in Linguistics from the Institute of Bulgarian Language, Bulgarian Academy of Sciences in 1999. She participated in BulTreeBank, a joint project with the University of Tuebingen. She had a one year postdoctoral project (2004) on measuring language contacts together with the Groningen University, Holland. In 2010 she won a Fulbright grant at Stanford University, where she developed a first version of a Deep HPSG grammar for Bulgarian. She has participated in EU projects on eLearning and Machine Translation. Her interests are in the area of natural language semantics and pragmatics, lexical knowledge bases, formal grammars, corpus linguistics, and question answering.</p>

5	Atos Spain SAU	
ATOS / ES/ COMPANY	<p>Organisation Profile</p> <p>Atos is an International Information Technology Services company with annual revenues of EUR 8.7 billion and 78,500 employees in 42 countries. Serving a global client base, it delivers hi-tech transactional services, consulting and technology services, systems integration and managed services. Atos focuses on business technology that powers progress and helps organizations to create their firm of the future. It is the Worldwide Information Technology Partner for the Olympic Games and is quoted on the Paris Eurolist Market.</p> <p>Atos Research & Innovation (ARI) is the research, development and innovation hub of ATOS and it is a key reference for the whole Atos group. ARI is located in six cities: Madrid, Barcelona, Bilbao, Santiago and Valladolid in Spain and Istanbul in Turkey. ARI is organised in fifteen Sectors addressing the needs of the 5 well established markets of the company (e.g. Public, Health & Transport, Finance Services, Telecom, Media & Technologies, Energy& Utilities, and Manufacturing, Retail & Services). From a technological point of view, ARI consists of eight Labs developing new technologies in their respective innovation fields. ARI is one of the key players of Future Internet in Europe, being member of the Steering Committee of the FI PPP, EFIA (European Future Internet Alliance) and Vice-President of ES.INTERNET, the Spanish Platform of Future Internet.</p> <p>ARI is a founding member of the several European platforms and initiatives, such as NESSI, eSafety Forum, Net!works, ARTEMIS, NEM, EOS and Nanomedicine, and also at a National Spanish level (LOGISTOP, eMOV, Railway, Maritime, eSEC PROMETEO, INES, etc.). ARI has performed around 300 R&D&I projects since 1987. ARI is represented in the project by the Knowledge Lab, specialized in Big Data, semantics and language technologies, with wide ample experience in EU projects (BIG, FIRST, Khresmoi, VIRTUOSO, VPH-Share, TaToo, NeOn, LUISA, TAO, SOA4All, IntelLEO, among others).</p>	

	<p>Role in the project: Leading WP6 on integration, scalability, and evaluation. Expertise in building big data analytics systems with Hadoop and Storm. Leaders of the exploitation board and the commercial exploitation activities in PHEME.</p>
	<p>CVs of key personnel</p> <p>Tomás Pariente is project manager and technical coordinator for EU-based projects in semantic and big data technologies in ATOS Research and Innovation, and he is currently Head of the Knowledge Lab. His technical expertise is mainly in Big Data, semantic technologies and knowledge management. Tomás is involved in several working groups in this technology. Tomás is currently coordinating the EU FP7 project FIRST, and participating in BIG, a FP7 coordination action on Big Data. He worked on EU projects such as Ontologging, SmartGov, OntoGov, INFRAWEBs, TAO, NeOn, LUISA, SOA4ALL, FIWARE and TaToo, among others. He also has participated and coordinated several Spanish R&D projects.</p> <p>Iván Martínez is a senior researcher in Atos Research and Innovation. He graduated in Computer Science from Technical University of Madrid. In recent years Iván participated in natural language processing, cloud computing, Semantic Web and Semantic Web Services related projects. He has contributed to national research projects such as PLATA, and other European Semantic Web Services related project, such as SUPER and SOA4ALL. Previously in GLOCAL project and currently in KHRESMOI and VPH-Share projects, he is leading in the latter's definition and integration of system architecture.</p>

6	<h2 style="margin: 0;">King's College London</h2>	 <p style="font-size: small; margin: 0;">University of London</p>
KCL / UK / RES	<p>Organisation Profile</p> <p>King's College London (KCL) is one of the top 30 universities in the world (2012/2013 QS international world rankings), is in the top seven UK universities for research earnings, and is the fourth oldest in England. King's has an outstanding reputation for providing world-class teaching and cutting-edge research. In the 2008 Research Assessment Exercise for British universities, 23 departments were ranked in the top quartile of British universities.</p> <p>KCL links closely with South London and Maudsley NHS Foundation Trust (SLAM) and with its two other neighbouring Acute Healthcare providers at Guys and St Thomas' Hospitals and Kings College Hospital under the Kings Health Partners Academic Health Sciences Centre – one of only five national centres dedicated to bringing together academic research with clinical practice. The close KCL-SLAM links on the NIHR Biomedical Research Centre for Mental Health (BRC) will bring the needs of medical practitioners into a tight loop with the project's science and technology development programme. SLAM is the largest mental healthcare provider in Europe, serving a geographic catchment of 1.2 million residents, and plays a significant role in the development of health policy at a national level. The SLAM electronic patient record thus forms the basis of Europe's largest mental health case register. The Clinical Record Interactive Search (CRIS) allows databases to be assembled from the fully electronic patient records system which has been operating across all services provided by SLAM since 2006, extracting bespoke anonymised data for secondary analysis including de-identified free text fields if required. The data resource contains full anonymised clinical records on over 200,000 service users, over 35,000 of whom are receiving active case management at any given time. The total size of the database in SQL is approximately 33gb, of which around 25gb is attributed to text fields (15 million rows). The BRC has demonstrated long-standing commitment to developing CRIS and a wider Clinical Informatics programme, including the incorporation of external data linkages, the development of a shared electronic record (in collaboration with Microsoft), and the application natural language processing to derive structure from patient records using GATE, the latter being the product of a long-standing and productive collaboration with the University of Sheffield (USFD) team. PHEME will show how the technologies developed can be applied to a health-related use case, and how social media</p>	


	analysis can be integrated with public health monitoring and with analysis of the EPR.
	Role in the project: KCL will lead WP7, which will adapt and apply the project technologies to improving patient care. The BRC team will provide technological requirements from the perspective of medical practitioners, help annotate relevant datasets, adapt the PHEME tools and algorithms to the needs of the case study, and evaluate the results using the CRIS register.
	<p>CVs of key personnel</p> <p>Prof. Robert Stewart leads the Clinical and Population Informatics theme of the SLAM BRC. He has worked as an Academic Psychiatrist at the Institute of Psychiatry, King's College London since 1996 specialising in Old Age Psychiatry and Psychiatric Epidemiology, assisted by a Wellcome Trust Research Training Fellowship. Robert Stewart was appointed as Clinical Senior Lecturer in 2003, as Head of the Section of Epidemiology in 2006, as Clinical Reader in 2007 and as Professor of Psychiatric Epidemiology and Clinical Informatics in 2012. Prof. Stewart has led the academic development of the SLAM BRC Case Register and the Clinical Record Interactive Search (CRIS) application since its development in 2007/8 overseeing its rapid expansion in size and output.</p> <p>Prof. Stewart will be assisted by the following members of his research team, all of whom have extensive experience both in research using CRIS and in the application of GATE for natural language processing in this data resource: Dr Alex Tulloch, Clinical Senior Researcher; Drs Chin-Kuo Chang and Richard Hayes, Senior Post-Doctoral Researchers; Dr Gayan Perera, Post-Doctoral Researcher; Mr Mike Denis, SLAM Head of Information; Mr Matthew Broadbent, CRIS Project Manager. In addition, the project will benefit from the wider cross-disciplinary infrastructure provided by the SLAM BRC, most notably expertise in individual clinical specialties, but also a mature BRC Patient and Public Engagement theme to ensure that project design, implementation and dissemination receive maximally effective service user input.</p>

7	iHub	
IHUB / KEN / SME	<p>Organisation Profile</p> <p>iHub - Nairobi's Innovation Hub for the technology community is an open space for the technologists, investors, tech companies and hackers in the area. This space is a tech community facility with a focus on young entrepreneurs, web and mobile phone programmers, designers and researchers. Two of iHub's divisions are relevant to PHEME: iHub Consulting and the iHub Cluster. iHub Consulting pools together top talent from the community to help organizations develop and implement technology strategies and solutions for long term growth. We leverage local skills to deliver tech innovation to organizations where it is needed. In this way, we believe we are catalysing the movement of innovation to where it matters most. The iHub Cluster is building an HPC cluster to achieve: ACCESS - Bringing Super Computing technology and knowhow that is high in academia or deep in institutions to iHub projects; SCALE - To be at the forefront of the changing computing landscape in Africa that delivers on the technology and develops the human capacity to maximise the technology</p> <p>Hub Consulting is part of the community behind the Ushahidi Platform: free, open source, web-based software for leveraging crowdsourced data for crisis response, civic engagement, and more. The roots of Ushahidi are in the collaboration of Kenyan citizen journalists during a time of crisis. The current team comprises individuals with a wide span of experience ranging from human rights work to software development.</p> <p>They will bring into the project expertise in developing and integrating Ushahidi 3.0, along with new versions of Crowdfunder and SwiftRiver, newer and complimentary software products, aimed at making sense of social data at large scale.</p> <p>Role in the project: Implementation of the open-source digital journalism dashboard, which</p>	

	<p>will connect the open-source SwiftRiver information filtering and content curation platform to the PHEME veracity intelligence tools. Outreach and exploitation activities aimed at existing media users of the Ushahidi platform, including the Guardian, the BBC, and Al Jazeera.</p>
	<p>CVs of key personnel</p> <p>Robert Baker brings over ten years of experience as a web and new media developer, trainer, and manager as IHUB's Project & Outreach Manager, responsible for documentation, logistics, software development, and working with clients. Before officially joining IHUB, his contributions to their community earned him the first ever induction to the Trusted Developer Network for his work as technical or project lead on dozens of IHUB deployments from crisis response to civic engagement around the world as well as the creation of the Ushahidi Community website. In addition, he acted as Director of the Universities for Ushahidi program, a 2011 initiative to train students from around the world on mobile and mapping technology. When he's not working on IHUB projects, Rob is also an active member of the Humanitarian OpenStreetMap Team (HOT).</p> <p>Nathaniel Manning is Director of Business Development. Nathaniel's work orbits around the theme of developing technology that makes the world a better place. He worked for the Clinton Climate Initiative (CCI) for three years throughout Asia, where he helped to craft business models that make clean energy financially feasible in emerging nations. He focused on solar in Malaysia, rural electrification in India, waste management in Indonesia, and energy efficiency in Thailand. He has advised companies such as First Solar and Virgin on market strategy and is a LEED (Leadership in Energy and Environmental Design) Accredited Professional. Nathaniel holds a BA in Philosophy and MA in International Environmental Policy from Brown University.</p>

8	swissinfo.ch	swissinfo.ch
SWI / CH / SME	<p>Organisation Profile</p> <p>swissinfo.ch is the international service of the Swiss Broadcasting Corporation (SBC), which is a private media organisation. Since 1999, swissinfo.ch has fulfilled the federal government's mandate to distribute information about Switzerland internationally, supplementing the online offerings of the radio and television stations of the SBC. Today, the international service is directed above all at an international audience interested in Switzerland, as well as at Swiss citizens living abroad.</p> <p>The online service offers a Swiss view of topics and highlights, Swiss positions on international events and developments, while reflecting the view of Switzerland from abroad. swissinfo.ch's coverage focuses on politics, business, culture, society and research. swissinfo.ch also provides specific information for the Swiss abroad to assist them in exercising their political rights in Switzerland (vote dossiers).</p> <p>As of 2013, swissinfo.ch will provide reports in Russian—in addition to English, German, French, Italian, Spanish, Portuguese, Chinese, Arabic and Japanese—thereby reaching more than 80% of the world's internet users.</p> <p>swissinfo.ch has a wide experience with social media platform: Facebook, Twitter, Google plus, Kaixin.</p> <p>swissinfo.ch has its offices in Bern (headquarter), Geneva and Zurich, and is represented in parliament's media centre in the nation's capital. swissinfo.ch is present with mobile APPs as Apple, Android and Windows.</p> <p>Swissinfo has an annual budget of 17 Mio CHF (14 Mio Euro). Staff head count is 110 (=86 FTE). The head count of journalists, working in 10 languages is 75, translators and freelancers not included. URL. www.swissinfo.ch.</p> <p>Role in the project: SWI will lead WP8, which seeks to bring the project technologies to bear</p>	

	<p>in the digital journalism domain by providing technological requirements from newsroom journalists, multilingual news content dating back to 2000, annotated gold standard data, and evaluating the algorithms against well known circulating rumours.</p>
	<p>CVs of key personnel</p> <p>Dr. Peter Schibli is Director of swissinfo since 2008. He graduated from the University of Berne with a Dissertation in Public Law and has a Masters Degree from George Washington University (DC). Mr. Schibli has a long career in Journalism. He worked as a political correspondent in Bern, Bonn, Berlin and Washington D.C. before becoming Editor in Chief of swissinfo. Besides his function at swissinfo.ch Mr. Schibli is National Multimedia Coordinator for the Swiss Broadcasting Corporation (SBC). In both functions he reports to the Director General of SBC, Mr. Roger de Weck. Peter.schibli@swissinfo.ch</p> <p>Christophe Bruttin is Product Manager at the Marketing Dept. of swissinfo.ch and a specialist in Social Medias. The SM-activities of swissinfo are run under his concept. Before he joined swissinfo, Mr. Bruttin was a Marketing expert for Alinghi (sailing) in Valencia until 2009. Christophe.bruttin@swissinfo.ch</p> <p>Hubert Zumwald is head of the IT department at swissinfo. He has worked for the IT department of SBC and for IT of Mobiliar Insurance Company, before he joined swissinfo.ch in 2012. Mr. Zumwald has a Master Degree in Business Innovation. Hubert.zumwald@swissinfo.ch</p>

