Small or medium-scale focused research project (STREP) proposal
ICT Call 5
FP7-ICT-2009-5

News Event Extraction & Description
When is a news items news, and when does it change* MY *world?

NEWS EVENT EXTRACTION AND DESCRIPTION

Small or medium scale focused research project (STREP)
Date of preparation: October 26th 2009
Version number *(optional)*: 1.0

Work programme topic addressed

Objective ICT-2009.4.3: Intelligent Information Management

Name of the coordinating person: Erik Mannens
e-mail: erik.mannens@ugent.be
fax: +32 9 331 48 96

Participant no. *	Participant organisation name	Part. short name	Country
1 (Coordinator)	Interdisciplinary Institute for Broadband Technology	IBBT	Belgium
2	EURECOM	EURECOM	France
3	Stichting Centrum voor Wiskunde en Informatica	CWI	The Netherlands
4	Agence France-Presse	AFP	France
5	GEIE ERCIM	ERCIM/W3C	France
6	TEMIS SA	Temis	France
7	CINECA Consorzio Interuniversitario	CINECA	Italy

Proposal abstract

With the emergence of both citizen-based media and social media, media have to fundamentally re-think their production and distribution workflow processes. In addition, traditional text-based products are losing market share for other media products, in particular video. At the same time, users have access to multiple news portals, which provide online access to different sources and services for commenting and debating on the news, and use social media to instantaneously spread news information. This results in large amounts of unreliable and repeated information, leaving the user exploring on their own to try to build their own version of a news event from large amounts of potentially related information, or simply to find the truth in the middle of an ocean of rumours or hoaxes.

NEED will provide services for analyzing streams of video and textual news data, provide interoperability solutions in established news workflows and offer user interfaces providing access to heterogeneous sources of information, in multiple media types and in multiple languages centred on a macro-view of news events. By developing knowledge models for news events, **NEED** will provide event-centric navigation interfaces. These models will be populated with annotations of news items, using semi-automatic text processing and multimedia content analysis. Scheduled and breaking news events will be detected, analyzed, described formally and linked to existing background knowledge available on the web. News events sharing across social media sites will be analyzed both technologically and socially providing feedback to journalists, decision makers and citizens on how news information is consumed and perceived across cultures and communities.

For this open-knowledge approach to news workflow to work on an industrial scale in practice, all information components need to conform to open standards. Consortium members have close ties with different standards bodies to ensure that the real problems from existing communities are addressed within the project and that the solutions developed within the project can feed into the relevant standardisation activities. This will ensure broad uptake for establishing metadata workflows in the production and consumption chains throughout the news industry.

Table of contents

Section 1: Scientific and/or technical quality, relevant to the topics addressed by the call

1.1 Concept and objectives

1.1.1 Motivation

Nearly every European citizen reads, watches or listens to the news, at home, while commuting to and from work, at work and even as part of their work. As voting citizens, we need to understand local, national and international politics to allow us to cast our vote. As company employees, we need to understand the state and development of local, national and international economies to enable us to understand our markets. As part of our leisure time, we want to know about our favourite sports teams, the lives of our soap idols or the most recent books available. Nowadays, this information is online, and hence easily accessible from anywhere.

Information is online and available, but through many different sources, including branded websites. Traditional news providers (e.g. journalists, news agencies, press and broadcasters) make use of the Web for distributing news. More recently, non-traditional news providers, often called citizen media or independent media, make use of Web technologies to publish alternative views and opinions of events. Furthermore, we observe an increasing tendency of the social media sites playing a crucial role either in spreading news at a speed never seen (e.g. the death of Michael Jackson) or simply in informing citizens in countries or situations where other traditional communication means fail (e.g. the coverage of the protests following the controversial Iran elections). Professional users have therefore access to continuous streams of incoming data from press agencies and archives of published news to tweets, images and videos posted by anonymous people. Lay users have access to myriads of web sites, offering push and pull sources of news information and the means of contributing to the expanding collection of news data (e.g. Twine[1]).

These billion of sources contain large amounts of isolated information that lack context, leaving users with the immense problem of manually finding related information. Typical tasks include: finding the same event covered from a different (political) angle, finding the role of the event in a wider historical perspective, finding and checking the original sources on which a story is based, etc. News aggregators only tend to amplify the problem by aggregating pointers to isolated articles, rather than analyzing and contextualizing events from the many continuous data streams. Faced with this upraise of citizen-based media and social networks, traditional media must re-think their fact-checking processes while citizen need also tools to enhance their user experience in terms of knowledge and confidence, when reading and watching news online.

In existing news workflow processes, news items are typically *i)* produced by news agencies, independent journalists or citizen media, *ii)* consumed and enhanced by newspapers, magazines or broadcasters then *iii)* delivered to end-users and finally *iv)* perceived by these end-users that further let a trace of what they have understood and felt facing the news event using blogs and tweets as means of expression. News items are typically accompanied by a set of metadata and descriptions that facilitate their storage, retrieval and lifecycle. However, **much of the metadata is lost because of interoperability problems occurring along the workflow**. In addition, at the end user interface, opportunities for making use of the available metadata are often lost.

News web sites such as Le Monde[2], El Pais[3] or La Repubblica[4] generally classify news in categories such as: *World*, *National*, *Politics*, *Business*, *Science and Technology*, *Sport*, *Entertainment* and *Health*, while

[1] http://www.twine.com/
[2] http://www.lemonde.fr/
[3] http://www.elpais.com/
[4] http://www.repubblica.it/

other services such as Google News[5] aggregate stories from multiple sources and offer personalised selections based on the user topics of interest. More advanced web sites such as SiloBreaker[6] or Newstin[7] provide more flexible access to news stories by *topic*, *person*, *organization* or *region*. Specific services attempt to extract trends from large amount of micro-blogging feeds. These support the user's information need more closely, but just add more sources of individual news items, which leads to overly complex interfaces.

Current systems have a number of limitations that force the user to explore news information in an environment that contains **large amounts of irrelevant, unreliable and repeated information, with insufficient access to background knowledge**. In particular, current systems:

- mainly deal with textual news articles in a single language (mostly English), and do not process audiovisual content at the same level of detail;
- are unable to provide explicit relationships between different news on the same event to help the user form his/her own opinion on a particular topic, e.g. cannot automatically link a quote in a news article to the original statement in a video clip, or link a statement to the subsequent reactions;
- are unable to handle the evolution of news events, e.g. do not link the first announcement of an explosion to its subsequent interpretation as a terrorist attack;
- are unable to provide a historic perspective of events, e.g. do not show the chain of events that led to the information the reader is focusing on, and do not highlight *editorial* news items summarizing events that took place years ago.

Consequently, users are overwhelmed by too many individual and disconnected pieces of information, and cannot situate news in a proper context. A news event is defined as a cluster of statistically related news items, but no ontological notion of event is supported. In contrast, the organization of news providers is centred on the notion of scheduled and breaking news events, but they currently lack the tools necessary to relate easily the news they produce to the events they manage on a daily basis.

Semantic processing of news information can improve the clustering and organization of individual news items – from heterogeneous sources, in multiple media types and in multiple languages – into meaningful events linked to appropriate background knowledge.

In this project, **we will build the technological infrastructure to allow the aggregation of large amount of multiple, distributed information sources** ranging from professional media content to user generated content, using encyclopaedic knowledge and micro-blogging feeds. Based on this, **we will provide *event based* user interfaces driven by semantic metadata** for searching and browsing multimedia news articles, independently of whether the news is expressed in text or audio/visual media.

1.1.2 Use Case Scenarios

To further exemplify the **NEED** vision, we provide below two representative use cases scenarios that will be developed during the project.

Scenario 1: Providing a portal watching environmental issues for EU citizens

A non-governmental organization such as Greenpeace is interested to get a real-time portal informing citizens about a specific theme in order to strengthen its lobbying activity. In this case, the theme deals with environmental issues. Such a portal should not only harvest, digest, and prioritize different input streams around environmental issues (i.e., stories, photos, graphics and videos), but it should also present to its users with the past and upcoming events around this theme with relevant background and context information.

Let us take for example the topic "*climate change*", a sub-topic of environmental issues, and in particular the coverage of the typhoon Parma which recently hit the Philippines. The real-time portal regarding environmental issues will report on different breaking-news items concerning Parma. These different incoming news events come from heterogeneous sources, being professional organizations or user-generated

[5] http://news.google.com/
[6] http://www.silobreaker.com/
[7] http://www.newstin.com/

content. In case of professional organizations, news agencies such as AFP, CNN or Reuters are a traditional source of information, but organizations such as the Red Cross[8] provide also breaking news feeds regarding natural disasters such as Parma. Aggregators such as BreakingNews.com[9] or AlertNet[10] gather additional breaking news regarding the typhoon. Finally, social networking sites such as Twitter, Plurk, or Facebook provide more and more real time information regarding the progress of the typhoon, the regions it has stroke and a very first estimate of the damage. These short messages sometimes refer to more user generated content such as videos post on YouTube that are also worth to watch and analyze in order to cover the event. For example, when the typhoon Parma stroke the Philippines, people were actively posting and re-posting information about missing people, where to find or give help and even posted reports on areas that badly needed government attention.

Next to breaking news, the portal could also provide contextual information related to the event, such as data about the past typhoons hitting the Philippines. Encyclopedia such as Wikipedia[11] or general news providers[12] gather and curate data about the history of typhoons hitting the Philippines.

For such an event, many input streams are available. The following excerpts are just an example of the large amount of data that **NEED** will process, analyze, and mashup into meaningful presentation:

- News item from AFP:

  "MANILA, Oct 3, 2009 (AFP) - Typhoon Parma ripped open houses and cut off power lines as it smashed into the northern Philippines on Saturday, bringing more devastation to the Southeast Asian country after deadly floods. However the typhoon veered away from Manila, sparing millions of people who were struggling to recover from massive rains that submerged most of the capital, claiming nearly 300 lives, last weekend. Parma, packing winds of 175 kilometres (110 miles) per hour and gusts of up to 210 kilometres per hour, began lashing the northern province of Cagayan and surrounding areas about midday (0400 GMT). '*The wind is very, very angry*,' Cagayan regional police chief Roberto Damian said in a radio interview from his headquarters, about 400 kilometres (250 miles) from the Philippine capital. *'I can see trees are being toppled inside our camp.... One sturdy Narra tree was uprooted and smashed a car and a house. We cannot go out,*' he said in a radio interview before his line went dead. Cagayan is a mainly rural area with coastal towns and a population of just over a million people. Parma caused major damage in Tuguegarao, the capital of Cagayan with a population of 130,000, according to the city's mayor, Delfin Ting."

From which named entities such as Typhoon Parma, Oct 3 2009, Cagayan[13], Tuguegaro[14] can be extracted leading to more encyclopaedic information. Similarly, multimedia analysis and in particular quote detection temporal alignment can be performed in order to link the quotes from this regional police chief with the video sequence where he actually pronounced this statement.

[8] http://newsroom.redcross.org/ and http://newsroom.redcross.org/2009/10/06/disaster-alert-typhoon-parma/
[9] http://www.breakingnews.com/
[10] http://www.alertnet.org/
[11] http://en.wikipedia.org/wiki/Typhoons_in_the_Philippines
[12] http://www.gmanews.tv/story/173677/past-super-typhoons-in-the-philippines
[13] http://en.wikipedia.org/wiki/Cagayan
[14] http://en.wikipedia.org/wiki/Tuguegarao_City

- Images showing affected cities in Cagayan:

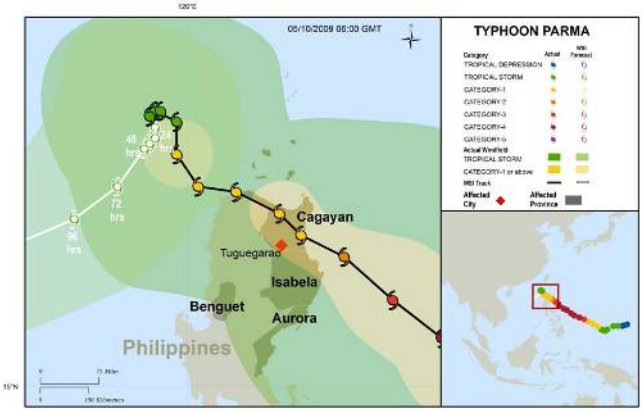


Figure 1: Credit: Guy Carpenter's Instrat® department provides CAT-i reports for major natural catastrophes worldwide. These reports cover catastrophes including worldwide tropical cyclones, earthquakes, major UK and European floods and any other natural event that is likely to incur a significant loss to the (re)insurance industry, see http://www.gccapitalideas.com/2009/10/05/update-typhoon-parma/

- Twitter / Plurk and micro-blogging feeds:

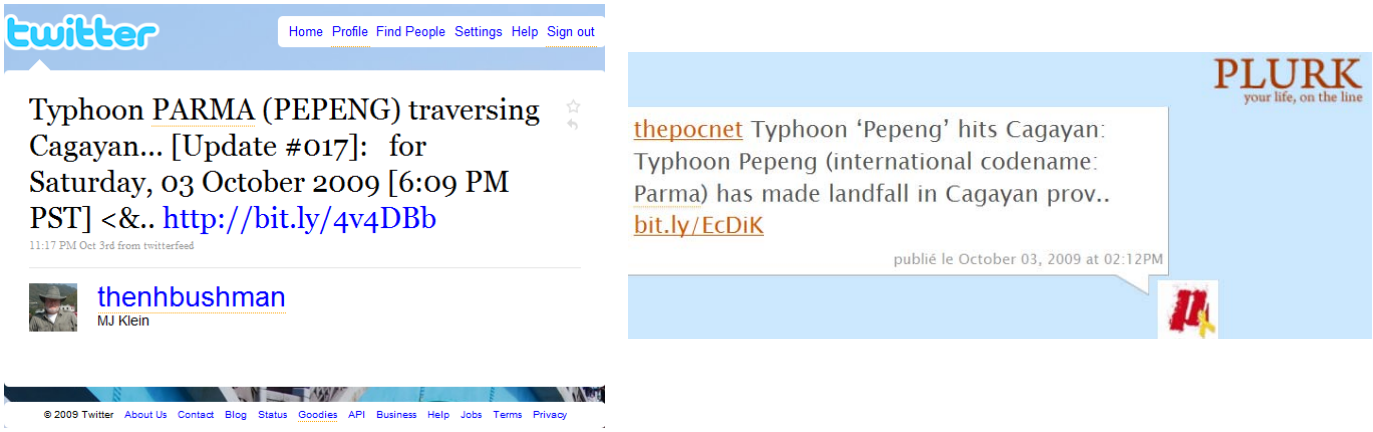


Figure 2: Micro-blogging feeds about Parma, http://twitter.com/thenhbushman/status/4587734968 and http://www.plurk.com/p/256q3m

This tweet posted on October 3rd indicates that typhoon Parma is traversing Cagayan at this moment. A link with more statistics regarding the typhoon is also provided. The Plurk status message points to a breaking news article about the typhoon.

- YouTube movie

Figure 3: Typhoon Parma hits Phillippines, http://www.youtube.com/watch?v=x_LabpBEgms

The description of this video reads "*... somewhere on the far northern coast of Luzon in the eyewall of Parma ...*" which provides provenance information about this video content. Again, encyclopedic information will highlight the fact that Cagayan is located in the Northern coast of Luzon[15] and therefore that this movie can be linked with stories about Cagayan damages.

These few examples just highlight the fact that for any type of scheduled or breaking news event, interested users can be easily overwhelmed by too many individual and disconnected pieces of information and therefore cannot situate articles in a proper context. A news event is defined as a cluster of statistically related news items, but no ontological notion of event is supported. In contrast, news agencies currently lack the tools necessary to relate easily the news they produce to the events they manage on a daily basis, let alone take the real-time streams (e.g. Twitter, YouTube) into consideration.

NEED will provide services and technologies to process these many sources and provide meaningful and contextualized presentation of the information so that end-users have all the keys to actually understand news events, their causes and consequences and have access to all the dimensions that constitute an event.

NEED will provide appropriate interfaces for supporting the end-user in these tasks, giving him/her the appropriate level of real-time information, contextualized with background knowledge but without overwhelming him/her by unnecessary information.

[15] http://en.wikipedia.org/wiki/Luzon

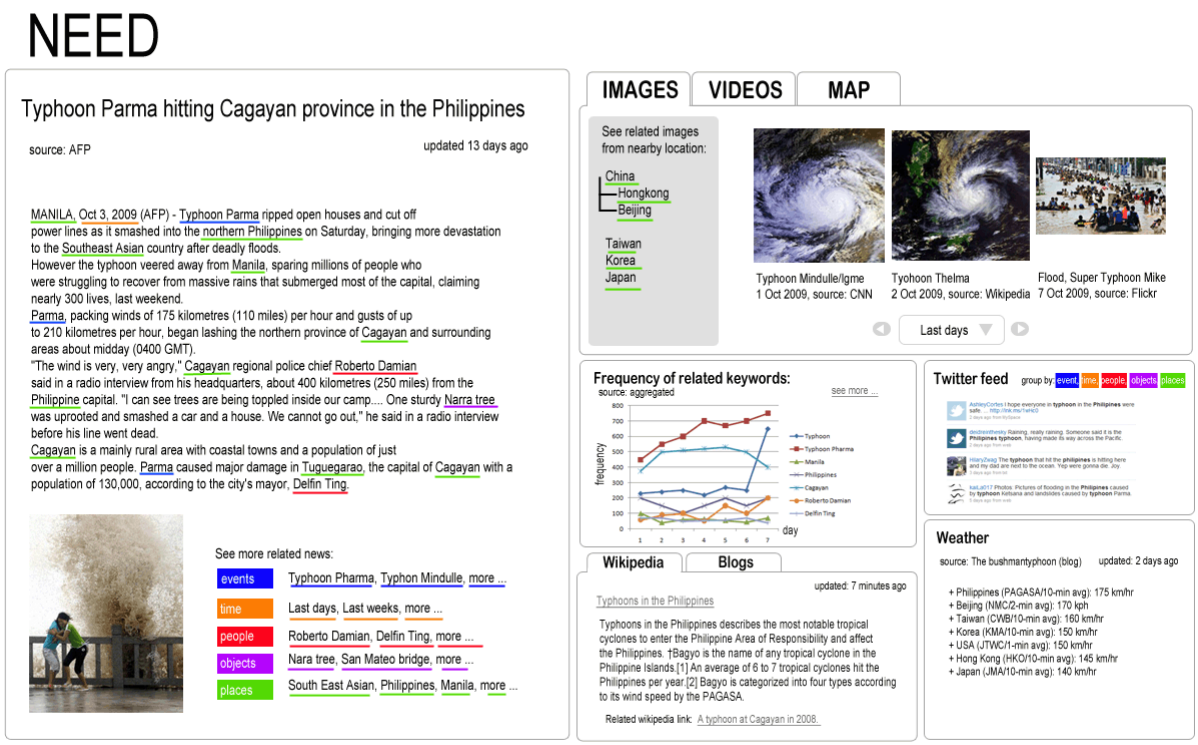


Figure 4: Sketch of a potential interface putting the Typhoon story in context

The Figure 4 shows a potential interface where different interface widgets visualize different information related to the Parma Typhoon from different sources. To promote transparency and trust, each widget clearly displays the source of the information as a link, clicking on that link would provide more provenance information about the data used. The functionality of the individual widgets shown includes:

- Central story: The left-side of the figure shows the central news item, with colour-coded matches created by the **NEED** event and named-entity detector.
- Space/time view: The top-right of the figure visualizes the typhoon Parma as a complex event, composed of a series of related sub-events that take place during a specific time range and in a specific space. The user can zoom in and out of in both dimensions, which allows both for selecting media items about events from different regions or time periods. It also allows for selecting media items with different topic granularity, e.g. a YouTube amateur video about a particular accident at a particular place versus a video summarizing the events related to Parma over the last week. The time/space interaction can be done in different modalities, e.g. the text-based interface shown here can be replaced by an interactive map or by a timeline view.

Figure 5: Map-based variant of the space-time view

The Figure 5 shows a map-based view, in this case showing the number of tweeds per location:

- Wikipedia/blog/Twitter views: The bottom widgets visualize other potentially related information aggregated from the Web. For example, the Twitter widget in the middle-right has grouped relevant tweets along the same event-related dimensions as used to colour-code the central story. The user selected one of the tweets, which provided a link to the wind-speed statistics on the bottom-right, showing that the Philippines suffer the highest wind speeds compared to other countries in the region.
- Aggregation view: Each widget above displays information from a specific source. We also envision use cases where users need aggregated statistics on data from many different sources. Here, the chart shows the frequency of related keywords in the last week. This not only provides insides in trends (e.g. the peek on day 7) but also suggests related keywords occurring often in related material.

These figures are just a sketch of potential interfaces that will be developed in **NEED** together with various user groups involved.

Scenario 2: Help professionals to determine the source and relate information

The media industry has largely written on the supposedly role of Twitter and other social sharing web sites in reporting the protests that have followed the controversial Iran Elections. Historians and sociologists will most likely analyze this phenomenon during the coming years, but at the heart of the journalist investigation work, there is the task of gathering pieces of information, of confirming their truthiness, of verifying all statements in order to deliver a fair report of the event. **NEED** will provide specific technologies and services that will support journalists in doing this daily activity.