9	University of Warwick	
UWAR / UK / RES	<p>Organisation Profile</p> <p>The University of Warwick is one of the UK's leading research-led universities and was ranked seventh overall in the UK by the latest Research Assessment Exercise (2008), and is consistently ranked in the Top Ten UK Universities in other league tables. Recent large-scale funding has resulted in the establishment of cross-faculty centres of research, which bring together emergent technologies and target disciplines. The Department of Computer Science is one of the leading Computer Science Departments in the UK for both research and teaching. The Department was awarded international excellence status in the last three UK Research Assessment Exercises (1996, 2001 and 2008) and has the state-of the-art computing facilities to support the needs of the project. Research in the Department encompasses a variety of topics, ranging from advancing the foundations of computing to exploring novel, interdisciplinary applications. The activity is strengthened by a range of collaborations, including within the University, nationally and internationally. The department is a partner in the Centre for Urban Studies and Progress (CUSP), a research institute created by New York University in conjunction with a consortium of world-class universities and world-leading tech companies, to tackle the array of complex challenges facing cities in the 21st Century.</p> <p>The Social Informatics Group (SIG) is a research group with the Department of Computer Science, led by Prof Rob Procter. The group's research focuses on the sociology of technology and innovation, in particular, how cognitive, organisational and social factors shape the design, development, adoption and use of information and communication technologies (ICTs). The applications strand of this research seeks to work with user organisations to develop and deploy innovative applications of distributed, digital infrastructure and tools. The social shaping strand adopts a social studies of science and technology approach to understand how distributed, digital infrastructure and tools are being developed, how they are used and their implications for individuals, organisations and communities. Prof. Procter will be playing a leading role in Warwick University's contribution to CUSP.</p> <p>A major focus of current work is on social media analysis. Procter's work with the Guardian Newspaper, which focused on analysing tweets from the time of the August 2011 riots in England, has been widely cited and has led to collaborations with various UK government</p>	

	<p>agencies, including police forces (via the National Policing Improvement Agency), Scottish Government Justice Analytic Services Unit, the BBC World Service and several third sector organisations. SIG has funding from the UK JISC to develop a scalable social media analysis workbench for harvesting and analysing social media for use by academic researchers. Procter has also worked with the Guardian Newspaper in a study of how journalists use social media as source of information for news stories and with the BBC World Service in an analysis of the use of social media as a tool for increasing audience engagement and programming impact.</p>
	<p>Role in the project: UWAR will contribute their expertise in social informatics. They will undertake the qualitative analysis of rumours across media and languages, using social science methods (T2.1) and also contribute to the journalism use case (WP8).</p>
	<p>CVs of key personnel</p> <p>Prof Rob Procter has more than 25 years experience of inter-disciplinary research, and has worked on more than 50 projects. He led the social media study of the August 2011 riots and is the Principal Investigator on the JISC funded analysing social media project and the ESRC DSR Community Fund projects 'Training and dissemination for Social Media Analysis' (hosting events to raise awareness of the potential of innovative social media analysis methodologies among UK researchers and build capacity in their application) and 'Helping the Police with their inquiries' (bringing together social media researchers and representatives of UK Police forces for a series of meetings to discuss lessons from the 2011 riots). He is a co-investigator on the ESRC-funded 'Hate' Speech and Social Media: Understanding Users, Networks and Information Flows.</p>

B2.2.1. PHEME Personnel List

Partner Name	Staff Name	Position	Permanent / Contract	Involvement	Percentage	Duration (m)	Email	Involvement in other EC project
USFD	Kalina Bontcheva	Senior Research Scientist	Contract	Principal Investigator	45%	36	k.bontcheva@dcs.shef.ac.uk	TrendMiner 10% (until Dec'14)
USFD	Leon Derczynski	Research Associate	Contract	Researcher	20%	24	l.derczynski@dcs.shef.ac.uk	
USFD	New RA	Research Associate	Contract	Researcher	50%	36		
USFD	PhD Student	Student	Contract	3 year Studentship	100%	36		
USFD	Lucy Moffatt	Research Administrator	Contract	Administrator	10%	36	l.moffatt@dcs.shef.ac.uk	Arcomem 40% (until Dec'13), Paths 20% (until Dec'13), AnnoMarket 10% (until May'14), Prowess 15% (until Sept'15)
USFD	Joanne Watson	Head of European Finance	Permanent	Finance	3%	36	joanne.watson@sheffield.ac.uk	
USAAR	Thierry Declerck	Senior Research Scientist	Contract	Principal Investigator	20%	36	declerck@dfki.de	TrendMiner 80% (until Dec. 14)
USAAR	Stephan Busemann	Professor	permanent	Researcher	5%	36	Stephan.Busemann@dfki.de	TrendMiner 5% (until Dec. 14)
USAAR	Vera Demberg	Professor, Head of Group	Contract	Researcher	15%	36		
USARR	Cronna Hahn	Research Administrator	Contract	Administrator	5%	36	c.hahn@eurice.eu	
MOD	Arno Scharl	Professor	Contract	Principal Investigator	10%	36	arno.scharl@modul.ac.at	DecarboNet 15%
MOD	Marta Sabou	Assistant Professor	Contract	Co-PI	5%	36	marta.sabou@modul.ac.at	DecarboNet 5%
MOD	Alexander Hubmann-Haidvogel	Senior Researcher	Contract	WP Leader	40%	26	alexander.hubmann@modul.ac.at	
MOD	New RA	Researcher	Contract	Researcher	80%	36		
MOD	New RA	Researcher	Contract	Researcher	40%	36		
ONTO	Georgi Georgiev	Information Extraction Architect	Permanent	Researcher	50%	36	georgi.georgiev@ontotext.com	AnnoMarket 30% (until May'14)
ONTO	Valentin Zhikov	Researcher	Permanent	Researcher	30%	36	valentin.zhikov@ontotext.com	
ONTO	Petya Osenova	Researcher	Permanent	Researcher	40%	36	petya.osenova@ontotext.com	
ONTO	Laura Tolosi	Lead Scientist	Permanent	Researcher	40%	36	laura.tolosi@ontotext.com	
ATOS	Tomas Pariente	Project Manager	Contract	Team Leader	15%	36	tomas.pariente@atos.net	First: 35%; VPH Share: 25%
ATOS	Ivan Martinez	SW Architect	Contract	Project Responsible	40%	36	ivan.martinez@atos.net	Khresmoi:50%; VPH Share: 10%
ATOS	Miguel Angel Tinte	SW Engineer	Contract	Project Co-Responsible	40%	36	miguel.tinte@atos.net	Khresmoi: 55%
KCL	Robert Stewart	Professor of Psychiatric Epidemiology & Clinical Informatics	Permanent	Principal Investigator	5%	36	robert.stewart@kcl.ac.uk	
KCL	New RA	Research Associate	Contract	Researcher	100%	36		
KCL	PhD student	Student	Contract	studentship	50%	36		
KCL	EU administrator	EU finance administrator	Permanent	administrator	3%	36		
iHub	Rob Baker	Operations Manager	Permanent	Principal Investigator	10%	36	robbaker@ushahidi.com	
iHub	Limo Taboi	Finance Director	Permanent	Administrator	5%	36	limo@ushahidi.com	
iHub	Nathaniel Manning	Director of Business Development	Permanent	Researcher	10%	36	nathaniel@ushahidi.com	
iHub	New Hire	Software Developer	Contract	Developer	100%	36		
SWI	Peter Schibli	Director	Permanent	Principal Investigator	10%	36	peter.schibli@swissinfo.ch	
SWI	Hubert Zumwald	Head IT	Permanent	Researcher	20%	36	hubert.zumwald@swissinfo.ch	
SWI	Christophe Bruttin	Specialist new editorial formats	Permanent	Researcher	40%	36	christophe.bruttin@swissinfo.ch	
SWI	New RA	Research Associate	Permanent	Researcher	40%	36		
SWI	Peter Zschaler	CFO	Permanent	Administrator	10%	36	peter.zschaler@swissinfo.ch	
UWAR	Rob Procter	Professor	Permanent	Principal Investigator	10%	36	rob.procter@warwick.ac.uk	
UWAR	New RA	Research Associate	Contract	Researcher	100%	20		

B 2.3 Consortium as a whole

The Consortium consists of 9 partners from 7 countries. The make-up of the Consortium guarantees that the project's objectives will be met, and its main strength is a firm pan-European (international) partnership, whose synergy will provide far more than each partner working independently. Within the EU, we have six national states represented (Austria, Bulgaria, Germany, Spain, Switzerland, and UK), including one recently joined EU country with a less resourced language (Bulgarian). The seventh country is Kenya, where we have attracted a world-leading SME (iHub) in developing open-source news crowdsourcing and collaborative filtering platforms, as well as providing paid-for web hosted versions and other services.

The consortium includes internationally established players in all key areas of required expertise:

- *large-scale web data collection, storage, and indexing* (ATOS, ONTO);
- *multi-lingual information extraction, temporal processing, semantics* (USFD, USAAR, ONTO);
- *Linked Open Data, reasoning, and use as knowledge sources for NLP* (ONTO, USFD, USAAR);
- *graph-based methods for modelling information diffusion patterns* (MOD, USFD);
- *web-based visualisations for document analytics* (MOD, IHUB, ONTO);
- *large-data analytics, Hadoop, Storm, and related infrastructure* (ATOS, ONTO, USFD);
- *crowdsourcing* (IHUB, MOD, USFD, USAAR)
- and two use case partners that require innovative tools for automatic intelligence gathering across news, social media, and other sources for their verticals in *veracity intelligence in patient care* (KCL) and *digital journalism* (SWI, IHUB).

The main scientific competences are not only provided by some of the leaders in each field, but highlight that the project will be a truly synergistic experience, utilising the best approaches from the whole consortium while promoting cross-fertilisation of ideas.

The involvement of 3 SMEs and 1 big data IT service provider puts PHEME into a strong position to promote and exploit the project results in diverse commercial applications and vertical markets.

Another strong point of the consortium is that most of the partners have collaborated in the past and/or are currently collaborating in EC funded projects.

SME and Corporate Involvement: The consortium includes three SMEs that provide innovative products and solutions in their target markets (text mining and semantic solutions for ONTO; crowdsourcing, information collection, visualization and interactive mapping for IHUB; and multilingual news for SWI); and a large software company working on big data analytics (ATOS).

The chosen partners are **particularly suited to this project**, due to their:

- **Strong knowledge transfer expertise:** ONTO, ATOS, and IHUB have a proven track record of collaborating with universities and publishing innovative research results;
- **Unique background knowledge and market positioning:** ATOS is uniquely positioned as an IT service provider with big data expertise; ONTO has track record both in building scalable text mining solutions (e.g., AstraZeneca, BBC, Press Association, the UK National Archive) and as leaders in Linked Data integration, reasoning, and use for text mining; SWI carry out manual analyses of multilingual social media and its relation to mainstream news on a daily basis. IHUB have a unique platform for crowdsourcing via multiple channels (SMS, Twitter, email, web) and also the SwiftRiver platform, which helps journalists filter and make sense of social media streams.

Their roles in the project have been chosen carefully to match their strengths and company strategies:

- **ONTO** will act as a text mining developer, provider of scalable semantic reasoning solutions (e.g., OWLIM is the most scalable semantic database on the market), Linked Data and reasoning specialists, and leaders of the exploitation activities.

- **ATOS** will be focused on the large data integration infrastructure, with focus on capturing and running large data analytics on social networks, multilingual social media, and authoritative content.
- **IHUB** will enhance the SwiftRiver platform with the new socio-semantic analytics and visualisation tools developed in PHEME, as well as tailor and deploy it in the digital journalism domain. They will chair the media user group, gather their requirements, carry out evaluations, and feed all this expertise and requirements into the scientific workpackages.
- **SWI** will be end-users for the journalism case study. They will engage with the scientific and technology development partners to feed in requirements, provide annotated gold standard data, customise the algorithms to their vertical domain, and evaluate the results.

In general, when setting up the consortium the following requirements were considered for the consortium building strategy, all of which are critical to the success of the project:

- **Competences required:** The best institutions were selected for their expertise in the necessary technological, scientific and business applications areas covered by the project. In particular, partners who show leading competences in interdisciplinary areas have been chosen, both for inter-disciplinarity within technological disciplines (information extraction and attitude detection, knowledge modelling and reasoning, social sciences, visualisation) and within areas outside of the technological disciplines (journalism and healthcare). At the same time, the choice has been optimised to maximise the diversity of competences, providing complementary expertise, while avoiding unnecessary duplication.
- **Roles in the project process:** Partners have been chosen that can successfully carry out all the tasks in the project from the research and development, to implementation of the outcomes in specific real-world situations, as well as the dissemination and exploitation of those outcomes at the transfer period at the end of the project.