Let us take the example of Mary, a video journalist for a TV Network, who is preparing a news report about the Iran elections. The video posted on the 20th of June 2009 on YouTube and titled "*They killed my sister. Is there any God to see?*" [16] was brought to her attention (Figure 6). The video reports allegedly about the death of a young woman on the streets of Teheran during the protest after the controversial Iran election.

[16] The video located at http://www.youtube.com/watch?v=RUPAdZ05wYA can be seen only by people over 18 due to particular violent content.

Figure 6: "They killed my sister. Is there any God to see?", http://www.youtube.com/watch?v=RUPAdZ05wYA

As a professional, Mary needs to accurately verify the source of this video before using the footage in a video report. She would also like to analyze how much this video has been viewed and shared across social networks, what the sociological impact is and how opinions differ across various countries. Consequently, Mary will perform the following tasks:

- Gathering evidence helping to source the event information:
  - The user that has uploaded this video on YouTube is R2D2Reports[17] and claims to live in Iran. The user "R2D2Reports" has created accounts on various social networks, for example on micro-blogging site with the IranPress[18] Twitter account. Mary also finds other videos of the same event on social web news site with this digg[19].
  - Mary performs cross-language search, using online translation service and finds other videos titled كشته شدن دختر جوان توسط لباس شخصی (a young girl being killed ... in Persian). Mary gets more information about the location, the Khosarvi street in Teheran (which can be confirmed with a mapping service on the internet), the context of the protest and the crime allegedly committed by Iranian forces.
  - Mary finds related video from other users such as hamedfrt[20], find more information about him querying specific person search engine such as 123people and ultimately re-build the social networks of these people in order to be able to get in touch with these users and get more information about the circumstances, people and location of the event.
- Gathering evidence to assess the trueness of the event:
  - Mary watch carefully the video to check any incoherence and if some particular building or monument appears. She submits random excerpts of the video to multimedia analysis processes in order to check if this video is not a montage of older footage material and if the buildings depicted are really located on the Khosarvi street in Teheran.

NEED services will be accessible from a common interface, run various analysis processes and offer mesh-up views of all the data acquired in order to support the journalist reporting task.

[17] http://www.youtube.com/user/R2D2Reports
[18] http://twitter.com/IranPress
[19] http://digg.com/world_news/Girl_gets_killed_by_Iranian_forces_at_protest
[20] http://www.youtube.com/user/hamedfrt

1.1.3 Project Goals

Our goal is to create an environment that facilitates end-users in seeing meaningful connections among individual news items (stories, photos, graphics, videos) through underlying knowledge of the descriptions of the items, their relationships and related background knowledge.

More specifically, existing metadata will be enhanced using named entities detection and knowledge bases. Multimedia content analysis will be performed to extract semantic concepts and to assess the provenance of the content. Real-time semantic processing of micro-blogging feeds together with natural language processing will support the clustering and organization of individual news items into meaningful events linked to appropriate background knowledge.

Enabling access to repositories containing any kind of media requires a system that can produce, collect, maintain and distribute media assets as well as aggregations of metadata associated with them. We will create metadata models to improve metadata interoperability along the entire news production chain. The underlying research challenges cover the two ends of the news workflow spectrum: how to model and represent semantic multimedia metadata along the news workflow and the consequences of this modelling at the user interface. At the same time, we will investigate the requirements the interface imposes on the modelling.

NEED will provide services for analyzing streams of video and textual news data, provide interoperability solutions in established news workflows and offer user interfaces providing access to heterogeneous sources of information, in multiple media types and in multiple languages centred on a macro-view of news events. By developing knowledge models for news events, **NEED** will provide event-centric navigation interfaces. These models will be populated with annotations of news items, using semi-automatic text processing and multimedia content analysis. Scheduled and breaking news events will be detected, analyzed, described formally and linked to existing background knowledge available on the web. News events sharing across social media sites will be analyzed both technologically and socially providing feedback to journalists, decision makers and citizens on how news information is consumed and perceived across cultures and communities.
For this open-knowledge approach to news workflow to work on an industrial scale in practice, all information components need to conform to open standards. Consortium members have close ties with different standards bodies to ensure that the real problems from existing communities are addressed within the project and that the solutions developed within the project can feed into the relevant standardisation activities. This will ensure broad uptake for establishing metadata workflows in the production and consumption chains throughout the news industry.

NEED will conduct its research and development following two case studies that will showcase the technologies developed and assessed during the project:

- The development of a web portal that will allow EU citizens to be kept up-to-date with environmental issues[21]. **NEED** services will be used to analyze any kind of news information related to some environmental issues (e.g. climate change and global warming, pollution, energy conservation and renewable energy, etc.) and offer interfaces structuring this information. A user group from the European Green Party will help in defining the themes to be covered, the level of information to report, the sources of information to monitor, etc. This user group will also evaluate the pertinence and accuracy of the news events presented to this portal (see Annex A).
- The development of a suite of integrated online services supporting a journalist in writing news reports. These services focus on real time analysis of information published on the web and include news events detection, news events provenance tracking, social network analysis, etc. Ultimately, **NEED** services will support journalists and decision makers providing them with tools performing strategic and technology forecasting.

The high-level goal is to deploy semantic metadata throughout the workflow in specific tools, while at the same time using the constraints of these tools to optimise the metadata provided all along the workflow.

[21] http://en.wikipedia.org/wiki/List_of_environmental_issues

More specifically, we will deploy metadata to improve the detection of overlapping feeds of information to enable clustering into events the user is looking for. By coupling metadata with controlled vocabularies, these can be used for finding relations among items in different repositories. By linking the metadata with background knowledge available on the Web, such as Wikipedia[22] or Britannica[23] for information on people, events, countries and general topics, we will provide end users with access to contextual information to help with understanding individual news items. The controlled vocabularies and background knowledge will in turn be used to "power" interfaces for end users to gain higher-level access to the repositories' contents. We will create an environment where media assets can be enriched with information based on usage patterns. We will finally develop interface design guidelines, based on the available metadata, which are useful for different types of applications that manage multimedia assets.

While we need to automatically extract metadata from textual and visual resources, our hypothesis is that sufficient amounts of metadata are already available, or can be extracted with existing techniques. The problem is that metadata is lost along the workflow due to interoperability problems and/or that metadata is not used in the end-user application.

Using general data modelling and processing techniques we will:

- ensure interoperability along the news workflow by integrating existing knowledge models such as the IPTC News Architecture (NAR) or the Core Ontology of Multimedia (COMM) and the Linked Open Descriptions of Events (LODE);
- find relations between news items coming from different sources, in different languages and on different media by using these models;
- cluster news items by using statistical analyses and integrate social media sources;
- rank the items of a cluster and find the representative of this cluster by analyzing the structure of the knowledge base.

In addition, we will use standard HCI methods (e.g. task analysis, user centred design) to ensure that the interface follows the maxims of conversation of Paul Grice[24] by:

- rendering an event and its related news items;
- providing the context needed to interpret a news item by providing links to related background knowledge at an appropriate level of detail;
- providing appropriate levels of information granularity, for example, a global view linking to more detailed information.

1.1.4 Meeting the challenges of the call

NEED particularly addresses four of the themes outlined in the work programme:

Targeted outcomes of FP7-ICT-2009-4.3	Contribution of NEED	Project objectives
a) Capturing tractable information: robust and performant technologies to acquire, analyse and categorise extremely large, rapidly evolving and potentially conflicting and incomplete amounts of information. These technologies will extract, correlate and integrate data from diverse sources and formats (multimedia and 3D content; heterogeneous databases; data streams from sensors and scientific equipment; social interactions and networked	**NEED** will provide an ontology-based infrastructure for representing news events, linked with appropriate existing background knowledge and exposed itself as formalized semantic web data linked to other datasets for being reused by third party applications.	• Deploy a knowledge infrastructure for news integration based on COMM and NAR • Enrich news metadata automatically using multimedia analysis tool-kits and text processing techniques • Detect and annotate

[22] http://www.wikipedia.org/
[23] http://www.britannica.com/nations
[24] http://en.wikipedia.org/wiki/Paul_Grice#Conversational_Maxims

appliances; information from business processes and software services) while tracing provenance, evaluating trust level and assessing reliability. The scalability, flexibility and performance of such methods and techniques will be demonstrated by rigorous empirical testing over large-scale testbeds.	**NEED** will mine in real time large amount of input streams being user generated content, micro-blogging and blogging feeds and cluster these sources into events linked to professional news content.	news events • Expose knowledge base of identified events as linked data
b) Delivering pertinent information: usable and customisable systems to improve the efficiency of the information lifecycle, starting from proactive diagnoses of information gaps and triggering goal-dependent search, acquisition, structuring and aggregation of relevant local, remote and streaming resources. Managing this information and making it actionable requires large-scale reasoning resulting in effective ranking, profiling and interpretation as well as versioning for time-dependent compliance and justification. Such systems will support the navigation, manipulation and consumption of digital information by means of adaptive user-information interactions based on the state of the art in the psychology of human perception and attention. The effectiveness of such systems will be validated with appropriately-sized groups or communities of representative users.	**NEED** will specifically design, implement, and evaluate novel interfaces for annotating, searching, browsing and presenting news events and related multimedia resources. Semi-automatic annotation interfaces for news will make use of existing web resources for suggesting and disambiguating metadata values.	• Evaluate the added value of the **NEED** technologies for journalists and lay users
d) Personal sphere: intuitive systems that help individuals secure, manage, visualise and interpret their personal information, attention trail and social history so as to enable the provision of personalised and context-dependent information from multiple sources and services. A specific requirement and design principle is that such systems preserve privacy and implement auditable information disclosure policies that are under user control and whose application can be verified at all times. Their usability and rate of uptake will be monitored by means of verifiable quantitative indicators.	**NEED** will address multilingualism and offer personalized interfaces for presenting news content. **NEED** will present contextualized information, adapted to various users, using different level of granularity	• Create personalized interfaces rendering the news items belonging to an event in various languages, and using various media • Provide real showcases for journalists and lay users to demonstrate innovation and usability
e) Impact and S&T leadership: networks and other initiatives designed to link technology suppliers, integrators and leading user organisations. These actions will help develop a common understanding, including vis-à-vis neighbouring disciplines, and ensure	A key contribution of **NEED** is that results will be developed in close cooperation with international standardisation bodies for web, multimedia and the news industry.	• Provide solutions combining Web and Metadata standards from W3C, IPTC, EBU and ISO • Support

proactive cross-fertilisation between EU projects and other relevant industrial and national activities. They will address barriers hindering a wider deployment of research results, work towards establishing or advancing widely recognised standards, reference architectures and benchmarks, and increase awareness of the potential of the technologies at stake within broader audiences	Dissemination and standardization work will be in close cooperation with upcoming W3C activities and address spatio-temporal addressing of audiovisual content fragments on the Web, metadata integration and media-specific metadata vocabulary.	standardization of spatio-temporal addressing of audiovisual content fragments on the Web in close cooperation with W3C