The design of the research objectives and the associated work packages has been arranged carefully to ensure maximum efficiency of input from each partner while ensuring a suitable distribution of responsibilities. We have concentrated on covering all aspects required for successful innovation, development and exploitation of the project.

In addition, an important feature of the project is that it is fundamentally multi-disciplinary. The interdisciplinary nature will ensure that the ideas that are usually dealt with in isolated communities can be presented and seeded into other ICT domains.

B2.3.1 Roles in the innovation process

The following roles have been identified for the innovation process:

- Scientific research, guaranteeing the necessary know-how to produce innovative research beyond the state of the art in all the areas relevant to the project.
- Technology developers, guaranteeing all the necessary know-how for linking back the research results into state of the art technologies as well as commercial routes to market for exploitation of the project results.
- Users, providing the necessary requirements and case studies for validating the results and guaranteeing all necessary skills and connections to the market to ensure future exploitation.

The allocation of the project partners to the different roles is reported below:

Partner	USFD	USAAR	MOD	ONTO	ATOS	KCL	IHUB	SWI	UWAR
Scientific research	√	√	√	√					√
Technology developer			√	√	√		√		
End Users						√	√	√	

A key part of the project is the dissemination of both scientific and technological results, both in industrial domains as well as academic domains. All partners will disseminate project results in their specific domains and key partners in each domain will be chosen to manage the dissemination activities. Additional dissemination and knowledge transfer will occur through the project's user focus group and advisory board, as well as through ongoing collaborations in other projects.

European Added Value of the Consortium

There are a number of reasons why PHEME needs to be carried out at a European level.

Firstly, it is widely accepted that SMEs are one of the strong points of the European economy and also that the speed of innovation is such that partnering is often the indicated strategy for companies who wish to compete effectively in their markets. Moreover, the European IT industry is very fragmented with a large number of participants not gaining effective access to sufficiently large markets. Furthermore, they cannot establish sufficient links with potential collaborators who are needed to help develop innovative products and services, because of the complexity of the technologies involved and the significant amounts of added value that must be created. These two European-wide issues, namely the speed of innovation/commercialisation and the overcoming of the fragmented industry base, provide an important justification for this project being carried out at the European level. In addition, the creation of open-source socio-semantic content analytics and visualisation tools will lower their cost of ownership and the barrier to entry, thus making them more affordable to SMEs.

Secondly, the objectives of the project are inherently multi-national, because it aims at developing efficient multi-lingual content analytics tools and services, tailored specifically to interlinking social and traditional media. In order to achieve this goal, input from diverse areas and world-leading expertise in all the areas listed above is needed. In addition, the impact and take-up of the results will be much more immediate and wider ranging if it is developed at a European level. The Consortium is composed of complementary partners from 5 different European countries, covering all required expertise for the project. The set of skills of the partners is both necessary and sufficient for carrying out the proposed work, and for the commercialisation of the project results.

Funding for Beneficiaries from Third Countries

The proposal includes a partner (IHUB) from Kenya, who are essential for the consortium, as they are the unique provider of two key technology platforms for the project:

- The Ushahidi platform for crowdsourcing information using multiple channels, including SMS, email, Twitter and the web.
- The SwiftRiver platform, which is aimed at journalists as an information filtering, curation, and analytics tool over diverse media streams (web, microblogs, email, text messages). It uniquely offers curation-based, shared information filtering through crowdsourcing, which will be used in PHEME as training and evaluation data for the socio-semantic content analytics algorithms.

IHUB already work closely with major news media, such as the Guardian, Al Jazeera, the BBC, and Huffington Post, which will give PHEME immediate access to additional end-users for testing and evaluation.

Focus Groups: Key Stakeholders and their Contribution to the Consortium

The project has a user group with key stakeholders, involved in the digital journalism use case (see Appendix A for letters of support). These will be expanded with new members and finalised before the project start. Below we list organisations who have expressed interest in the project and others with whom partners have already established collaborations and other ongoing projects:

- *Newspapers and news media*, including the Guardian, the BBC World Service, Deutsche Welle and Südwestrundfunk;
- *News providers* (e.g. the Press Association, UK);
- *Open news initiatives*: Open society media programme (www.mappingdigitalmedia.org), the Knight Mozilla Open News program (<http://www.mozillaopennews.org/>).
- The UK Health Protection Agency is supporting the veracity intelligence for patient healthcare use case (WP7), where Dr Tim Brooks will be an HPA end user advisor to the project. From April 2013, HPA has become part of Public Health England and a staff member from their “Knowledge” department will also join.

B 2.4 Resources to be committed

B2.4.1 Budgetary Analysis

The ambitious set of objectives pursued by the project, as well as the relevant amount of resources to be set into motion for their attainment are adequately reflected in the financial plan of PHEME. In estimating the project’s budget, the consortium has taken full advantage of previous experience in RTD projects, particularly, those co-financed by the European Commission within FP6/FP7 and other previous programs. To this respect, budget has been drafted with the contribution from all partners.

Therefore, the adequacy of the financial planning to ensure the integration of resources and the proper development of the Implementation Plan has been secured. The total costs budget for the completion of the PHEME project is estimated at 4,269,938 € and the requested EU contribution, according to the rules established for this call is of 2,916,000 €

The balance of effort between research and technology development (76.8%), case studies with end users (12.2%), dissemination and exploitation (5.6%), and management (5.4%) reflects, on the one hand, the scientific and technological nature of this endeavour but, at the same time, the significant effort devoted to the activities oriented to maximise the outputs of PHEME. The allocation of 5.4% of the effort to dissemination and exploitation activities will secure the transfer of knowledge created by the project, as well as provide clear impacts and plans for future exploitation and usage of PHEME’s outputs. Previous experience has allowed streamlining the efforts required for project management (5.4%), while at the same time ensuring that resources are adequate both in qualitative and quantitative terms.

From the perspective of budget breakdown by partners and type of activity, budget structure reflects the following elements:

- The scientific partners represent 45% of the total budget, deemed as sufficient to secure the critical mass required for achieving PHEME’s ambitious research objectives.
- The participation of industrial/end user partners is a particular strength (55%), allowing the project to pursue two user-driven case studies in key verticals – healthcare and journalism.
- The effort of the Project Coordinator (17%) reflects both their activities in project management as well as their strong involvement in key RTD activities.

In more detail, the effort breakdown by partner and activity is shown below:

Partner	Name	PMs (RTD)	PMs (MGT)	% of PMs	€ Request	% Request
1	USFD	67	12	17%	581,378 €	20%
2	USAAR	50	3	11%	389,400 €	13%
3	MOD	58	1	13%	387,472 €	13%
4	ONTO	56	1	12%	315,648 €	11%
5	ATOS	79	3	18%	303,750 €	10%
6	KCL	40	1	9%	330,239 €	11%
7	iHub	27	1	6%	170,056 €	6%
8	SWI	43	2	10%	268,920 €	9%
9	UWAR	22	1	5%	169,137 €	6%
	Total:	442	25	100%	2,916,000 €	100%

From the point of view of main cost headings, the initial estimated breakdown is the following:

TYPE OF ACTIVITIES	Euros	%
RTD	2,677,761 €	92%
MANAGEMENT	238,239 €	8%
TOTAL	2,916,000 €	100%

The direct costs are further broken down as follows:

TYPE OF ACTIVITIES	Euros	%
PERSONNEL	2,615,047 €	89%
TRAVEL & SUBSISTENCE	192,000 €	6%
OTHER COSTS	147,401 €	5%
TOTAL	2,954,448 €	100%

The above costs were arrived at in the following way.

Direct costs

Personnel costs: Personnel costs are calculated on the basis of person months allocated to the Project by the partners and the corresponding average person-months rates of each partner.

Subcontracting: Within this chapter, the costs of audit certificates have been included. For partners who receive an EC contribution of more than 375,000 € one certificate has been calculated. More specifically, USFD, USAAR and MOD have audit costs included. KCL have also budgeted 24,000 € which is to be incurred under WP7. WP7 will involve integration of temporal and spatial relationships pertaining both to the newly measured internet content within each case study and to the EHR data with which it will be linked and evaluated. Although the necessary data are available in the EHR on time and place (e.g. the person's residence and the timing of healthcare contacts and events), these are cumbersome to extract and analyse. Resources are therefore requested for developing the CRIS system (which provides researcher access to mental health EHR data) so that temporal and spatial relationships can be more robustly archived and readily accessed by researchers while maintaining data security and confidentiality (particularly concerning address-linked data).

Travel and subsistence: These costs cover all participants travel and daily allowance that will be likely needed to assure adequate participation of partners in bilateral and consortium meetings and working sessions, as well as in key industrial and scientific conferences for dissemination purposes. In general, travel costs have been reviewed and minimized as far as practically possible. Most partners have a travel budget of 20,000€ which is for 12 meetings (at 1,000€each) + 8,000€for conferences (5-6 conferences at

1200-1500 euros each). The 12 meetings are 4 per annum, over 3 years (2 consortium meetings, 1 project review, 1 integration/technical meeting). From previous FP7 project experience we have established that dedicated integration meetings are necessary for the technical work to progress and integrate well.

USFD, as the coordinator, has a travel budget of 30,000€ to enable additional dissemination activities for the entire project and the participation of 2 people at project meetings (the coordinator Dr. Bontcheva and the WP3 work package leader). We have budgeted for 20 meetings for the 2 USFD participants and 10,000€ for conferences (2,000€ more than other partners, so the coordinator can represent the project at ICT events and similar meetings, organised by the EC and/or major NLP networking projects).

The travel budget of ATOS (24,000€) and SWI (26,000€) are also slightly higher in comparison to the other partners (20,000€ or less), since ATOS lead the exploitation activities so will need slightly higher budget to present the project at trade fairs and industrial conferences. SWI are leading the case studies so we anticipate they will have more travels for site visits.

The project teams will be led by internationally leading, experienced researchers, as follows: USFD - Dr Kalina Bontcheva and Dr. Trevor Cohn; USAAR – Dr. Thierry Declerck; MOD – Prof. Arno Scharl and Dr Marta Sabou; ONTO – Dr. Georgi Georgiev; ATOS – Tomás Pariente Lobo; KCL BRC – Prof. Robert Stewart; IHUB – Robert Baker; SWI – Dr. Peter Schibli; UWAR – Prof. Rob Procter.

Setting up, testing and making operational the integrated PHEME veracity intelligence framework requires suitable hardware resources, as well as access to cloud computing platform(s). Overall, in the PHEME budget, these resources are relatively small, in proportion to the data sizes and processing power required for mining large-scale media streams in real time. We plan to invest about 1.1% of the total budget for hardware infrastructure and cloud computing costs (estimated at 46,500 Euros). About the 46,500 in hardware and cloud compute time, this is divided as follows. USFD, ATOS, ONTO, MOD, and IHUB request 9,000€ each to cover cloud computing costs, as well as the purchase of a server each for data collection and running experiments locally (needed due to the large volume of social media data to be collected and analysed). Lastly, UWAR has 1,500€ for hardware or cloud computing costs, as required

The overall budget has been carefully constructed so that it is both adequate for the work packages and robust for the full 3-year duration of the project. The budget will be reviewed annually, in the light of progress and as part of the risk management to ensure that all major tasks are successfully completed and the coherence of PHEME is maintained throughout the entire project.

B2.4.2 Project co-financing

The requested EC contribution is complemented by co-financing by each of the partners. The following table shows a breakdown of the grant by activity type:

TYPE OF ACTIVITIES	Budget		EU Contribution	
	Amount	Percentage	Amount	Percentage
RTD	4,031,698 €	94%	2,677,761 €	92%
MANAGEMENT	238,240 €	6%	238,239 €	8%
TOTAL	4,269,938 €	100%	2,916,000 €	100%

The majority of the scientific, industrial and end user partners in PHEME have an established track record in investing in this kind of shared-cost RTD projects in the past and are at present capable of securing a sound financial contribution to PHEME, as the project is well aligned with their medium and long term corporate/institutional objectives.

B3. IMPACT

B3.1 Strategic impact

B3.1.1 Contribution to Expected Impacts as Listed in the Call

As stated in Section 1.1.6 regarding the relevance to FP7 and objectives of the call, **PHEME addresses Objective ICT-2013.4.1 Content analytics and language technologies, a) Cross-media content**

analytics. PHEME aims at investigating and implementing cross-media, multilingual solutions for gathering actionable socio-semantic veracity intelligence. Such tools are needed in the European society, to ensure effective communication and ingestion of information across media, borders and language barriers. PHEME will help by supporting European citizens, businesses and institutions make sense of multilingual, contradictory, and interlinked online content, coupled with models of authority and information propagation in social networks. A detailed discussion on the impact on these different target communities follows in Section 3.1.2.

The FP7 ICT work programme identifies several foci and related impacts for Challenge 4 "Content analytics and language technologies". Our contribution to each of them is as follows.

- **Outcome 1:** *“Strong participation of private-sector players, including SMEs, well above the FP7 ICT average.”*

Four of the PHEME partners are companies, including 3 SMEs. ATOS, ONTO, IHUB are leading technology providers (ATOS in big data analytics; ONTO for semantic-based language technology and Linked Data; and IHUB in open-source crowdsourcing and information filtering platforms). Our use case partners are SWI (a multilingual online news provider). Consequently, the project has been designed from its very inception to have strong participation of industrial partners. Even more importantly, the presence of ATOS and the two SME technology providers will ensure that the new research methods arising from PHEME are tested and evaluated on large data sets, and ultimately delivered as reliable, industrial-strength language technology services. The partners will also integrate them within existing products and platforms (e.g. SwiftRiver, KIM). User requirements capture and evaluation are a particular strength of PHEME, through the two user groups in our two case studies (healthcare and journalism), which bring together over 10 user organisations with diverse needs, working practices, and from diverse application areas. In this way, we will ensure that PHEME methods and technologies are developed in accordance with real user requirements, are made easily customisable to different user applications, and are evaluated with real users, in real use case scenarios.

- **Outcome 2:** *“Technological leadership and increased innovation capacity...”*

The partners in the PHEME consortium are established as scientific or technological leaders in their respective fields: language technology, web science, social science, big data analytics, information visualisation, healthcare, and online media. The inter-disciplinary nature of PHEME will result in cross-fertilization of methods across those disciplines, and bringing new advances in each. For example, through our cross-disciplinary, socio-semantic approach to intelligence gathering from contradictory content, we will open up language technology and big data analytics providers towards recent advances in social computing and semantic technologies. This will ensure that the industry and research partners maintain and further their world-leading position in language technology, web science, big data analytics, and biomedical text mining, to name just a few.

In addition, PHEME will develop strategies to **allow SMEs and other private sector players** (in the first instance those in the consortium and in our user groups) to not only **sell new products and services**, but to also use the innovative content analytics tools in-house to **improve their efficiency, lower their costs**, and **widen** their pan-European **market share** through better utilisation and provision of multilingual content.

- **Outcome 3:** *“...supporting Small and Medium Enterprises (SMEs) developing innovative applications in structured and unstructured digital content ...and, particularly, in the reuse of open data.”*

The content analytics market for software tools, business solutions, and services currently stands at \$850 million, with a wider addressable market through business intelligence, knowledge management, and customer relations analytics (estimated by Gartner at \$10.5 billion globally in 2010). At present, no single provider dominates and over 50% of text analytics vendors are SMEs.

In this context, the impact of PHEME will be significant. Firstly, the project will directly support the participating SMEs and those in our user groups to offer more competitive products and applications in the content analytics market (see Sections 3.1.2 and 3.2.2 for exploitation details). Reuse of extra-linguistic knowledge from Linked Open Data resources and its integration with linguistic and social information, are a particular strength.

Secondly, by open-sourcing most of PHEME's RTD results, we will support other SMEs in adopting and productising the project results. More specifically, further SME providers will be attracted during the project, capitalising specifically on USFD's strong user base for their open-source GATE text analytics toolkit, which includes over 50 SMEs and entrepreneurs worldwide.

Thirdly, the integration of PHEME's novel content analytics and visualisation methods within IHUB's open-source SwiftRiver platform will lower further the adoption barrier for SMEs targeting the journalism, brand monitoring, and customer intelligence vertical domains.

B3.1.2 Impact on the Target User Communities

Currently companies in diverse areas (e.g. business intelligence, market research, campaign and brand reputation management, customer relationship management, enterprise search and knowledge management) are analysing and comparing social media streams and authoritative content, in a labour intensive and expensive manner. For instance, separating the fake images on Twitter from the real eyewitness photos taken during Hurricane Sandy or during the London riots in 2011 was essential for journalists, but very laborious to accomplish. In more detail, PHEME will target the following areas:

Digital Journalism and Semantic Publishing: The journalist profession has been undergoing a radical change in recent years, in response to the emergence of user-generated content and social networks: "I have seen the future, and it is mutual," says Alan Rusbridger, editor of the Guardian. This has given rise to the idea "networked journalism enables an increasingly wide range of different viewpoints, languages, cultures, values and goals to be encountered" (Beckett & Mansell, 2008).

In order to make the best use of social media and remain competitive, news media, such as SWI, BBC, and other members of our user group, are now looking to enrich their primary product with semi-structured data and also to adopt advanced technologies that help in the gathering and verification of news from UGC.

Consequently, PHEME will have an economic impact on digital journalism and semantic publishing, firstly by lowering the cost of monitoring user-generated content and social networks during news production and secondly by enabling semi-automatic press clipping services, tailored to individual clients. By basing this use case on Ushahidi's news crowdsourcing platform and their SwiftRiver information filtering platform, the impact of PHEME will also go beyond the user organisations directly involved in the project.

Healthcare: South London and Maudsley NHS Foundation Trust (SLAM) is the largest health-care provider in Europe, serving a population of 1.1 million. SLAM's Case Register Information System (CRIS) allows searches to be made of the 11 million textual patient notes and letters to physicians, and extracts anonymised data for secondary analysis. In 2011 SLAM assessed that 80% of the data they require for research and statistical purposes is hidden into these textual records. KCL and USFD have an ongoing collaboration with SLAM, around text mining and mental health research. The technology developed in PHEME will enable them to correlate the issues patients discuss with their physicians (as recorded in CRIS) against medical rumours and information broadcast in the press around the same time. The PHEME methods for contradiction detection will enable also automatic detection of patient misconceptions.

PHEME is also supported by the UK Health Protection Agency (HPA) and will be of relevance to Public Health England (which will absorb the HPA in April 2013). They are interested in PHEME's rumour intelligence technology, to enable them to monitor patient forums and social networks and identify new rumours and scares, so appropriate counter-balancing actions and policies can be undertaken.

Marketing and public relations: Managers are constantly looking for innovative ways of capturing individual and public viewpoints to guide their strategic and operative decisions. PHEME will develop methods for these managers to "listen" to their customers on both the micro and the macro level by gathering and analysing rumours being spread through user-generated content across social media platforms. As demonstrated by Toyota's unnecessary recall, arising from the rumour of Prius has breaking problems, identifying, monitoring, and reacting to rumours in a timely fashion is absolutely essential. PHEME will demonstrate how automated content analytics and the automated tracking of rumours can provide valuable marketing intelligence - e.g. by investigating how brand image and perceived product quality can be inferred from social media, including rapid feedback on the origin and drivers of

perceptions and observable misconceptions. ONTO and MOD already have clients in this vertical, who will be the first impact targets.

Search and knowledge management: USFD's and ONTO's existing text mining tools are already used in electronic archives and digital libraries to enable more powerful, semantic searchers, e.g. for person names and locations. For instance, USFD and ONTO completed recently a project with the UK National Archives, which developed a bespoke 'intelligent discovery tool' to improve searches of archived UK Government websites (around 7TB of data). Secondly, USFD is currently working with the British Library on customising and integrating LOD-based information extraction tools into their Envia information discovery tool (scheduled to go public in 2013). All these projects involved search over trusted, internally held content. There is now strong market interest in linking such content to relevant social media and social networks, which is the focus of PHEME. ONTO and USFD will seek to attract further significant industrial interest and generate exploitation opportunities, firstly by presenting project results to existing industrial contacts (e.g. AltaPlana, Innovantage) and secondly, by targeting new ones through webcasts, YouTube presentations, and participation in relevant industry-oriented forums (e.g. Intelligent Content, Text Analytics, KDNuggets).

Society and Citizens: In addition to its high commercial relevance, PHEME will benefit society and citizens by enabling healthcare professionals, politicians and government representatives to monitor social media and respond to misinformation and rumours. PHEME will show how phemes diffuse in social networks, how they are understood, remembered, and propagated. By uncovering rumours and authoritative sources, it will engage stakeholders and reveal hidden knowledge to decision makers.

B3.1.3 European Dimension

Due to its focus on multilinguality, PHEME is best undertaken on pan-European level. In addition, no single European country can marshal the resources necessary to undertake the project. A single country project would suffer the risk of being driven by use case practice of that country alone and the linguistic properties of the local language. PHEME requires a high level of technical and scientific expertise in a diverse range of areas: knowledge-based fact extraction, reasoning and contradiction detection, social network analysis, information diffusion, big data analytics, etc. No single country leads in all these areas. In addition, it brings together major contributors and developers of a variety of open source software required for the project.

The PHEME project is eminently European, and gains by bringing the deep specialisations of many European research leaders together from many countries together into a shared unit. A project such as this needs this kind of critical mass - of research, domains and data - to be feasible. In addition to drawing together research communities, the project creates technology which not only permits companies to adopt big data content analytics (e.g. monitor social media for drug reactions), but also makes these advanced tools available to SMEs, non-profit organisations, journalists, bloggers and citizens to see and explore social networks, media, and authoritative sources, without being inhibited by language and cost. While not covering all languages, as is the limit of a medium sized project, the algorithms and frameworks are at every stage built with extension in mind; and the inclusion of the Kenyan partner looks outwards even beyond this European scope.

B3.1.4 Impact through Open Source and Standardisation

Europe is a leader in Open Source Software and its importance as a vehicle for innovation and competitiveness, especially for SMEs, is well established (see <http://www.flossimpact.eu>). Many PHEME partners are actively coordinating or contributing to the development of open source products, many of which are already widely used in the world. The research and development activities in PHEME will be reflected into these open source products, increasing their available features, reliability, scalability, and performance, and hence their attractiveness and suitability for wider deployment. Members of the PHEME consortium active in the **open source** area are:

- USFD: the GATE toolkit contains the most widely used information extraction components in existence.
- MOD: they have developed the open-source ThreadViz visualisation tools.

- ONTO: has incorporated open source software development as part of its technology development strategy. One of Ontotext's flagship products, OWLIM, is open source software, implemented on top of the Sesame RDF repository. They have also contributed to development of Sesame and GATE.
- IHUB: the Ushahidi multi-channel crowdsourcing platform and the SwiftRiver collaborative information filtering platform are unique world-wide.

USAAR and USFD are involved in **ISO TC37/SC4** on language resource management, and USFD's GATE implementation forms part of the reference implementation of the **standards**. USAAR and USFD will feed their expertise from PHEME into the decisions and pre-normative documents of this ISO committee.

ONTO are also actively involved in W3C standardisation activities. The experience from using RDF(S), OWL, and Linked Data in PHEME will be fed back into the relevant standardisation groups.

The project will also make extensive use and be compliant with relevant standards in language technology, Linked Data, RDF, and OWL. For further details see Section 2.1.3.

B3.1.5 Cooperation with and Relationship to European and National Projects

Where applicable, PHEME consortium members involved are indicated.

TRENDMINER (www.trendminer-project.eu/) (USFD, ONTO): work on identifying entities, opinions, and trends in social media streams. PHEME will use the results of this project as initial building blocks for the rumour, misinformation, and contradiction extraction algorithms.

X-LIKE (www.xlike.org/): focused on cross-lingual knowledge extraction from news sources and social media, with applications to publishing and brand media monitoring. PHEME differs in its focus on contradictory and misleading information, social networks, modelling of information diffusion, and the discovery of implicit information exchange networks across media.

EXCITEMENT (www.excitement-project.eu/): research on textual entailment for customer interactions, an open-source textual inference platform, and industrial CRM applications. PHEME will benefit from any corpora, algorithms for textual entailment, and the text inference platform. We will seek active cooperation and knowledge exchange with EXCITEMENT. PHEME's contribution will be in new methods for the detection of contradictions, rumours, and unverified claims, as well as in the integration of social networks.

OPENNER: will create sentiment lexicons and open-source methods for named entity recognition with Lined Data, with applications to tourism. PHEME will reuse the OpenNER results, to create sentiment and entity features for the learning algorithms for rumour, contradiction, and claim detection.

LIMOSINE (limosine-project.eu/): focused on web opinion mining and social media, semantic personalisation, and relation extraction. These are all necessary semantic features for the PHEME learning algorithms and, therefore, active collaboration and exchange of tools will be pursued.

EUROSENTIMENT (www.eurosentiment.eu/): creating a multilingual LR pool for sentiment analysis, connected to EmotionML. Similar to OpenNER and Limosine, the results of EuroSentiment will be directly reusable in PHEME, to provide sentiment features for PHEME learning algorithms.

META-NET (www.meta-net.eu/) (USFD): META-NET aims at an open infrastructure of resources, especially sharing of large repositories of processed and annotated language data. META-NET and its related projects, will serve as a source of multilingual resources and tools for PHEME, as well as a dissemination network.

LarKC (www.larkc.eu/) (ONTO, USFD) developed a platform for large scale integrated reasoning. ONTO's Linked Life Data approach to reasoning will be extended in PHEME, where we will apply it to other types of knowledge, for the detection of rumours, contradictions, and misinformation, which were not dealt with in LarCK.

FIRST (<http://project-first.eu/>) (ATOS): Another complementary project, concentrating on merging of factual information that can guide financial decision making. PHEME's focus is on veracity of automatically discovered rumours. A close cooperation on the use of Hadoop and Storm for big data analytics is envisaged. ATOS are developing both infrastructures, which will ensure reuse and avoidance of duplicated effort.