1.2 Progress beyond the state-of-the-art

Given the context of news production and consumption and its particular workflow, and assuming news content is produced in the form of audiovisual material and/or text, setting up a news environment as suggested by the project goals, involves the following areas of expertise:

- *Knowledge models for News Integration*: the ontology infrastructure for representing multimedia and news metadata, the various dimensions that composed an event, and a schema for attaching provenance information to all metadata [**Ontology Engineering field**];
- *Human Language Technology (HLT) for Event Detection*: natural language processing techniques for extracting named entities and more generally structured data from textual documents [**HLT field**];
- *Multimedia Analysis*: the interpretation of the rich media, including the ability to recognize identical or modified copies, the intelligent analysis required for topic and event detection, and for establishing relationships between news items [**Signal Processing and Computer Vision field**];
- *Semantic Interaction with News Content*: applications designed to make optimal use of available metadata, ontologies and background knowledge for providing semantic search interfaces and for generating multimedia news presentations [**Semantic Web User Interaction field**].
- *Social Media analysis*: the understanding of the social role of new communication tools, the generation of groups and communities and the creation of trust and engagement in online communities [**Social Media Analysis field**];
- *Large scale analysis of data*: trend detection from zillion of tweets [**Data Mining field**].

We discuss the current state of the art regarding these six research pillars. We then review a number of commercial initiatives related to the project objectives and we finally present the EU projects related to **NEED**.

1.2.1 Knowledge models for News Integration

For easing the exchange of news, the International Press Telecommunication Council (IPTC) has developed the **News Architecture**[25] (NAR), an XML-based language that provides a generic model for exchanging all kinds of newsworthy information and gives birth to a set of IPTC G2-Standards for news exchange. NewsML is a media-type agnostic news exchange format for general news and **NewsML-G2**[26] is its latest version. NewsML-G2 provides exchange formats for various media including textual news, articles, photos, graphics, audio and video (the News Item); a flexible mechanism for packaging news in a structured way (the Package Item); information about concepts, used for values in controlled vocabularies (the Concept Item); a format to exchange full controlled vocabularies as a single file (the Knowledge Item); and a wrapper around items to transmit them by any electronic means (the News Message). **EventsML-G2**[27] aims to be a comprehensive, flexible and extensible standard for conveying event information in a news industry environment and may be used for receiving or publishing all (or a subset of) facts about an event from the event organizer, adding information regarding the coverage of an event by a news provider, and storing facts about knowledgeable events in archives.

Even though NAR provides a general framework, some **interoperability problems** still occur. News are about the world and their metadata use numerous specific controlled vocabularies such as the IPTC News Codes or other thesaurus coming from industry that provide controlled terms further used as values of the metadata in the News Architecture. All these resources, however, come in XML-based formats and with no machine processable semantics which make their integration really difficult.

From the media point of view, the pictures taken by a journalist come with their EXIF[28] metadata. The **MPEG-7**[29] standard, formally named "Multimedia Content Description Interface", can be used to describe the structure and semantics of the multimedia content. The flexibility of MPEG-7 is based on allowing

[25] http://www.iptc.org/NAR/
[26] http://www.iptc.org/G2-Standards/newsml-g2.php
[27] http://www.iptc.org/EventsML/
[28] http://www.exif.org/
[29] http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

descriptions to be associated with arbitrary multimedia segments, at any level of granularity, using different levels of abstraction. The downside of the breadth targeted by MPEG-7 is its complexity and ambiguity: MPEG-7 XML Schemas define 1182 elements, 417 attributes and 377 complex types, making the standard difficult to manage. Moreover, the use of XML Schema implies that a large part of the semantics remains implicit. For example, very different syntactic variations may be used in multimedia descriptions with the same intended semantics, while remaining valid MPEG-7 descriptions. Given that the standard does not provide a formal semantics for these descriptions, this syntax variability causes serious **interoperability issues** for multimedia processing and exchange [van Ossenbruggen *et al.*, 2004], [Troncy and Carrive, 2004], [Nack *et al.*, 2005].

The profiles introduced by MPEG-7 and their possible formalization [Troncy *et al.*, 2006] address only a subset of the whole standard. For alleviating the lack of formal semantics in MPEG-7, four multimedia ontologies represented in OWL and covering the whole standard have been proposed [Hunter, 2001], [Tsinaraki *et al.*, 2004], [Garcia and Celma, 2005], [Arndt *et al.*, 2007] and further compared [Troncy *et al.*, 2007].

In parallel, the W3C Media Annotations Working Group[30], part of the Video in the Web Activity, aims to provide an ontology and an API designed to facilitate cross-community data integration of information related to media objects in the Web, such as video, audio and images. The goal is to create a simple ontology containing a concise set of terms extracted from such metadata standards. Additionally, mappings will be defined between a representative set of metadata standards and this ontology. As such, annotations of (news-related) media objects can be accessed in a uniform way, disregarding the original format of the metadata. The challenge is now how to reconcile multimedia ontologies with news ontologies and domain specific vocabularies for solving these interoperability issues along the news workflow.

In the context of online news items the origin of the news source is very important. In this aspect, knowledge on the creator or issuer of the news can allow to give more contextual information to the specific news item. Additionally, social links of the creator to other people or organizations are highly relevant when determining the context of a news item. OpenSocial is a set of common APIs for building social applications across many websites. The use of APIs allows to access the data, however to integrate this information according to the Linked Data initiative, a formal representation is needed. A first initiative has been made to see how the Social web can be integrated with the Semantic Web by the W3C Social Web Incubator Group[31].

Finally, journalists often stress the absolute need of representing the provenance of all type of information in order to trigger confidence regarding the truthiness of an event report. The Open Provenance Model is a model for provenance which meets the following requirements: (1) To allow provenance information to be exchanged between systems, by means of a compatibility layer based on a shared provenance model. (2) To allow developers to build and share tools that operate on such provenance model. (3) To define the model in a precise, technology-agnostic manner. (4) To support a digital representation of provenance for any ‘thing' whether produced by computer systems or not. (5) To define a core set of rules that identify the valid inferences that can be made on provenance graphs. Recently, the creators of the model have started the W3C Provenance Incubator Group[32] in which we participate and that will introduce the Open Provenance Model in the Semantic Web.

[30] http://www.w3.org/2008/WebVideo/Annotations/
[31] http://www.w3.org/2005/Incubator/socialweb/
[32] http://www.w3.org/2005/Incubator/prov/

These standards will be our starting point for designing a system that can produce, collect, maintain and distribute news media assets as well as aggregations of metadata associated with them. We will create knowledge models to improve metadata interoperability along the entire news chain production. Our approach will be to implement a knowledge infrastructure on top of: *i)* **COMM**[33], a Core Ontology for MultiMedia compatible with existing (semantic) web technologies based on both the MPEG-7 standard and the DOLCE foundational ontology [Arndt *et al.*, 2007]; *ii)* an OWL ontology of the News Architecture complemented by an appropriate Events Ontology such as **LODE**[34]; and *iii)* a number of SKOS thesauri either widely used in the news domain or specific to particular theme such as environmental issues.

1.2.2 Human Language Technology (HLT) for Event Detection

The considerable development of multimedia communication goes along with an exponentially increasing volume of textual information. Developing intelligent tools and methods, which give access to document content and extract relevant information, is more than ever a key issue for knowledge and information management.

- Information Extraction (IE) is the task of identifying relevant information from texts and presenting it in a structured form (Wilks, 1997). The IE field has been initiated by the DARPA's MUC program (Message Understanding Conference in 1987). MUC has originally defined IE as the task of (1) extracting specific, well-defined types of information from the text of homogeneous sets of documents in restricted domains and (2) filling pre-defined form slots or templates with the extracted information. Thus MUC has inspired a large amount of work in IE and has become a major reference in the text-mining field. Even as such, it is still a challenging task to build an efficient IE system with good recall (coverage) and precision (correctness) rates.
- IE Process: IE relies on document pre-processing and extraction rules (or extraction patterns) to identify and interpret the information to be extracted. The extraction rules specify the conditions that the pre-processed text must verify and how the relevant textual fragments can be interpreted to fill the forms. In a typical IE system, three processing steps can be identified (Hobbs et al., 1997; Cowie and Wilks, 2000):

1. text pre-processing, whose level ranges from mere text segmentation into sentences and sentences into tokens to a full linguistic analysis;
2. rule selection: the extraction rules are associated with triggers (e.g. keywords), the text is scanned to identify the triggering items and the corresponding rules are selected;
3. rule application, which checks the conditions of the selected rules and fills the forms according to the conclusions of the matching rules.

- IE and Ontologies : Ontology and IE are closely connected by a mutual contribution. The ontology is required for the IE interpreting process and IE provides methods for ontological knowledge acquisition.

An ontology is a description of conceptual knowledge organized in a computer based representation while information extraction (IE) is a method for analyzing texts expressing facts in natural language and extracting relevant pieces of information from these texts. IE and ontologies are involved in two main and related tasks:

- Ontology is used for Information Extraction: IE needs ontologies as part of the understanding process for extracting the relevant information;
- Information Extraction is used for populating and enhancing the ontology: texts are useful sources of knowledge to design and enrich ontologies.

These two tasks are combined in a cyclic process: ontologies are used for interpreting the text at the right level for IE to be efficient and IE extracts new knowledge from the text, to be integrated in the ontology (Nedellec and Nazarenko, 2005):

- Mapping an ontology with extraction results. The ontology and its knowledge base are used to store and exploit the information in a rigorous and constraining way. Formalisms such as OWL and RDF(S) are used to define the semantic representation of the knowledge along with the required quality for normalizing the instances. IE produces a set of semantic tags organized as a conceptual tree, mostly represented as a XML document. IE tends to extract a maximum of information without a constant worry about the normalization of the extracted data because the result is usually exploited by search engines and the formalism is thus less crucial.

[33] http://multimedia.semanticweb.org/COMM/
[34] http://linkedevents.org/ontology/

We will first compare Temis technology in terms of named entities detection and disambiguation with other deployed services such as OpenCalais[35], Zemanta[36] or GATE[37]. Further, we will develop unique technology for macro event detection. The goal of event extraction is identifying and tracking events and the entities that participate in them. The main purpose of event detection and tracking (TDT) is to detect, group, and organize news items reporting on the same event. Since an event is a reported occurrence at a specific time and place and the unavoidable consequences, TDT can benefit from an explicit use of time and place information.