RENDER (render-project.eu/) (ONTO): This ongoing project provides a conceptual framework and technological infrastructure for managing and exploiting information diversity on the Web. ONTO will bring their expertise on highly scalable data management, enriched with machine-understandable descriptions and links referring to the Linked Open Data Cloud. PHEME's focus on Social Semantic Intelligence, across media and languages, is therefore complementary to RENDER. In PHEME, we will go beyond modelling facts and opinions (RENDER), towards modelling contradictions and misinformation, the temporal validity of facts, the reliability of the information sources, and the implicit information exchange networks. Additionally, PHEME focuses on more concrete and socially crucial use-cases: digital journalism and patient care.

CLARIN (Common Language Resources and Technology Infrastructure) (www.clarin.eu) (ONTO, USFD): PHEME fits into the vision of the pan-European initiative CLARIN to enable lower thresholds to multicultural and multilingual content. PHEME will serve as a use case that shows how the language and semantic technologies would combine to process, analyse and visualize social digital content with respect to rumour/opinion distinction in various domains.

GATE Cloud Exploratory (bit.ly/gate-cloud-jisc-project) (USFD): A UK-funded project, which delivered a prototype platform for deploying GATE-based text analytics applications on the Amazon EC2 cloud infrastructure. PHEME will benefit from USFD's expertise, for the design and implementation of the big data content analytics.

CUBIST (www.cubist-project.eu/)(ONTO): This is an ongoing project that aims at Combining and Uniting Business Intelligence and Semantic Technologies with a special focus on unstructured data mining. Being central to the project goals, the semantic technology supports a persistent layer – a semantic Data Warehouse. PHEME might adapt the proposed semantic model and make use of the visual analytics in accordance with its own aims.

DICODE (dicode-project.eu/): focused primarily on large-scale data aggregation and mining, through MapReduce, Hadoop, and Mahout. PHEME will reuse expertise and results with respect to parallelisation, scalability and relevant algorithms. Our approach is different in its emphasis on mining unstructured social media content for rumours and misinformation, based on extra-linguistic knowledge.

ROBUST (robust-project.eu/): A complementary project working on describing and modelling of user communities. PHEME's focus is on social network analysis and discovery of implicit information diffusion networks will be of relevance to ROBUST, as well as benefit from tools developed in ROBUST.

COCKPIT (www.cockpit-project.eu/): focused on citizen collaboration and co-creation in public service delivery, including some citizen opinion mining research, where cooperation and reuse of results will be pursued. PHEME's focus however is not opinions, but modelling rumours, contradictions, and misinformation.

B3.2 Plan for the use and dissemination of foreground

B3.2.1 PHEME Dissemination Plans

For maximum impact, PHEME will adopt a multi-channel dissemination approach.

There will be a full programme of scientific papers and presentations at technical and scientific conferences. These will be aimed at sharing the results of PHEME with the scientific community, particularly the European scientific community, to encourage their incorporation into the work of other scientists and technologists. The conferences to be targeted are detailed in WP9.

Complementing this, there will be a programme of papers and articles in the information technology and general business literature, as well as presentations at IT and business seminars and conferences (see WP9 for a detailed list). We will also target key stakeholder groups for the two case studies. Again, this will be directed towards European community and associated countries.

The project will also set up and maintain a **public project web site**. The project website will function both as a project dissemination tool and as a server running some of the software produced by the project. The website will also provide examples, API documentation and video tutorials explaining how the

software tools can be used by end users, integrated in applications, or used as component services by SMEs.

We will also harness **social media as a dissemination channel** and will actively promote the project and its results to the large online communities in the areas of language processing³, text analytics, semantic web, business intelligence, new media, and other relevant groups on LinkedIn, Facebook, and blogs. The project will also have a Twitter dissemination account.

Within the limitations imposed by appropriate protection of intellectual property, **all scientific results of this project will be made available to the research community**. In addition, a number of content analytics tools and resources from PHEME will be made available as **open-source** to facilitate take-up. Another vital part of the dissemination and exploitation activities will be the **support of a community of early adopters** of the open-source technology, through a mailing list, code examples, training workshops, etc.

Training materials (e.g. state-of-the-art reports) resulting from the project will be gathered and used by the academic partners as part of their course materials at postgraduate level, as well as, used by all partners to provide courses to companies. In addition, the research activities in the project will involve PhD students for some of the participants (e.g. USFD, USAAR).

PHEME will also produce **results relevant to other ongoing projects**, with which **it will cooperate**. In addition, project partners have already established links to complementary projects in all key RTD areas:

- **Language Technologies:** META.NET, ARCOMEM, TRENDMINER, ANNOMARKET, OPENER;
- **Social Media Analysis:** EXCITEMENT, X-LIKE, OpeNER, ROBUST
- **Linked Data, Reasoning, and Semantics:** LARCK, LOD2, ROBUST
- **Use cases:** EU-PATI, SELCOH, ANNOMARKET, VISTA-TV

Both the industrial and the academic partners will play active roles in dissemination. The industrial partners will form a key part of the dissemination activity, as they will provide **outreach towards businesses in diverse sectors**. Both individual and joint dissemination and exploitation actions will be undertaken. Within the academic partners, consultancy to business and contractual research and development will be a significant part of dissemination, besides the normal channels of publications, conference papers, co-organised workshops and courses given, and the web.

A range of other instruments will also be employed for dissemination including press and media, further participation in appropriate networks of excellence and some training activities aimed at both the academic and industrial sectors (e.g. delivery of lectures at summer schools, participation in industry-oriented dedicated workshops and events). More specifically, we have planned to **collaborate with the META network** as part of joint dissemination events, e.g. info days for researchers and businesses, workshops, conferences, etc.

B3.2.2 PHEME Exploitation Intentions

A 2011 survey of the text analytics market (Grimes, 2011) put its size at \$850 million, with 25% growth potential, and wider applications in business intelligence, knowledge management, customer experience management, market research, etc. The study also showed that social media are the most important kind of unstructured information for which automatic solutions are needed; followed by news articles, email, and online forums. In addition, 49% of respondents needed multilingual tools (English and German are in the top 3 most needed languages).

PHEME will not only develop **novel algorithms for mining socio-semantic veracity intelligence**, but will also deliver them as **high-throughput, scalable tools and services** for processing textual and social media streams, as well as **integrate and test** these in working use case prototypes. The PHEME partners have been selected carefully to enable a quick transformation of project results into working prototypes and to investigate the business opportunities in the target business sectors. The novelty of the project and the expected results will allow the partners to showcase the project achievements and ease the market barriers for this new technology, opening up more commercial opportunities.

³ For instance, the NLP LinkedIn group has more than 2,300 members; the Semantic Web one – close to 6,000.

The project will allow the participating industrial and end user partners to **reinforce their position** in the market and enrich their content analytics product and service portfolio, and also give opportunity to explore new areas and collaborations. The PHEME cross-media and multilingual veracity intelligence methods and tools will allow for **immediate exploitation** of the project results and the know-how acquired from the RTD activities. While each partner will develop their own exploitation plan, a common plan will be defined during the course of the project, in order to build on the jointly acquired know-how and allow for strong commercial and marketing exploitation based on the strengths of each individual partner.

Research organisations will also pursue commercialisation of their research through consulting services, inclusion as modules in commercial products or via spin-off companies.

In summary, the **concrete PHEME exploitable results** will include both tangible and intangible assets, which are summarised in the following table:

Tangible Assets include:	<ul style="list-style-type: none"> ✓ Multilingual methods for spatio-temporal grounding and user and content geolocation ✓ Methods for contextual interpretation and cross-media linking ✓ Methods for computing veracity, across media and languages ✓ The PHEME visual analytics dashboard ✓ The large-scale content and semantics storage tools ✓ The integrated veracity intelligence framework ✓ The PHEME use case prototypes in digital journalism and patient care ✓ The annotated corpora produced by the two use cases
Intangible Assets include:	<ul style="list-style-type: none"> ✓ The knowledge for multilingual social web analysis, mining and visualisation ✓ The expertise from customising the PHEME tools to new under-resourced languages and to new application domains

Furthermore, different partners will exploit these and any other results through different avenues:

- Actual commercialisation of the PHEME tangible assets.
- Utilisation in the partners' own solutions and commercial portfolios.
- Earning royalties from the incorporation of the research and development results into commercial tools and applications.
- Offering a range of consultancy and training services to other interested parties.

Partner Exploitation Intentions

ONTO's business model combines the development of products (including some open source versions) with the provision of research, consultancy and development services. Many commercial projects combine all four elements. For **Ontotext** PHEME will bring the unique opportunity to strengthen its position in the semantic technologies and knowledge-driven text analytics market, with development and adoption of cutting edge social intelligence methods.

More precisely, PHEME fits into Ontotext's strategy in the following ways:

- Experiment with applying its technology (in this case, the KIM Semantic Annotation Platform, <http://www.ontotext.com/kim>; and Publishing tools) to a new challenging problem. Cross-media analytics is a typical case of business intelligence. It is also related to other projects where KIM and Publishing tools are used for multimedia content analysis and search (e.g. ANNOMARKET, TNA).
- Allow ONTO to go beyond the semantic and world-knowledge infrastructure modules. Namely - towards the detection, analysis and extraction of rumours in a cross-media context. ONTO's

existing LOD resources (e.g. *Linked Life Data*) would be applied for different socially aware domains and across languages. For example, the entity extraction tool LUPedia (<http://lupedia.ontotext.com>) and the linked data concept store FactForge (<http://factforge.net>) will be used for a priori knowledge.

- The existing methods and techniques would be developed further with respect to the management of large amounts of rumours and phemes, in the direction of source's trustworthiness and impact.
- The task of discovering conflicting information in socio-media contexts would test and enrich the OWLIM reasoning platform, developed and maintained at Ontotext.

ATOS is interested in integrating the project results in its current commercial products and services. The operating assets of R & D projects move through the Trade Committee of ATOS to the sales network spread across different market sectors. The node of R & D in Atos (Atos Research & Innovation) has a special interest in applying innovation to our language technologies portfolio and to increase our knowledge of handling big data in the scope of language technologies. We plan to improve our big data prototypes by applying a sound architecture that potentially will be presented internally to our own customers and especially the company account managers, who are in charge of the commercial exploitation and engagement with customers. Our aim is to reach a wide internal dissemination of the results of the project to our commercial staff in order to promote and demonstrate relevant PHEME tools to different customers with potential interest in language technologies. In addition to this commercial work, promotional material (brochures, business presentations, etc.) will be disseminated to customers or to relevant workshops and conferences.

IHUB will exploit project results through their integration within IHUB's open-source SwiftRiver platform. The digital journalism use case in PHEME will lend critical insight to SwiftRiver's strengths in manual and algorithmic analysis and verification of crowdsourced content, as well as opportunities to explore and expand other business applications.

iHub's business model is centred on providing consulting and development services on top of our existing portfolio of free and open source software, including offering paid-for hosting of the SwiftRiver platform. In 2013 IHUB is launching a freemium model that would additionally incorporate "VIP" service models for paid subscribers.

In more detail, PHEME plays an important role in Ushahidi's SwiftRiver development strategy through:

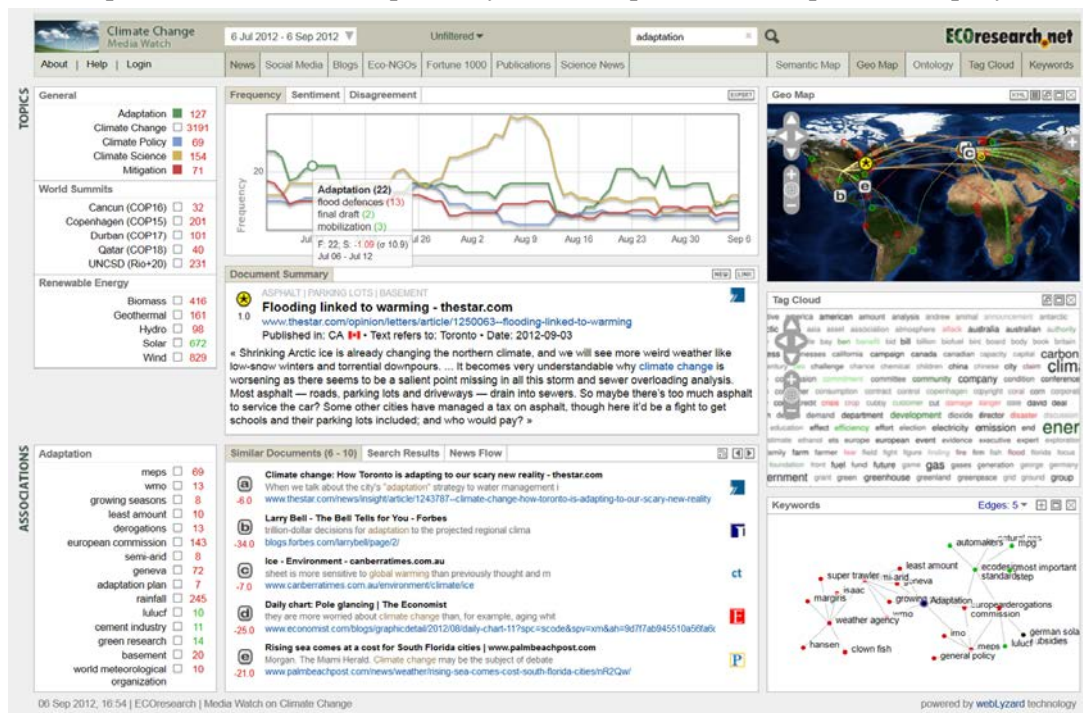
- Integration of new language processing-based functionality to create actionable knowledge;
- New customization and integration of text mining and information visualisation tools, to create a domain-specific bespoke customised application aimed at journalists and media outlets;
- Utilizing SwiftRiver's plugin architecture, to extend the platform and create a series of open source plugins for broader use by journalists and the larger Ushahidi user community;
- Assist Ushahidi and its community of thousands of global users to gain new insights into the unique needs and goals of journalists harnessing social and mobile data gathering methodologies. Ushahidi publishes logistic toolkits along with its software to help the community understand developing strategies around the software itself.

SWI (swissinfo.ch) is aiming to keep its pioneer status among online news platforms. The project would contribute to the editorial team's capabilities in tracking specific themes and the status of the discussion on social media and news platforms, helping to find relevant news/information (track rumours), create something like a relevance aggregator to enhance accuracy and relevant contacts and figures to produce quality content, contribute to the discussion where it happens or getting the main arguments on the issue.

With multilingual aggregations (ontology mapping) news/information could provide a better understanding what is discussed beyond language frontiers and strengthen swissinfo's position as a multilingual provider. Also information diffusion is critical, to remain in a competitive position among traditional and social media.

In addition, PHEME will help journalists to: (i) gain further insight into using social media as a working instrument: communication with their audience and user-generated content; (ii) find relevant and interesting stories, competences and sources for their articles; (iii) offer more added value to the readers

USFD has proven track record in developing and promoting open-source content analytics and language technologies, as well as providing training and consultancy services around these. Over the past 15 years, USFD has developed the world-leading open-source GATE NLP toolkit (gate.ac.uk), which has thousands of users at hundreds of sites. USFD provides annual training courses and certification in its technology. A number of new tools for content analytics will be developed in PHEME and released as plugins within the GATE platform. This will ensure automatically that PHEME results are compliant with relevant text and data standards, such as ISO 24611 and 24615, TEI, OASIS UIMA, RDF, and OWL. USFD also plans to pursue impact beyond the bounds of this project, via the Sheffield University Knowledge Transfer Partnerships (KTP) account and EpiGenesys (the Computer Science spin-out company).



MOD will actively pursue exploitation of the research and technological outputs of the PHEME project through their spin-out company. **webLyzard** technology builds bespoke web platforms for analysing online media. It currently operates Web intelligence platforms for a range of international and national clients including the NOAA Climate Program Office of the U.S. Department of Commerce, the National Cancer Institute of the U.S. Department of Health and Human Services, Hutchison 3G Austria, and the Austrian Federal Chamber of Commerce. To measure and assess the impact of public relations and marketing campaigns, the webLyzard platform captures, aggregates and annotates large archives of Web content from multiple stakeholder groups. The ability to track rumours and misconceptions will significantly improve the competitive position of webLyzard, as the content diffusion metrics to be developed within PHEME will go beyond a mere characterization of language structures and provide actionable intelligence and decision support for guiding marketing and public outreach campaigns.

B3.2.3 Management of Knowledge

The consortium partners are ready to bring in their individual expertise and knowledge to make the PHEME project a success. Knowledge created during the project will be distributed within the consortium - and where not subject to confidentiality, beyond the consortium - to enable a targeted and coordinated development towards the project goals. This also requires active knowledge exchange between the project partners and the target user communities, via user group and participation in scientific and industry-oriented events and workshops. In addition to creating an open, creative and flexible project environment, which is built on mutual trust and respect, knowledge exchange will be fostered by meetings, teleconferences, mutual technology demonstrations and the publication of information on the project web page.

As an entry point to project related knowledge a comprehensive project web site will be created, which will also house a project-internal area for trusted exchange of documents and resources. Multiple access protection levels will be set to allow appropriate access to project partners, user group members, and European Commission officers. All intermediate results (i.e. milestones and WP progress reports) will be documented in this area. All meeting preparation, administrative and technical management, discussion groups and deliverable drafts will also be stored and accessed through the restricted area of the web site.

More formal issues of Management of Knowledge will be addressed in the Consortium Agreement, which will be signed before the start of the project.

B3.2.4 Management of Intellectual Property

It is expected that substantial assets in terms of knowledge and technology will be created during the PHEME project. The effective management of knowledge, its dissemination and transfer is a key need for carrying out project related issues properly.

The PHEME consortium is aware that a careful and tailored management of intellectual property rights is needed to meet the requirements and interests of the different types of partners in the project, as well as of the different types of assets created in the project. A well-designed IPR management will contribute to the smooth functioning of the consortium and to the commitment of individual project partners to the project.

For the design of this IPR strategy the consortium will follow a systematic approach. IPR issues within the project will be structured along **three dimensions**: (i) the **type of the asset** considered (i.e. content or technology); (ii) the **point of creation** of the asset (i.e. pre-existing assets vs. assets created during the project); and (iii) the type of **intended use** (i.e. use by consortium partners during the project or after the project, commercial use outside the project, non-commercial use outside the project).

For developing an IPR management strategy, first a general approach will be agreed upon for each of these dimensions. This can then be stepwise refined for groups of assets or individual assets. The definition of an elaborated IPR management strategy will be the subject of the consortium agreement that will be agreed upon and signed by the consortium partners before the beginning of the project. The Consortium agreement will encompass potential issues concerning internal organisation and management of the consortium, the IPR agreement for the pre-existing rights as well as those generated by the work of the consortium, processes for the settlement of internal disputes and commercial exploitation of results.

This section already summarises some important base lines of the IPR management strategy within the PHEME project that are especially critical.

Non-Commercial Use after the Project of Technology Created during the Project: Regarding the core part of the content analytics technology, i.e. algorithms created in WP2, WP3, and WP4, it is our goal to keep their future non-commercial use free of charge. Different licensing models such as LGPL, dual licensing, and freemium will be evaluated during the project to identify the most adequate one(s) for fostering wide use and further extension of the different PHEME tools and components. In some cases, the licensing of pre-existing know-how may influence the choice of possible licensing models.

Commercial Use after the Project of Technology Created during the Project: The option of commercial exploitation of technologies developed within European projects and the creation of new services and products on top of such technologies is one of the major reasons for larger companies and SMEs to join European projects. Thus for some of the technology, especially the ones created as part of the use cases and applications, it might be decided to choose other IPR approaches rather than distributing the source code free of charge to enable systematic commercial exploitation of the technologies by the companies. Details again will be regulated as part of the consortium agreement.

Non-Commercial Use of Content Created during the Project: The main type of content that the project deals with is social media and other Web content, i.e. content not created within the project. Here the existing and evolving IPR rules for the web apply. Much of the content that is created in the project results from content aggregation, de-duplication, extraction and network analysis. Access to this type of content is envisaged via web services with well-defined APIs, which will provide other companies and researchers with access and reuse, under well-defined licensing terms.

Use of Pre-Existing Technology: Pre-existing know-how will naturally be contributed by each partner in order to enable the success of the PHEME project. In principle, the IPRs for pre-existing technology will stay with the original owners. However, the use of these technologies will be free of charge for the consortium partners during the project (for purposes of the project). Further regulations for the use of these technologies will be defined in the consortium agreement.

B3.2.5 Open Source Licensing

The PHEME consortium includes leaders in the provision of open source software for text mining (GATE by USFD), crowdsourcing (iHub), information visualisation (MOD) and curated information management of media streams (SwiftRiver). Most of the core outputs of the project will be open source under a suitable license, such as LGPL (the Lesser GNU Public Licence).

Open source alone is not enough -- the software needs to be supported, maintained and promoted, and the communities of developers nurtured and organised. PHEME partners (USFD, IHUB, ONTO) have a proven track record in this area over the past 15 years and will continue to perform these functions during the project lifetime as part of the exploitation and dissemination activities, and beyond..

B4. ETHICAL ISSUES

The project will involve data sets of humans by using medical data of patients. A large part of the used information is freely available on the Internet already and for its collection privacy issues had been taken into account. In particular, content from publicly accessible patient forums and websites (e.g. PatientsLikeMe) may contain medical images or descriptions of cases. The scientific medical literature is another example where we will reuse data that had been anonymised and for which ethics approval had been obtained ahead of time.

For KCL's patient care use case KCL researchers will also trial the PHEME veracity intelligence tools on the BRC CRIS case register (Stewart *et al*, 2009). CRIS allows searches to be made of the fully electronic patient records system, which extracts anonymised data for secondary analysis including free text fields. The data resource contains full anonymised clinical records on over 150,000 service users, over 30,000 of whom are receiving active case management. CRIS received research ethics approval as an anonymised dataset for secondary analyses by Oxford REC C in September 2008. Even though anonymised, CRIS will only be accessible to KCL researchers, over secure connections. No patients will be involved in the use case, since it is focused entirely on assisting clinicians and other medical practitioners and regulatory bodies.

Using strict procedures of hospital ethics committee in our case from KCL and SLAM assures that all national and international ethical guidelines are strictly followed.

Other privacy issues may arise due to the project's use of social media streams and other Web content. The PHEME tools will perform statistical analysis and produce aggregated veracity intelligence from the collected content, which will result in the data being anonymised. Even though PHEME develops tools for automatic content geolocation and learning the users' likely home location (e.g. from users' public posts on Twitter and Facebook), the obtained information will be used only internally, as input to the veracity intelligence algorithms. No automatically derived data on user behaviour or location will be made public, neither will users be tracked or observed through sensors, cameras, mobile phones, or other technological devices. We will not collect private usage data and other non-public content, but nevertheless, should the need arise, there are tools and expertise in the consortium (USFD) in automatic methods for content anonymisation.

PHEME includes a partner from outside the European Union and Associated states (IHUB is based in KEN). However, they are involved purely in software development, involving open-source tools and publicly available content and data from the web. IHUB will have no access to any copyright or confidential material.

Finally, none of the PHEME case studies will make public any data that raise ethical or privacy issues.

Ethical issues and data protection

The proposed work plan of the consortium involves data collection of humans and the processing of personal data. No additional data will be acquired in the course of the project. All the data will be fully anonymised. Data of children can appear.

All work on data collection of humans will be conducted under the rules and legislation in place within the respective countries of the partners, which are based on:

- the Declaration of Helsinki (informed consent for participation of human subjects in medical and scientific research, 2004) and the IHC guideline for Good Clinical Practice (1996),
- European Directive 2001/20/EC (April 4, 2001) on Good Clinical Practice for clinical trials,
- Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 (amended 2003) on the protection of individuals with regard to the processing of personal data and on the free movement of such data;
- Regulation (CE) No 45/2001 of the European Parliament and of the Council of 18 December 2001, on the protection of individuals with regard to the processing of personal data by the institutions and bodies of the Community and on the free movement of such data.

The Opinions of the European Group on Ethics in Science and New technologies (EGE) (especially Opinion Nr.13 30/07/1999 - Ethical issues of healthcare in the information society) will also be taken into account.

Ethical committees

All protocols will be submitted for approval to local ethical committees or institutional review boards before the start of the project. These operate in accordance with international ethical guidelines and the national laws on research and protection of the human rights of subjects and privacy.

APPENDIX A: LETTERS OF SUPPORT

The University of Sheffield
Department of Computer Science
Dr. Kalina Bontcheva
Regent Court, 211 Portobello
Sheffield S1 4DP
United Kingdom

20 Dec 2012

Dear PHEME Coordinator,

On behalf of my organisation, Open Society Foundations Media Program, I am writing to express my strong support for the PHEME proposal on automatic extraction and visualization of socio-semantic intelligence, across multiple languages, modalities, and content types (social media, online news, structured data, and other authoritative sources).

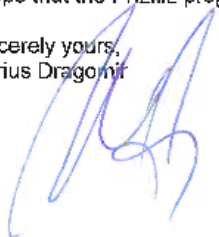
The program supports, among other things, research as the basis for subsequent activities in policymaking, advocacy, and training.

Journalists and online media increasingly monitor and integrate information from social media and user-generated content, alongside the more traditional and authoritative sources, such as research agencies, statistical offices, governmental organizations and others in order to produce complete and up-to-date articles for their audiences. The technologies and methods to be developed in PHEME, namely those for automatic detection and visualisation of rumours, misinformation, authority, and information diffusion, would be of major help to gain timely, strategic knowledge and intelligence from conflicting information streams.

The Media Program is happy to support PHEME through participation in the user group and advice on the user needs and requirements of our industry.

I hope that the PHEME proposal is successful, so we can initiate this exciting work.

Sincerely yours,
Marius Drăgoi





SWR 76522 Baden-Baden

Südwestrundfunk
Anstalt des öffentlichen Rechts

The University of Sheffield

Department of Computer Science
Dr. Kalina Bontcheva
Regent Court, 211 Portobello
Sheffield S1 4DP
United Kingdom

Dr. Robert Fischer
Information, Documentation and Archives
Head of Dept. Digital Archive Systems

Hans-Bredow-Straße
78530 Baden-Baden
Telefon +49 7221-929-24124
Telefax +49-7221-929-24104
robert.fischer@swr.de
www.swr.de

11. Januar 2013

Dear PHEME Coordinator,

On behalf of my department, Südwestrundfunk (SWR) Digital Archive Systems, I am writing to express my strong support for the PHEME proposal on automatic extraction and visualization of socio-semantic intelligence, across multiple languages, modalities, and content types (social media, online news, structured data, and other authoritative sources).