1.2.3 Multimedia Analysis for News Content

Integration of multimedia and semantic web technologies for describing, extracting and retrieving semantic information from multimedia has been identified as one of the most promising approaches to narrow down the semantic gap. Multimedia news processing starts with low-level features extraction and semantic concept detection in order to structure the content and perform quotes detection. Mining techniques on existing metadata and background knowledge can be further used to automatically detect and annotate *events*, which cluster several news items.

Shot boundary Detection in Audio Visual News data. Algorithms for shot boundary detection can broadly be classified into two major groups, depending on whether the operations are performed in the pixel domain, or whether they rely directly on compressed-domain features. Although most of the proposed techniques for the detection of gradual changes work on uncompressed video, compressed-domain algorithms are gaining importance. [Zang *et al.*, 1993] presented a compressed domain approach based on DC coefficients and motion vectors. As the H.264/AVC video coding standard performs significantly better than any prior standard in terms of coding efficiency, it can be expected that a significant amount of future video content will be encoded in this format. As this specification contains a number of new coding tools, several new shot boundary detection algorithm are introduced which anticipate on these new coding tools [Liu *et al.*, 2004], [Kim *et al.*, 2005], [Zeng and Gao, 2005].

Key frame extraction from Audio Visual News data. Key frames generally represent the content of one shot and are selected according to an analysis method that optimizes the semantic coverage of the video content. Based on these key frames, content analysis algorithms can extract semantic information. Many of the systems described in the literature use a constant number of key frames for each detected shot. Often, the first, middle, or last frame is selected as representative. Other algorithms adjust the number of key frames to the content of the shot. [Zhang *et al.*, 1993] propose to select the first frame of a new shot as key frame. Consecutive frames are then compared against this candidate key frame, applying a difference metric. When a frame significantly differs from the candidate, this frame is added to the set of key frames as well. Other techniques select the key frame by determining the location where frame differences are minimal. Motion information or frames closest to the cluster center are often used for this purpose [Wolf, 1996].

Extraction of semantic meaningful information from Content-Based Image Retrieval (CBIR). In typical CBIR systems, the visual contents of the image are extracted and described by multi-dimensional feature vectors. To retrieve images, example images or sketched figures are provided to the system. The system then changes these examples into its internal representation of feature vectors. The similarities/distances between the feature vectors of the query example or sketch and those of the images in the database are then calculated and retrieval is performed with the aid of an indexing scheme. Visual content can be very generally described with color [Faloutsos *et al.*, 1994], [Mehrotra *et al.*, 1997], [Rui *et al.*, 1998], [Smith and Chang, 1998], texture [Haralick, 1979], [Martens *et al.*, 2007], shape [Lowe, 1999], spatial relationship or domain specific like human faces [Viola and Jones, 2002].

Web and Text Mining techniques in blogs and news stories. The European Commission Joint Research Centre's language technology activities (JRC) in Web Mining have recently produced the NewsExplorer[38]. A number of text gathering (retrieval), analysis and visualisation tools have been developed with a focus on multilingual and multi-document information aggregation, and on tools to provide cross-lingual information

[35] http://www.opencalais.com/
[36] http://www.zemanta.com/
[37] http://gate.ac.uk/
[38] http://press.jrc.it/NewsExplorer/home/en/latest.html

access. These text analysis tools have been integrated with the news gathering engine Europe Media Monitor (EMM) to produce several complex, high-level applications. The current tool set consists of three main components with the following functionality: multilingual and cross-lingual retrieval of potentially user-relevant documents, analysis of documents and extraction of different information aspects from these documents plus language-neutral representation of this information where possible, and finally visualisation of all this content.

BLEWS is a framework from Microsoft Research that uses blogs to provide context for news articles. While typical news-aggregation sites cluster news stories according to topic, they leave the reader without information about which stories figure prominently in political discourse. BLEWS uses political blogs to categorize news stories according to their reception in the conservative and liberal blogospheres. BLEWS also offers a "see the view from the other side" functionality, enabling a reader to compare different views on the same story from different sides of the political spectrum. BLEWS achieves this goal by digesting and analyzing a real-time feed of political-blog posts provided by the Live Labs Social Media platform, adding both link analysis and text analysis of the blog posts [Gamon *et al.*, 2008].

Retrospective news event detection (RED) is defined as the discovery of previously unidentified events in historical news corpus. Although both the contents and time information of news articles are helpful to RED, most researches focus on the utilization of the contents of news articles. Few research works have been carried out on finding better usages of time information. In that sense, Microsoft Research is doing some explorations on both directions based on the following two characteristics of news articles. On the one hand, news articles are always aroused by events; on the other hand, similar articles reporting the same event often redundantly appear on many news sources [Li *et al.*, 2005].

We will build on these results to implement multimedia analysis tool-kits tailored to the news domain. **NEED** will reuse the current state of the art techniques in multimedia content analysis and will go further in three directions:

- *Detection of visually similar news items*:

Recently, thanks to Web 2.0 applications and social games such as Wikipedia, Flickr, YouTube and ESP games, users are able to contribute tags about various media. They are therefore providing rich and meaningful additional information about multimedia documents which can be used as additional knowledge about the content and eventually its context. It is clear from looking at the success of Flickr that users are willing to provide these semantic descriptions through manual annotations. The difficulty when using such information is that it will inherently contain some noise which may confuse the classifier. However, through the large number of annotation provided by user and the presence of duplicate and near duplicate documents is may be possible to discard the outliers. There are been very limited efforts so far concerning the use of such collective information in order to infer new knowledge about both content and context. The methodologies proposed in [Lu et al., 2008] and [Wang et al., 2006] focus on image content, while we propose to study how additional information collected from numerous web sources can contribute to the generation of new, enriched, and continuously evolving concept models. To this end, we will investigate the use of data mining approaches rather than more conventional methods based on learning from manually annotated ground truth. Furthermore, current approaches are tackling each modality (text, visual, audio) separately. Either combining the various features before classification (early fusion) or at a later stage after individual classification have been performed (late fusion). We believe the combination should not be made at a specific point in time but rather used to infer additional context information and therefore ensure higher accuracy and richer concept representation.

- *Quote detection and alignment* :

Thanks to the multimodal representation and methodology employed within **NEED** the alignment between news quotes and multimedia document gathered through web search will achieve level of performance yet unattained. The task consists in searching predefined web services (YouTube, Flickr, CNN, Wikipedia…) for relevant information according to a given "news quote" or eventually multiple quotes. As the returned list of multimedia items will contain both relevant and irrelevant documents a filtering stage should be applied in order to prune totally irrelevant documents, and rank the remaining ones according to some automatically computed confidence level.

- *Assess the provenance of news item based on visual clues*:

When searching and collecting multimedia information on the web for enhancing news services, the origin (the source) and the relevance (the spatio-temporal occurrence of the event) should be validated. Concerning

the source, it is obvious that multimedia items originating from the CNN archive are more likely to be genuine than those downloaded from YouTube. YouTube (and similarly Wikipedia) documents will have to be further checked in order to assess their authenticity. A set of multimodal algorithms will be researched and developed in order to detect visual clues allowing to verify with a high level of confidence the temporal and spatial origin of a document. This will consist in searching for visual cues strongly correlating the documents originating from "trusted" sources with the ones from socially contributed repositories.

1.2.4 Semantic Interaction with News Content

The presence of semantic annotations allows browsing interfaces to use them to tailor selections to be presented to the user. The presentation of news can be thus personalized and include some background knowledge so that the user can fully understand an event. One of the key issues is then the use of an appropriate time model for presenting temporal information at the right level of granularity.

The **TERQAS** (Time and Event Recognition for Question Answering Systems) project has provided TimeML[39], a specification language for events and temporal expressions. Based on this language, the **TIDES** (Translingual Information Detection, Extraction, and Summarization) project has produced Temporal Annotation Guidelines[40] which not only offers a way to mark inline dates but also "directions in time".

Microsoft Research has also analyzed timeline visualizations for displaying the results of queries on an index. Results of searches are presented with an overview and detailed timeline visualization. A summary view shows the distribution of search hits over time, and a detailed view allows for inspection of individual search results. In a user study, [Ringel *et al.*, 2003] explore the value of extending a basic time view by adding public landmarks (holidays and important news events) and personal landmarks (photos and important calendar events) in the hope that this added context will help people in locating the target of their search.

In the cultural heritage domain, the **MuseumFinland** portal [Hyvönen *et al.*, 2005] allows users to explore using pre-specified facets (or characteristics) of artefacts, allowing them to select subsets, without being confronted with queries that return zero results. The **/facet** (pronounced "slashfacet") browser [Hildebrand *et al.*, 2006] provides similar end-user browsing techniques, but enables any Semantic Web collection to be browsable with the tools. The **CHIP** project [Aroyo *et al.*, 2007] takes the use of semantic annotations further and allows users not only to browse the collection, but give their opinion on each artwork and use their profile to refine the artworks chosen to be presented to them.

Pure schema-less Web applications and truly generic Linked Open Data browsers

In traditional application development, the design of the user interface of the application and the design of the underlying data structures are typically tightly linked, as both are based on the functionality provided by that single application. The data structures are designed in a way that they optimally reflect the needs of the application's functionality, and the UI is designed to provide users access to that same functionality. That balance has slowly changed in the last decades. Especially on the Web, where the ability to exchange data across different applications of different vendors has been a key success factor, the days that a single application dictates the structure of a data or file format are over: many data schemata and serialization languages are designed by standardization committees, which application developers have to implement. In recent years, we have seen the introduction of "schema-less" Web applications. These applications that do not commit to a fixed set of schemata, but only commit to a set of meta-languages used to define these schemata. Parsing of, and reasoning with, this data can now be done with mature off-the-shelf tools and libraries. Unfortunately, it proved much more difficult to design effective user interfaces for data of which the schema is not known at the time the interface is developed. Often, these interfaces provide little more than a very generic interface based on a metaphor that matches the underlying meta-datamodel: expandable trees for XML applications, graph and triple-based visualizations for RDF applications, and various class/instance viewers for RDFS and OWL.

[39] http://timeml.org/site/index.html
[40] http://fofoca.mitre.org/annotation_guidelines/timex2_annotation_guidelines.html

For example, the Tabulator by Berners-Lee et al. provides an interface that allows end-users to browse linked data published on the Web. It shows the raw triples related to a resource as hyperlinks, allowing the user to surf from an RDF resource on one location to a semantically related resource on another location. While the idea is similar to surfing from one HTML page to another in a normal browser, there is a crucial difference: where HTML pages typically contain information that is written, packaged and styled for human consumption, the raw data contained in most RDF data sources is not. This makes the number of realistic end-users tasks that can be supported by Tabulator extremely limited. Because of their goal to be truly generic, applications like tabulator can make no assumptions about the data or about the end user's task. As a result, the same limitations apply to other generic linked data browsers such as the DISCO browser by Bizer & Gauß.