SWR Archives support cross-media journalists for their TV-, Radio-, and Online publications, providing real-time access to all varieties of audio, audiovisual and web-based resources.

Journalists increasingly monitor and integrate information from social media and user-generated content, alongside the more traditional and authoritative sources, such as the Television-, Radio-, and Press-Databases provided by SWR Archives. The technologies and methods to be developed in PHEME, namely those for automatic detection and visualisation of rumours, misinformation, authority, and information diffusion in the web 2.0 sphere, would be of major help to gain timely, strategic knowledge and intelligence from conflicting information streams.

The SWR Digital Archive Systems Department is delighted to support PHEME through participation in the user group and advice on the user needs and requirements of the media environment.

We hope that the PHEME proposal is successful, so we can evaluate the outcoming results and compare them with our existing webarchive infrastructure.

Sincerely yours,



the guardian

THE UNIVERSITY OF SHEFFIELD
DEPARTMENT OF COMPUTER SCIENCE
DR.KALINA BONTCHEVA
REGENT COURT, 211 PORTOBELLO
SHEFFIELD S1 4DP
UNITED KINGDOM

KINGS PLACE, 90 YORK WAY, LONDON N1 9GU
TELEPHONE 020-3353 2000
GUARDIAN.CO.UK

JANUARY 14 2013

DEAR PHEME COORDINATOR

ON BEHALF OF MY ORGANISATION, THE GUARDIAN, I AM WRITING TO EXPRESS MY SUPPORT FOR THE PHEME PROPOSAL ON AUTOMATIC EXTRACTION AND VISUALIZATION OF SOCIO-SEMANTIC INTELLIGENCE, ACROSS MULTIPLE LANGUAGES, MODALITIES, AND CONTENT TYPES (SOCIAL MEDIA, ONLINE NEWS, STRUCTURED DATA, AND OTHER AUTHORITATIVE SOURCES).

THE GUARDIAN HAS MADE A COMMITMENT TO WORK IN AN OPEN, TRANSPARENT AND COLLABORATIVE WAY IN ITS NEWSGATHERING AND REPORTING PROCESSES. THIS INCLUDES WORKING TO IDENTIFY NEW SOURCES OF INFORMATION THROUGH SOCIAL PLATFORMS.

JOURNALISTS INCREASINGLY MONITOR AND INTEGRATE INFORMATION FROM SOCIAL MEDIA AND USER-GENERATED CONTENT, ALONGSIDE MORE TRADITIONAL SOURCES. THE TECHNOLOGIES AND METHODS TO BE DEVELOPED IN PHEME, NAMELY THOSE FOR AUTOMATIC DETECTION AND VISUALISATION OF RUMOURS, MISINFORMATION, AUTHORITY, AND INFORMATION DIFFUSION, WOULD BE OF MAJOR HELP TO GAIN TIMELY, STRATEGIC KNOWLEDGE AND INTELLIGENCE FROM CONFLICTING INFORMATION STREAMS.

THE GUARDIAN IS HAPPY TO SUPPORT PHEME THROUGH PARTICIPATION IN THE USER GROUP AND ADVICE ON THE USER NEEDS AND REQUIREMENTS OF OUR INDUSTRY.

I HOPE THAT THE PHEME PROPOSAL IS SUCCESSFUL, SO WE CAN INITIATE THIS WORK.

YOURS SINCERELY,

JOANNA GEARY
SOCIAL & COMMUNITIES EDITOR
JOANNA.GEARY@GUARDIAN.CO.UK
+44(0)20 3353 3243
@GUARDIANJOANNA
SOCIAL & COMMUNITIES TEAM, EDITORIAL

GUARDIAN NEWS & MEDIA LIMITED A MEMBER OF GUARDIAN MEDIA GROUP PLC
REGISTERED OFFICE PO Box 68164, KINGS PLACE, 90 YORK WAY, LONDON N1P 2AP REGISTERED IN ENGLAND NUMBER 908396

The University of Sheffield
Department of Computer Science
Dr. Kalina Bontcheva
Regent Court, 211 Portobello
Sheffield S1 4DP
United Kingdom

14.1.2013

Dear PHEME Coordinator,

On behalf of our organisation, BBC World Service, we are writing to express our strong support for the PHEME proposal on automatic extraction and visualization of socio-semantic intelligence, across multiple languages, modalities, and content types (social media, online news, structured data, and other authoritative sources).

BBC World Service is an international broadcaster, producing news in 28 languages across the Globe, who are increasingly reliant on digital distribution for their content- in both traditional and non traditional digital platforms.

Journalists within the World Service, as well as our product development teams, increasingly monitor and integrate information from social media and user-generated content, alongside the more traditional and authoritative sources, such as surveys, web analytics, proprietary research and internal sources, in order to ensure that their research is timely, involving and engaging to the audiences, as well as enabling them to pro-actively adapt and create content in a speedier manner than perhaps more traditional sources can provide. The technologies and methods to be developed in PHEME, namely those for automatic detection and visualisation of rumours, misinformation, authority, and information diffusion, would be of major help to gain timely, strategic knowledge and intelligence from conflicting information streams.

The BBC World Service is happy to support PHEME through participation in the user group and advice on the user needs and requirements of our industry.

We hope that the PHEME proposal is successful, so we can initiate this exciting work.

Sincerely yours,

Jemma Ahmed, Digital Insights Manager
Mohammed Abdul Qader, Languages and Social Media Editor

BBC Research & Development

Media Centre
201 Wood Lane
London
W12 7TQ
UK

The University of Sheffield
Department of Computer Science
Dr.Kalina Bontcheva

11/01/2013

Regent Court, 211 Portobello
Sheffield S1 4DP
United Kingdom

Dear PHEME Coordinator,

On behalf of my organisation, the BBC, I am writing to express my strong support for the PHEME proposal on automatic extraction and visualization of socio-semantic intelligence, across multiple languages, modalities, and content types (social media, online news, structured data, and other authoritative sources).

The BBC is one of the world's leading media and news organizations with many national and international TV, radio and online services. BBC News is the largest broadcast news operation in the world with more than 2,000 journalists and 48 newsgathering bureaux, 41 of which are overseas. BBC News is respected both in the UK and around the world for the strength of its journalism and impartiality. The BBC is fully committed to strong research collaborations to ensure the BBC can become the most valued, open, digital media services in the world. As a result it is our aspiration to support a rich network of research projects as well as developing ground-breaking opportunities for collaboration on projects of significant strategic benefit.

Journalists increasingly monitor and integrate information from social media and user-generated content, alongside the more traditional and authoritative sources, such as official news wires or personal contacts in order to produce high quality, original and trusted journalism. The technologies and methods to be developed in PHEME, namely those for automatic detection and visualisation of rumours, misinformation, authority, and information diffusion, would be of major help to gain timely, strategic knowledge and intelligence from conflicting information streams.

The BBC is happy to support PHEME through participation in the user group and advice on the user needs and requirements of our industry.

I hope that the PHEME proposal is successful, so we can initiate this exciting work.

Yours sincerely,

Tristan Ferne
Executive Producer, BBC R&D



Health Protection Agency

Microbiology Services

Rare and Imported Pathogens,
Porton Down
Salisbury
Wiltshire, SP4 0JG

Tel +44 (0)1980 612774
Fax +44 (0)1980 612695
www.hpa.org.uk

The University of Sheffield
Department of Computer Science
Dr. Kalina Bontcheva
Regent Court, 211 Portobello
Sheffield S1 4DP
United Kingdom

17th January 2013

Dear PHEME Coordinator,

On behalf of my organisation, the **Health Protection Agency**, I am writing to express my strong support for the PHEME proposal on automatic extraction and visualization of socio-semantic intelligence, across multiple languages, modalities, and content types (social media, online news, structured data, and other authoritative sources).

The Health Protection Agency is an independent UK organisation that was set up by the government in 2003 to protect the public from threats to their health from infectious diseases and environmental hazards. It does this by providing both practical support, and impartial advice and information on health protection issues, to the public, to professionals and to government. In April 2013 the Health Protection Agency will become part of a new organisation called Public Health England, an executive agency of the Department of Health. The activities of the Health Protection Agency will transfer to this new body to become part of its broad public health remit to protect and improve the nation's health and wellbeing.

To predict and prepare for public health threats requires continuous horizon scanning and knowledge surveillance. Public health practitioners have traditionally monitored professional resources such as scientific citation databases (e.g. PubMed) and disease alert systems (e.g. ProMed). However, to be ahead of the curve in identifying, preparing for and managing public health threats, there is an increasing need to monitor and integrate information from less authoritative sources such as social media and user-generated content. The technologies and methods to be developed in PHEME, namely those for automatic detection and visualisation of rumours, misinformation, authority, and information diffusion, would be of major help to gain timely, strategic knowledge and intelligence from conflicting information streams, to be used in addition to, or in advance of, information from more formal authoritative sources.

The **Health Protection Agency** is happy to support PHEME through participation in the User Group and to advise on the user needs and requirements of our sector.

I hope that the PHEME proposal is successful, so we can initiate this exciting work.

Sincerely yours,

A handwritten signature in black ink, appearing to read "Tim Brooks", written over a thin horizontal line.

Dr Tim Brooks
Head & Clinical Services Director
Rare and Imported Pathogens Department

APPENDIX B: REFERENCES

- Abel, F; Gao, Q; Houben, GJ; Tao, K (2011) Semantic enrichment of Twitter posts for user profile construction on the social web. In *ESWC (2)*, pages 375–389.
- Adams, B., Phung, D., and Venkatesh, S. (2011) Eventscapes: Visualizing events over time with emotive facets. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1477–1480.
- Angeletou, S., Rowe, M., and Alani, H. (2011) Modelling and analysis of user behaviour in online communities. In *Proceedings of the 10th International Conference on the Semantic Web, ISWC'11*, pages 35–50. Springer-Verlag.
- Anderson, J. and Rainie, L. (2012) *The Future of Big Data*. Pew Internet Research. <http://www.pewinternet.org/Reports/2012/Future-of-Big-Data>. Accessed on January 11th, 2013.
- Bansal, N. and Koudas, N. (2007) Blogscope: Spatio-temporal analysis of the blogosphere. In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 1269–1270.
- Barbour, A. and Mollison, D. (1990). Epidemics and random graphs. In Gabriel, Lefevre, and Picard, editors, *Stochastic Processes in Epidemic Theory*, number 86 in *Lecture Notes in Biomaths*, pages 86-89. Springer.
- Beckett, C; Mansell, R (2008) Crossing boundaries: New media and networked journalism. *Communication, Culture & Critique*, 1(1), 92-104.
- Belák, V., Lam, S. & Hayes, C., (2012) Targeting online communities to maximise information diffusion. In *Proceedings of the 21st international conference companion on World Wide Web. WWW '12 Companion*. New York, NY, USA: ACM, pp. 1153–1160. Available at: <http://doi.acm.org/10.1145/2187980.2188255> [Accessed November 20, 2012].
- Bendersky, M. & Croft, W.B. (2009) Finding text reuse on the web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining. WSDM '09*. New York, NY, USA: ACM, pp. 262–271. Available at: <http://doi.acm.org/10.1145/1498759.1498835>.
- Bostock, M., Ogievetsky, V. and Heer, J. (2011). “D3: Data-Driven Documents”, *IEEE Transactions on Visualization and Computer Graphics*, 17(12): 2301-2309.
- Broder, AZ (2000) Identifying and Filtering Near-Duplicate Documents, *Combinatorial Pattern Matching (CPM) Symposium*, Montreal, Canada.
- Bruns, A. (2006). Wikinews: The Next Generation of Online News? *Scan Journal* 3(1).
- Castillo, Mendoza, & Poblete (In press). Information Credibility in Time-Sensitive Social Media. *Internet Research*.
- Cha, M., Mislove, A. & Gummadi, K.P. (2009) A measurement-driven analysis of information propagation in the flickr social network. In *WWW '09: Proceedings of the 18th international conference on World wide web*. New York, NY, USA: ACM, pp. 721–730.
- Chambers (2012) Labeling Documents with Timestamps: Learning from their Time Expressions. *ProcACL*.
- Chang, H.-F. & Mockus, A. (2008) Evaluation of source code copy detection methods on freebsd. In *Proceedings of the 2008 international working conference on Mining software repositories. MSR '08*. New York, NY, USA: ACM, pp. 61–66. Available at: <http://doi.acm.org/10.1145/1370750.1370766> [Accessed September 27, 2012].
- Chimera, R. and Shneiderman, B. (1994). “An Exploratory Evaluation of Three Interfaces for Browsing Large Hierarchical Tables of Contents”, *ACM Transactions on Information Systems*, 12(4): 383-406.
- Clough, P., Gaizauskas, R., Piao, S., Wilks, Y. (2002). Measuring Text Reuse. *ACL*: 152-159.
- Collier & Doan (2011). Syndromic Classification of Twitter Messages. *eHealth 2011*.
- Costa & Branco (2010) Temporal inf. processing of a new language: Fast porting with minimal resources. *Proc. ACL*
- Derczynski & al (2012) Massively Increasing TIMEX3 Resources: A Transduction Approach. *Proceedings LREC*.
- Donald, E.E. & Jones, P.E. (2001) US Secure Hash Algorithm 1 (SHA1). <http://tools.ietf.org/html/rfc3174>.
- Dork, M., Gruen, D., Williamson, C., and S. Carpendale. (2010) A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138.
- Dowman, M; Tablan, V; Cunningham, H; Popov, B (2005) Web-assisted annotation, semantic indexing and search of television and radio news. In *Proceedings of the 14th International World Wide Web Conference (WWW)*.