Hybrid Web applications based on light schema mapping

While we expect **NEED** data to be highly heterogeneous, the requirements of **NEED's** users seem to call for a hybrid, and more flexible approach than that of the truly generic interface to RDF data. Applications such as MIT's Exhibit and MultimediaN's ClioPatria can also be applied to arbitrary RDF data sets, but focus on ways in which the user interface can be enriched nevertheless. In most realistic data sets, there are often extra small pieces of information that are known about the data that can be exploited to improve the interface. For example, if only the application could be told how to recognize geographical locations in the data set, this tiny bit of information already allows map mash-ups and other ways to enrich the interface beyond a purely generic triple interface. Other ways to enrich the UI are to exploit standard RDFS techniques to map domain-specific schema elements to elements from more common, cross-domain schemata such as Dublin Core, SKOS, FOAF, etc. If the application's UI supports effective interaction with data conforming to the these common standards, all the domain expert needs to do is provide a light-weight mapping from the schema elements in her domain to the more common elements known to the application. This technique has already proved to be effective for dealing with domain-specific variants of schema elements that tend to recur in many heterogeneous data sets across various domains, including domain-specific schemata for locations, time and dates, persons, physical objects, financial data and thesaurus concepts. By exploiting such light-weight schema mappings, applications such as **NEED** can build high-level user interfaces on top of very heterogeneous data sets.

Event-based interaction with news

The observations that persons, objects, locations, dates, etc. tend to be recurring elements in many domains can even be taken one step further. In many disciplines, from the humanities to the exact sciences, researchers have come to the conclusion that the notion of an "event" can, on the one hand, be seen as yet another element that tends to recur in many data sets across several domains. On the other hand, however, the event has the potential to meaningfully relate previously isolated mentions of persons, objects, locations and dates. As such, a sufficiently generic, but not overly abstract definition of event allows users to express information needs that are both sufficiently abstract to be applicable over a wide range of domains and data sources, yet sufficiently concrete to allow be automatically translated into the query languages that the underlying applications can execute effectively.

The K-Space demonstrator can be seen as an early example of this principle being applied to the news domain. For example, named-entity recognition applied to news articles may detect that several of them somehow "mention" the person Angela Merkel, the place Heiligendamm and the date June 2007. The notion of event however, will add these co-occurrences in the text in a meaningful context: these articles are all about the 33rd G8 summit that was attended by Merkel in her role of chair and taking place in Heiligendamm, in early June 2007. Such information can be exploited to improve the user interface even further. It allows for smart recommendations and navigation suggestions, it allows for automatic compositions of queries to find information of other events, it allows events to be compared on different dimensions, it allows news articles published in the early days of the summit that speculate over the possible outcomes to be enriched with forward links to the actual results, etc. etc.

Many of the interfaces discussed above are purely data-driven, often focusing solely on giving power-users direct access to the raw data. While we agree that this approach in an important part of application settings like those of **NEED**, we believe a successful interface also needs to take the user's task into account, and provide selection, ranking and grouping and visualization of the data that suits the current user's tasks. The semantic interfaces developed by the **NEED** consortium will extend the state of the art by focusing on the latter aspect: effective visualizations of heterogeneous Web data that convey the richness of the data in a way that matches the user's task. We believe that for all professional applications working with data aggregated from the web conveying provenance context will be key. In Web 1.0, and even in many Web 2.0 applications, users can base their level of trust on the reputation of the website where they found the information. With Web 3.0, machines will aggregate information from many sources, where some will be more reputable than others. Being able to the trace the origin of a unit of information, and being able to convey that trace effectively to the professional user will be a key contribution of the **NEED** project.

1.2.5 Social Media Analysis

Social media such as Facebook, YouTube, Flickr or Twitter reduce the technical complexity of Internet use (Harrison & Barthel, 2009; Cormode & Krishnamurthy, 2008; Depauw, 2008), turning it into an instrument for the mass. They make it easier for users to consume (read, listen, watch, download, search and buy), create (personalize, aggregate and contribute), share (publish, upload), facilitate (tag, recommend, filter, subscribe to channels and items through RSS) and communicate (send messages, post comments, rate and chat) online (Slot & Frissen, 2008). These new possibilities bring about significant social and cultural changes.

Online digital information has become ubiquitous and accounts for a major part of the economic and cultural activities in western society as the removal of the physical constraints on information production has made creativity and the economics of information the core structuring facts in the new networked information economy (Benkler, 2006). Chad Hurley (CEO YouTube) notes that: "Every minute 13 hours of video is uploaded to the site – equivalent of Hollywood releasing 57.000 movies a week". These societal and technological changes are the focus of (new) media studies and the study of computer mediated communication (CMC) in specific.

In recent years, the discourse on Internet use has been dominated by theories on *'user participation'*. (O'Reilly, 2005; Andersen, 2007; Aguiton & Cardon, 2007). Social media enable and strongly encourage user-generated content (UGC). UGC is defined as content made publicly available on the Internet, reflecting creative effort. UGC is created outside of professional routines and practices (OECD, 2007) and takes various shapes. These media can be taken up by regular users, wishing to express their creativity or opinions, but are also by social movements, pression groups or political parties given that these technologies provide a means for easy and relatively cheap campaigning. Moreover, increasingly journalists and reporters rely on these kinds of internet sources functioning as real-time information for the coverage of the topics discussed.

The *creation of online (virtual communities)* as "(…) social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace" (Rheingold, 1993: 5). Armstrong and Hagel (1996) distinguished four generic types of virtual communities: communities of transaction, communities of fantasy, communities of relationships and communities of interest (Gupta & Kim, 2004, p. 2682). Communities of transaction facilitate the buying and selling of products and/or services. Communities of fantasy create new environments and stories. Communities of relationships crystallize around particular life events. Finally, communities of interest join people working and communicating intensively with each other on a specific topic (e.g. sports, music, lifestyle, politics, environment etc.).

The *understanding of group or team processes* (e.g. collective knowledge management and knowledge sharin) and the design, creation and evaluation of technological systems that support group interaction. In this field, often called CSCW, researchers endeavour to understand the nature of collaborative work with the objective of creating efficient computer-based technologies that support group or team work. Within this field 'Enterprise 2.0' tries to use the emergent social software platforms within companies, or between

companies and their partners or customers (McAfee, 2006a). Andrew McAfee (2006b) defined Enterprise2.0 using the mnemonic SLATES: **S**earch (the discoverability of information); **L**inks (using URIs to create interconnections between information); **A**uthorship (providing authoring tools to all levels); **T**ags (bottom-up organisation of information); **E**xtensions ('extending' knowledge through data harvesting & mining); **S**ignals (making information consumption more efficient by providing (social) signals)

The *trustworthiness of information and the creation of trust and engagement in online communities*. Techniques such as reputation management and collaborative filtering, gatekeeping (Goldberg, Nichols, Oki, & Terry, 1992) are the focal points of attention here. However, internet not only provides the hugest information storage ever, but is also an information retrieval system. On the Internet, people access content via the evaluations that other people have already left on that content. From search engines to reputational systems, Internet is developing as a giant ranking system in which people's advice and preferences leave tracks that orient other people's advice and preferences. Often evaluation precedes or even replaces information. Here, the social information retrieval (SIR) approach is very relevant. SIR assists people in obtaining relevant information by harnessing the knowledge or experience of colleagues, friends, peers, and others. SIR relies on social network analysis and the use of subjective relevance judgments such as tags, annotations, ratings and evaluations.

The *exponential growth of (online) information and how it 'overloads' human users* (information overload, see e.g. Jones, Ravid, & Rafaeli, 2004) and the findability of the relevant information in this context (information paradox). However, as Clay Shirky points out, the problem is not the enormous amount of information which is now available, but that we don't have proper filters for it. One way to cope with this is through the use of folksonomies. The term folksonomy, generally attributed to Thomas Vander Wal (Smith, 2004), refers to online tagging systems intended to make information increasingly easy to search and navigate over time. Folksonomies move a user from a 'binary' in-or-out classification system to an 'analogue' one, only requiring a conceptual association with a resource (Shirky, 2005).

Although the above mentioned perspectives are very useful in the study of the 'user' in a new media or web 2.0 context, state-of-the art social science and Computer Mediated Communication research often stops there. The computer or the Internet is determined to be, both in name and in function, an instrument or medium through which human interlocutors exchange information and interact. However, this form of instrumental transparency is interrupted and resisted by the mechanisms of today's Internet. Increasingly, social media actively participate in communicative exchanges as a kind of additional agent and/or (inter)active co-conspirator (e.g. they approach the user with recommendations bases on 'read-wear' or 'exhaust data'). Our research aim is to go beyond the current state-of-the-art by exploring the idea of assigning the Internet the position of a social actor with whom one communicates and interacts (Gunkel, 2009). This approach should result in a better understanding of the user's current requirements and should optimise their information ontology in order to make an efficient aggregation, selection and retrieval of information possible.

1.2.6 Trend Detection from Micro-Blogging feeds

The objective of Emerging Trend Detection (ETD) is to provide an automated alert when new developments are happening in a specific area of interest. An ETD application takes as input a collection of textual data and identifies topic areas that are either novel or are growing in importance within the corpus. Current applications in ETD fall generally into two categories "fully automatic" and semi-automatic. The fully-automatic systems take in a corpus and develop a list of emerging Topics. A human reviewer then peruses these topics and the supporting evidence found by the system to determine which are truly emerging trends. These systems often include a visual component that allows the user to track the topic in an intuitive manner. Semi-automatic systems rely on user input as a first step in detecting an emerging trend. These systems then provide the user with evidence that indicates whether the input topic is truly emerging usually in the form of user-friendly reports and screens that summarize the evidence available on the topic. Most of the existing ETD systems employ some text mining techniques for detecting topics, then monitor these topics over time and define whether these topics are emerging or not (either applying machine learning techniques or relying on visualization techniques). Examples of trend detection using classical information retrieval or text mining methods, are Technology Opportunities Analysis System (TOAS), Constructive collaborative Inquiry-based Multimedia E-Learning (CIMEL), ThemeRiver, Envision, TimeMines, Hierarchical Distributed Dynamic

Indexing (HDDI), PatentMiner, Emerging Topic Tracking System (ETTS), Sequence-Based Self-Organizing Map (SBSOM). Trend detection system based on Web 2.0 services are emerging based on Hadoop and MapReduce to supports data-intensive distributed application.

The objective of Emerging Trend Detection (ETD) is to provide an automated alert when new developments are happening in a specific area of interest. An ETD application takes as input a collection of textual data and identifies topic areas that are either novel or are growing in importance within the corpus.

Current applications in ETD fall generally into two categories: fully automatic and semi-automatic. The fully-automatic systems take in a corpus and develop a list of emerging Topics. A human reviewer then peruses these topics and the supporting evidence found by the system to determine which are truly emerging trends. These systems often include a visual component that allows the user to track the topic in an intuitive manner. Semi-automatic systems rely on user input as a first step in detecting an emerging trend. These systems then provide the user with evidence that indicates whether the input topic is truly emerging usually in the form of user-friendly reports and screens that summarize the evidence available on the topic.

At the moment there are few evidences in the literature for trend detection system using Semantic Web technology. In **NEED**, we intend to leverage the knowledge infrastructure developed within WP2, to identify relevant topics. In our approach, the semantic tags associated to each document by the Event detection and annotation tool, developed in WP3, together with its timestamp, will provide the basis for trend detection. This approach will automatically filter the resulting trends from noise. Techniques for identifying real emerging trends among all discovered trends will also be investigated.

1.2.7 Commercial Tools

A number of commercial tools and web sites provide more advanced and flexible access to news stories. However, they are often restricted to the textual media only while the vision of the project is to consider that news will be more and more multimedia. Furthermore, the added value of these tools is often to aggregate more sources or languages resulting in too much information and overly complex interfaces.

Autonomy Virage[41] is world leader in Rich Media Management software and Video Analytics that automatically captures, encodes and indexes television, video and audio content. By automatically generating a comprehensive range of metadata, including keyframes, face recognition, speaker information and a full transcript of the audio stream, Virage ensures that rich media is immediately fully searchable and accessible by any user. It uses Bayesian statistics in order to find out the context of an item. Although Virage is a renowned product within the broadcast world, their visual tools are not State-of-the-Art and their searching is purely text-based starting from the transcripts. We will provide a richer suite of contextual and content analysis techniques for enriching the semantic descriptions of news photos and videos. In particular, we will develop analysis algorithms beyond state-of-the-art for structuring videos and detecting a chosen set of semantic entities useful for event detection and organization.

SiloBreaker[42] is an online search service for news and current events that delivers meaning and relevance beyond traditional search and aggregation engines, which is communicated via several exploratory information graphics. They support the user's information need more closely, but just add more sources of individual news items, which leads to overly complex interfaces. Throughout our knowledge infrastructure, we aim to develop interface design guidelines based on the real metadata requirements.

Twine[43] applies semantic analysis techniques to multimedia content (notes, videos, photos and contacts) and creates tags for each resource. The tags match up to concepts that Twine's algorithms associate with each piece of content, regardless of whether that concept is specifically mentioned in the content being tagged. Twine promises to be the first mainstream semantic web application to hit the market but full-featured social bookmarking is information-dense so adding all the semantic features and recommendations from Twine turns its information architecture and user experience into huge challenges. Twine's user experience is confusing. It's hard to keep track of all the levels and types of information available and site navigation is

[41] http://www.virage.com/content/home/index.en.html
[42] http://www.silobreaker.com/
[43] http://www.twine.com/

dizzying while **NEED** promises to develop clear interfaces displaying the right amount of information to the user.

Newstin[44] provides multilingual news access and allows searching, browsing or reading news by topics and not just by keywords. The topic structure allows switching languages and having foreign-language stories translated into a native tongue. The StarTreeNavigator is a novel way to find clustered items, but the hierarchy behind it is fixed and does not follow IPTC subject codes. The main limitation of Newstin is that the system deals only with textual news stories while we aim to smoothly integrate all media types and particularly video news.

NewsAtSeven[45] is an automatic system that builds personalized news shows. After selecting a news stories, the system finds relevant images, videos or external opinions and generates a virtual news bulletin read by an animated character while providing the original sources of the information. The avatar makes the presentation of the news very appealing. However, the system cannot link together several stories on the same topic (for example coming from several sources), nor present an event globally or provide contextual background knowledge for understanding the news.

ClearForest Gnosis[46] can extract particular information such as named entities from textual web resources and highlight them while browsing the web. The named entities extracted remain pure strings as they are not linked to any knowledge base while we want to enrich existing semantic web datasets and contribute to the linked data community. Furthermore, we will extract relationships between the entities.

YouTube has proposed the **Warp**[47] interface to display video clips under animated nucleus satellite graphs. The bubbles are linked together when they share the same subject or have some semantic proximity. Albeit this user interface looks nifty at first, it is difficult to understand why videos are linked together, and the interface looks sometimes very confusing if we get into a network that is very dense. The timeline view representation has been also largely studied in various contexts. **Google Experimental Labs**[48] proposes to see search results on a timeline. Google's technology extracts semantic concepts such as key dates, locations, and measurements to present the information along several dimensions.

None of the above websites and tools really solves the issue of the individual user being overwhelmed by too many individual and disconnected pieces of information. Some of them, however, present unique and renewed ways of presenting information to end-users. We will take these interface ideas to a higher level by really addressing all media (text, image, graphics and video) and by incorporating the temporal dimension of the news. We will provide the news event granularity the user is looking for with appropriate contextual information and background knowledge.

1.2.8 Related EU Projects

A number of EU projects funded under the Framework Programmes FP6[49] and FP7[50] are related to **NEED**. We show below how we will differentiate from these projects while building on their main results.

NEWS[51] developed an automated multi-lingual textual news classification and annotation engine that is able to categorize information (using the IPTC subject taxonomy) and extract named entities in English, Italian and Spanish. Temis now offers the same level of functionalities in 16 languages as a commercial tool, and this system will soon be in production in French and English as part of the AFP editorial system. Furthermore, the project never achieves to build an event extraction engine while we will use these annotations for inferring news events from content annotations. The NEWS project also provided a prototype

[44] http://www.newstin.com
[45] http://www.newsatseven.com
[46] https://addons.mozilla.org/en-US/firefox/addon/3999
[47] http://www.youtube.com/testtube
[48] http://www.google.com/experimental/
[49] http://cordis.europa.eu/ist/kct/fp6-projects-alpha.htm
[50] http://cordis.europa.eu/ist/kct/fp7_projects.htm and
http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-call1_en.html
[51] http://www.news-project.com/

ontology model applicable to the news domain [Fernández *et al.*, 2007], [Zap *et al.*, 2005]. We will build on the results of this work for creating a reference OWL description of the news domain, based on the NewsML-G2 IPTC standard. We will also tackle multiple types of media, in addition to text, as it is our belief that video will gradually become the dominant medium for news content.

Citizen Media[52] enables lay users to consume, author and publish their content as part of a networked audiovisual system. The project focuses on automatic analysis of visual information and on scalability issues since the system has to be able to handle a massive amount of user-generated content in different formats in real time, and annotate and store this content in huge databases in order to better reuse all these pieces of user-generated content. We will build on the experiences with the interfaces developed for lay users but in addition will target professional users from news agencies and broadcasters as well as independent journalists. We will also emphasize the role of semantic metadata for solving interoperability problems and empowering end-user interfaces.

MESH[53] aims to extract, compare and combine content from multiple multimedia news sources, automatically create advanced personalized multimedia summaries, syndicate summaries and content based on the extracted semantics, and provide end users with a multimedia mesh news navigation system. While the project has made progress across this broad set of goals, it focuses mainly on news distribution on mobiles. We concentrate, however, on the use of the underlying technological semantic infrastructure to reduce the amount of information exposed to the user in a simplified interface.

SEMEDIA[54] overall objective is to develop a collection of audiovisual search tools that are user driven, preserve metadata along the chain, and are generic enough to be applicable to different fields (e.g. broadcasting production, cinema postproduction or social web). We will also stress on the preservation of the metadata along the workflow and even solve the current interoperability problems using a knowledge infrastructure based on current standards and practices in the media industry. Given the specificity of the news domain and the existence of the IPTC subject codes, we will concentrate on the macro events detection problem together with the implementation of clustering and ranking algorithms for news items.

PENG[55] (Personalized News Content Programming) aims at providing news professionals with an interactive and personalized tool for multimedia news gathering and delivery. This was achieved by developing a flexible prototype for a personalized filtering, retrieval and composition of multimedia news.
The personalization is obtained by "tuning" the contribution of the distinct content types and sources of information by associating a trust score to each information source. However, the hierarchical organization and categorization of the news remains fuzzy while we will find automatically relationships between news items coming from different sources, in different languages and on different media by using our knowledge infrastructure and multimedia analysis tool-kits. This will result in an intelligent and intuitive clustering and ranking of news items.

PAPYRUS[56] (Cultural and Historical Digital Libraries dynamically mined from News Archives) aims at creating a cross-discipline digital library engine that allows for drawing content from one domain and making it available and understandable to the users of another. PAPYRUS starts from the historical perspective (within existing digital libraries) and tries to relate that back with legacy content from News Providers. In that sense, the definition of an event pre-exists and even if it can be controversial, some facts are unambiguous (why it happened; what were the consequences; and how it may have been avoided). We will rather take the point of view of news providers for defining formally the notion of an event and will have to deal with the dynamicity of the news where events are not necessarily expected. We will also build on the news ontology developed by AFP and CINECA, both partners of Papyrus.

[52] http://www.ist-citizenmedia.org/
[53] http://www.mesh-ip.eu/?Page=Project
[54] http://www.semedia.org/
[55] http://www.peng-project.org/
[56] http://www.ict-papyrus.eu/

aceMedia[57] aims to create a framework for combining advances in knowledge, semantics and multimedia processing technologies. The main output of the project is the aceToolbox, a content analysis platform for low-level image feature extraction. However, the project has not considered complementary resources and existing background knowledge for adding semantic information, while it will be the primary focus of **NEED** to perform events detection and annotation.

The **K-Space**[58] Network of Excellence brings together the content analysis and Semantic Web communities and aims at narrowing the semantic gap with semantic inference for automatic annotation and retrieval of multimedia content. As a NoE, the focus is on research resource creation via multi-partner collaborative activities, including PhD and researchers exchange. A number of knowledge-based multimedia content analysis technologies have been made available on which we will build **NEED**. Specific examples include automatic and semi-automatic object segmentation and concept detectors.

None of the above projects addresses the problems of finding how different news items are related to each other and of establishing explicit relationships between different sources for the same event in order to help the user to form his/her own opinion. Moreover these projects never achieved to build an event extraction engine giving an overview of individual pieces of information and making sense of individual news items by providing links to background knowledge.