- Ennals, R; Trushkovsky, B; Agosta, JM (2010) Highlighting Disputed Claims on the Web. WWW'10.
- Gaizauskas & al (2012) Applying ISO-Space to Healthcare Facility Design Reports. Proc. ISA-7 workshop.
- Gomez-Rodriguez, M., Leskovec, J. & Krause, A., (2012) Inferring Networks of Diffusion and Influence. ACM Trans. Knowl. Discov. Data, 5(4), pp.21:1–21:37.
- Gomez-Rodriguez, M., Leskovec, J. & Krause, A., (2012) Inferring Networks of Diffusion and Influence. ACM Trans. Knowl. Discov. Data, 5(4), pp.21:1–21:37.
- Grimes, Seth. (2011).Text/Content Analytics 2011: User Perspectives on Solutions and Providers. AltaPlana.
- Guerin, B; Miyazaki, Y (2006) Analyzing rumours, gossip, and urban legends through their conversational properties. The Psychological Record. vol. 56, pp. 23-34.
- Gruhl, D. & al (2005) The predictive power of online chatter. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. KDD '05. New York, NY, USA: ACM, pp. 78–87. Available at: <http://doi.acm.org/10.1145/1081870.1081883> [Accessed January 21, 2012].
- H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics, 45(5): 885 – 892.
- Gyongi & al (2004) Combating web spam with trustrank. Proc VLDB
- Hänska-Ahy, MT; Shapour, R (2013) “Who’s reporting the protests?”, Journalism Studies, Vol. 14, Issue 1, 2013, <http://eprints.lse.ac.uk/41674/1/Whos%20reporting%20the%20protests%20%28Isero%29.pdf>
- Harabagiu, S; Hickl, A; Lacatusu, F (2006) Negation, Contrast and Contradiction in Text Processing. AAAI.
- Hatch, Frissa, Verdecchia, Stewart, et al. Identifying socio-demographic and socioeconomic determinants of health inequalities in a diverse London community: the South East London Community Health (SELCoH) study. BMC Public Health 2011; 11: 861.
- Hey, T., Tansley, S. and Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft.
- Hirschman, D., Rich, L. (2012) Local news gets automated. <http://www.niemanlab.org/2012/12/local-news-gets-more-automated/>. Visited on Jan 5th, 2013.
- Hubmann-Haidvogel, A., Brasoveanu, A. M. P., Scharl, A., Sabou, M., and Gindl, S. (2012) Visualizing contextual and dynamic features of micropost streams. In Proceedings of the #MSM2012 Workshop, CEUR, volume 838.
- 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.
- Hullman, J. and Diakopoulos, N. (2011). “Visualization Rhetoric: Framing Effects in Narrative Visualization”, IEEE Transactions on Visualization and Computer Graphics, 17(12): 2231-2240.
- Ji & al (2011). Overview of the TAC2011 Knowledge Base Population task.
- Katz, E. and Lazarsfeld, P.F. (1955). Personal Influence. Glencoe, IL: Free Press.
- Kawahara, D; Inui, K; Kurohashi, S (2010) Identifying Contradictory and Contrastive Relations between Statements to Outline Web Information on a Given Topic. COLING.
- Kimura, M. & al (2010) Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery, 20(1), pp.70–97.
- Kolya & al (2012) A Hybrid Approach for Event Extraction. Polibits 46.
- Krieger, H-U (2010) A Temporal Extension of the Hayes and ter Horst Entailment Rules for RDFS and OWL. In Proceedings of the AAAI 2011 Spring Symposium "Logical Formalizations of Commonsense Reasoning.
- Krippendorff, K. (2004). Content Analysis: An Introduction to Its Methodology. 2nd edition, Thousand Oaks, CA: Sage.
- Kunegis, J. (2013). KONECT - the Koblenz Network Collection. konect.uni-koblenz.de.
- Lampos, V., De Bie, T., Cristianini. (2010). Flu Detector - Tracking Epidemics on Twitter. ECML/PKDD (3): 599-602
- Levin and Alexander. 1998. Attack resistant trust metrics for public key certification. USENIX Security Symposium.

- Li, Bontcheva and Cunningham (2009). Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. *Natural Language Engineering*, 15(02), 241-271.
- Lim, S.-H. & al (2009) Determining content power users in a blog network. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis. SNA-KDD '09*. New York, NY, USA.
- Llorens & al (2012) TIMEN: An open temporal expression normalisation resource. *Proc. LREC*.
- Lotan, Graeff, Ananny, Gaffney, Pearce, and Boyd (2011). The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions, *International Journal of Communication* (5) Feature: 1375-1405.
- Mahmud, J; Nichols, J; Drews, C (2012) Where is this tweet from? Inferring home locations of twitter users. In *Proc. ICWSM*, pages 511–514.
- Manes, M. (2012) Breaking is broken. <http://www.niemanlab.org/2012/12/breaking-is-broken/>. Visited on 14/01/2013.
- Mani; Wilson (2001) Robust temporal processing of news. *Proceedings of ACL*.
- Magnani, M., Rossi, L. (2011). The ML-Model for Multi-layer Social Networks. *Int. Conf. on Advances in Social Networks Analysis and Mining, ASONAM 2011*.
- Marneffe, M-C; Rafferty, A; Manning, C (2008) Finding Contradictions in Text. *Proceedings of ACL-08: HLT*.
- Marsono, M.N. (2011) Packet-level open-digest fingerprinting for spam detection on middleboxes. *Int. J. Netw. Manag.*, 22(1), pp.12–26.
- Mary McGlohon, M.H., (2009) Community Structure and Information Flow in Usenet: Improving Analysis with a Thread Ownership Model.
- Mendoza, M; Poblete, B; Castillo, C (2010). Twitter under crisis: Can we trust what we RT? *1st Workshop on Social Media Analytics (SOMA)*.
- Miller, M. & al (2011) Sentiment Flow Through Hyperlink Networks. In *Fifth International AAAI Conference on Weblogs and Social Media*.
- Morency, L., Mihalcea, R., and Doshi, P. (2011). Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web, in *Proceedings of the International Conference on Multimodal Interaction, Alicante*.
- Myers, S.A., Zhu, C. & Leskovec, J., (2012) Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '12*.
- Naaman, M. , Boase, J., and Lai. C. (2010) Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, pages 189–192.
- Nagarajan, M., Gomadam, K., Sheth, A., Ranabahu, A., Mutharaju, R., and Jadhav, A. (2009) Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In *Web Information Systems Engineering*, pages 539–553.
- Nel & al (2010) Rumour detection and monitoring in open source intelligence: understanding publishing behaviours as a prerequisite. *Proc. Terrorism and New Media conference*
- Newman, M.E.J. (2002) The spread of epidemic disease on networks. *Physical Review E*, 66(016128). Available at: <http://arxiv.org/abs/cond-mat/0205009> [Accessed November 23, 2012].
- Kang, N., Afzal, Z., Singh, B., van Mulligen, E.M., and Kors, J.A. (2012) Using an ensemble system to improve concept extraction from clinical records. *Journal of Biomedical Informatics*.
- Krestel, R., Bergler, S., Witte, R. (2008). A Belief Revision Approach to Textual Entailment Recognition. *TAC 2008*.
- Krestel, R., Witte, R., Bergler, S. (2010). Predicate-Argument EXtractor (PAX). *Proc. of the First Workshop on New Challenges for NLP Frameworks. LREC 2010*.
- Qazvinian, V; Rosengren, E; Radev, D; Mei, Q (2011) *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1589 - 1599.
- Quercia & al (2011) In the mood for being influential on Twitter. *IEEE Int. Conf. on Social Computing (SocialCom)*.
- Paulevé, L., Jégou, H. & Amsaleg, L. (2010) Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11), pp.1348–1358.

- Pérez-Díaz, N. & al (2012) SDAI: An integral evaluation methodology for content-based spam filtering models. *Expert Systems with Applications*, 39(16), pp.12487–12500.
- Procter, R., Vis, F. and Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, Special Issue on Computational Social Science: Research Strategies, Design & Methods.
- Procter, R., Voss, A. and Brooker, P. A Study of Using Social Media in Journalism. Internal report. Univ. of Warwick.
- Pustejovsky & al (2010) ISO-TimeML: An international standard for semantic annotation. *Proceedings LREC*.
- Pustejovsky & al (2011) ISO-Space: The annotation of spatial information in language. *Proceedings ISA-6 workshop*.
- Quercia, D., Ellis, J., Capra, L., Crowcroft, J. (2011) 3rd IEEE Int. Conference on Social Computing (SocialCom).
- Ratkiewicz, J; Conover, M; Meiss, M; Gonçalves, B; Flammini, A; Menczer, F (2011) Detecting and Tracking Political Abuse in Social Media, in 'Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)' .
- Ritter, A; Downey, D; Soderland, S; Etzioni, O (2008) It's a Contradiction - No, it's Not: A Case Study using Functional Relations. *EMNLP' 2008*.
- Ritter, A; Clark, S; Mausam; Etzioni, O (2011) Named Entity Recognition in Tweets: An Experimental Study. *Proceedings of EMNLP 2011*.
- Rivest, R.L. (1992) The MD5 Message-Digest Algorithm. Available at: <http://tools.ietf.org/html/rfc1321>.
- Romero, D.M., Meeder, B. & Kleinberg, J. (2011) Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proc. of the 20th Int. Conference on World Wide Web*.
- Romero & al (2010) Influence and Passivity in Social Media. *Proc ACM WWW*
- Rossi, L and Magnani, M (2012) Conversation Practices and Network Structure in Twitter. *ICWSM*
- Sadilek, A; Kautz, H; Silenzio, V. (2012) Modeling spread of disease from social interactions. *Proc. ICWSM*, 322–329.
- Salathé, Khandelwal (2011) Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Comput Biol* 7(10).
- Scerri, S., Cortis, K., Rivera, I., and Handschuh, S. (2012) Knowledge Discovery in Distributed Social Web Sharing Activities. In *Proceedings of the #MSM2012 Workshop*, CEUR, volume 838.
- Scharl, Hubmann-Haidvogel, et al. (2013). Media Watch on Climate Change – Visual Analytics for Aggregating and Managing Environmental Knowledge from Online Sources. *46th Hawaii Int. Conf. on Systems Sciences (HICSS-46)*.
- Scharl, A., Weichselbraun, A. and Liu, W. (2007). “Tracking and Modelling Information Diffusion across Interactive Online Media”, *International Journal of Metadata, Semantics and Ontologies*, 2(2): 136-145.
- Segel, E. and Heer, J. (2010). “Narrative Visualization: Telling Stories with Data”, *IEEE Transactions on Visualization and Computer Graphics*, 16(6): 1139-1148.
- Shamma, D.A., Kennedy, L. and Churchill, E.F. (2010) Tweetgeist: Can the Twitter timeline reveal the structure of broadcast events? In *Proceedings of CSCW 2010*.
- Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, Hotopf M, Thornicroft G, Lovestone S. (2009). The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) Case Register: development and descriptive data. *BMC Psychiatry*; 9: 51.
- Strötgen, J; Gertz, M (2010) HeidelTime: High quality rule-based extraction and normalization of temporal expressions. *SemEval*.
- Tang, J. & al (2009) Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '09*.
- TrstRank (2010) Importance Score. <http://www.infochimps.com/datasets/twitter-census-trst-rank>
- Verhagen & al (2010) SemEval-2010 task 13: TempEval-2. *Proceedings of SemEval*.
- Voorhees, E (2008) Contradictions and Justifications: Extensions to the Textual Entailment Task. *ACL-08: HLT*.

- Wang, C., Chen, W. & Wang, Y., (2012) Scalable influence maximization for independent cascade model in large-scale social networks. *Data Mining and Knowledge Discovery*, 25(3), pp.545–576.
- Weenig & al (2001) Bad news transmission as a function of the definitiveness of consequences and the relationship between communicator and recipient. *J Personality and Social Psychology*, vol. 80 3 449-461.
- Weerkamp & de Rijke. (2012). Credibility-inspired Ranking for Blog Post Retrieval. *Information Retrieval*. 15.
- Weiss, E. S. (2012) Mobile, location, data. <http://www.niemanlab.org/2012/12/mobile-location-data/>.
- Wu, S., Hofman, J.M., Mason, W.A. and Watts, D.J. (2011). Who says what to whom on Twitter. *Proceedings of WWW'11*.
- Xu, J., Bhargava, A., Nowak, R., and Zhu, X. (2012) Robust spatio-temporal signal recovery from noisy counts in social media. *Arxiv preprint arXiv:1204.2248*.
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., and Li, X. (2011) Comparing Twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR)*.
- Zhou & et al (2004) Automatic Linguistics-Based Cues for Detecting Deception in Text-based Asynchronous Computer-Mediated Communication. *Group Decision and Negotiation* 13: 81 – 106, 2004.