[57] http://www.acemedia.org/
[58] http://www.k-space.eu/

1.3 Concept and objectives

1.3.1 Overall strategy of the work plan

NEED aims to create an environment that facilitates end-users in seeing meaningful connections among individual news items (stories, photos, graphics, videos) through underlying knowledge of the descriptions of the items, their relationships and related background knowledge, centred on the notion of scheduled or breaking news events.

NEED will address *metadata interoperability* along the multimedia news workflow by designing, leveraging and integrating relevant multimedia and news ontologies. The resulting knowledge infrastructure will be thus ontology-mediated and expressed in knowledge representation web-based standards (OWL, RDF, SKOS). Several initiatives have proposed a way to represent news events in a structured way. The latest is the new IPTC working draft EventsML-G2. This standard relies on the News Architecture (NAR), an IPTC framework for the management and description of news and associated information. NewsML-G2 is also based on this architecture and can be further linked with an event model. Both models make use of multiple (and sometimes overlapping) controlled vocabularies, maintained by different authorities. Finally, other work on bridging W3C and MPEG-7 related standards for the description of multimedia content has resulted in the development of a core ontology for multimedia (COMM) [Arndt *et al.*, 2007]. We will integrate these knowledge models in order to preserve the metadata along the global workflow of the multimedia news production and consumption chain. We will develop a provenance ontology to attach provenance information to all metadata generated and display a truthiness score the news presented.

Event detection is at the heart of the **NEED** project. Some events might be obvious (e.g. Tour de France, UEFA Cup, World War II) while others are definitely not (e.g. USA mortgage crisis, wedding of Carla Bruni) which make the automatic detection and semantic annotation of events a real research challenge. Our approach will rely on a suite of contextual and content analysis techniques: entities extracted and categorised from textual stories, visual cues and structuring detected by a multimedia analysis tool-kit, audio processing for quotes detection. Events detected will feed a knowledge base of events, represented using Semantic Web languages and exposed on the web as a semantic web repository linked to existing open datasets.

NEED will address *interface interoperability* along the multimedia news workflow by integrating the existing platforms used by AFP based on the knowledge infrastructure described above. Novel interfaces supporting semantic search and presentation from rich and heterogeneous news datasets will be designed and implemented, for example using a facetted browsing paradigm. These interfaces will support temporal media, multilingual and heterogeneous sources of news information. They will be evaluated using real data from news providers among both journalists and lay users.

NEED will actively participate in and contribute to various standardisation bodies such as W3C, IPTC, EBU and ISO, with the expectation of having significant impact on their development. Consortium partners are already participants and co-chair specific technical working groups. **NEED** aims thus to further develop multimedia news standards and provide and distribute reference implementations of them.

1.3.2 Structuring and timing of the work packages

The overall strategy of the project is broken into nine work packages (WP), which in principle represent different aspects of the project and can be considered logical units of work. The tasks and deliverables of each work package have been defined so that the work packages are able to be run largely in parallel. The master plan identifies the milestones as specific points in time where a formal exchange of information between work packages is required and these exchanges have been identified as specific deliverables. Being able to formally manage the concurrency of individual work packages will substantially speed up the project.

The reasons behind the WP structure are the following:

- The **Project Management** is handled by the project coordinator IBBT in a single work package **WP9**.

- A work package deals with the understanding of user needs and **WP1** will therefore provide concrete **use case scenarios** that will later be **evaluated** by professional journalists and lay users among our user groups in **WP6**.
- Four work packages deal with the core research questions: **WP2** designs a **Knowledge Infrastructure for News Integration** for solving interoperability issues in the current news workflow. **WP3** provide tools for **Event Detection and News Enrichment** based on multimedia analysis techniques, natural language processing including entity recognition, and mining algorithms. **WP4** generates **Semantic Multimedia News Interfaces** offering personalized access to multilingual and multimedia news corpora, linked to relevant and appropriate background knowledge in order to make sense of the news. **WP5** provides the general **infrastructure** for integrating the various technologies and for **scaling with the large amount** of data to deal with.
- A **Standardisation Activities** work package **WP7** will immediately disseminate the technologies promoted by the project and ensure their uptake by the whole web and media industry. In liaison with an original Standardisation Advisory Board, it will allow the project members to have quick feedback with respect to the adoption of the technologies and inquiries towards future needs.
- **Exploitation and Dissemination** of the results are bundled in **WP8**, lead by the SME of the project which already foresees future products ready to enter the market.

The list of all work packages as well as their precise timing is given below in the Table 1.3a and in the Gantt charts in the Figures 6 to 14. The comprehensive description of each work package, tasks and sub-tasks, deliverables and milestones is given in the section 1.3.5.

1.3.3 List of work packages

Work package No	Work package title	Type of activity	Lead partic no.	Lead partic. short name	Person-months	Start month	End month
WP1	Use Case Scenarios	RTD	1	IBBT-MICT	31	1	30
WP2	Knowledge infrastructure for news integration	RTD	1	IBBT-MMLab	56	1	30
WP3	Event Detection and News Enrichment	RTD	2	EURECOM	78	1	36
WP4	Semantic Multimedia News Interfaces	RTD	3	CWI	42	1	36
WP5	Framework Architecture Design	RTD	7	CINECA	44	6	36
WP6	Evaluation	RTD	4	AFP	24	6	36
WP7	Standardization & Outreach	RTD	5	ERCIM/ W3C	26	1	36
WP8	Exploitation & Dissemination	RTD	6	Temis	19	1	36
WP9	Project Management	MGT	1	IBBT-MMLab	26	1	36
	TOTAL				346		

Tableau 1.3a – List of Work Packages

Task Id	Task Name
WP1	**Use Case Scenarios**
WP1.1	**Management of the expert user panel**
WP1.1	Management of the expert user panel
D1.1.1	List of initial members of the expert user panel
D1.1.2	**Interim reports of the virtual (online) panel consultations**
	Interim report 1 of the virtual (online) panel consultations
	Interim report 2 of the virtual (online) panel consultations
	Interim report 3 of the virtual (online) panel consultations
WP1.2	**Analysis of the user needs of different stakeholders**
WP1.2	Analysis of the user needs of different stakeholders
D1.2.1	Report of the qualitative research on user needs for use cases 1 and 2
D1.2.2	Report of the lead user survey for the use case 2
WP1.3	**Elaboration of the use case scenarios**
WP1.3	Elaboration of the use case scenarios
D1.3.1	Initial report on functional user requirements for NEED technologies and services
D1.3.2	Final report on functional user requirements for NEED technologies and services
D1.3.3	Report on two extended use case scenarios

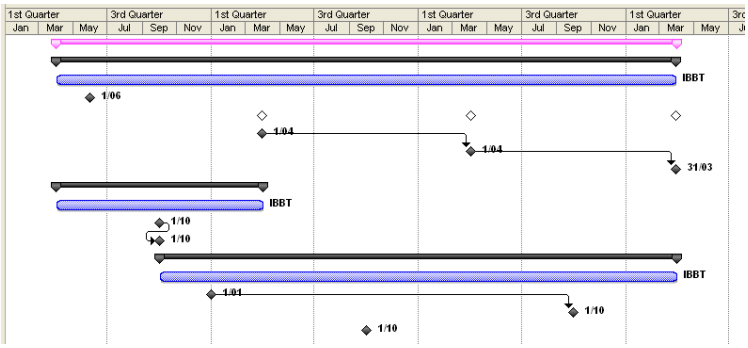


Figure 7: Gantt chart of WP1

Task Id	Task Name
WP2	**Knowledge infrastructure for news integration**
WP2.1	**Identification of News Context Sources**
WP2.1	Identification of News Context Sources
D2.1	Report on existing information sources relevant for news items
WP2.2	**Ontology Engineering**
WP2.2	Ontology Engineering
WP2.2.1	**Specification of the News ontology**
WP2.2.1	Specification of the News ontology
D2.2.1	Initial version of the News, Event and Provenance ontologies
WP2.2.2	**Specification of the Events ontology**
WP2.2.2	Specification of the Events ontology
D2.2.2	Final version of the News, Event and Provenance ontologies
WP2.2.3	**Specification of the Media Ontology (in cooperation with WP3)**
WP2.2.3	Specification of the Media Ontology (in cooperation with WP3)
WP2.2.4	**Specification of the Provenance ontology**
WP2.2.4	Specification of the Provenance ontology
WP2.2.5	**Integration of the various models with social media ontologies (e.g. FOAF, SIOC)**
WP2.2.5	Integration of the various models with social media ontologies (e.g. FOAF, SIOC)
WP2.3	**Ontology Alignment, Integration and Reasoning over Multimedia News**
WP2.3	Ontology Alignment, Integration and Reasoning over Multimedia News
D2.3.1	Initial version of the knowledge infrastructure, containing inter-linked ontologies
D2.3.2	Final version of the knowledge infrastructure, containing inter-linked ontologies

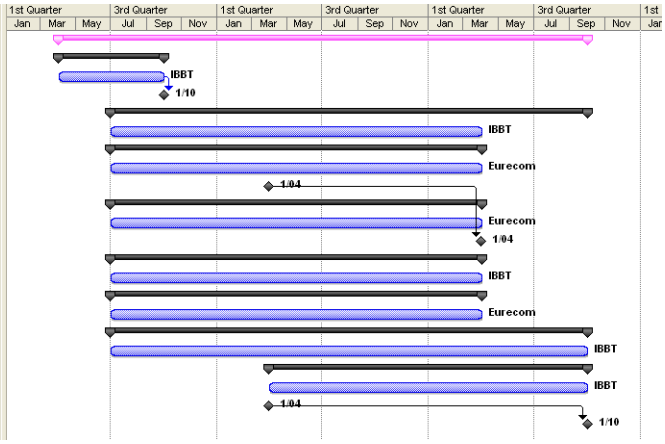


Figure 8: Gantt chart of WP2

Task Id	Task Name
WP3	**Event Detection and News Enrichment**
WP3.1	**Text Processing for Entity Extraction and Categorisation**
WP3.1.1	**Named Entity Extraction**
WP3.1.1	Named Entity Extraction
D3.1.1	Skill Cartridge™ of Named Entities
WP3.1.2	**Entity Categorisation and Evaluation**
WP3.1.2	Entity Categorisation and Evaluation
D3.1.2.1	Specifications for the Event Skill Cartridge™
D3.1.2.2	Event Skill Cartridge™
WP3.1.3	**Toolbar for Enriching Documents**
WP3.1.3	Toolbar for Enriching Documents
D3.1.3	Temis cross web browser toolbar
WP3.2	**Content and Context Analysis for Multimedia News Items**
WP3.2.1	**Concept Detection Toolbox**
WP3.2.1	Concept Detection Toolbox
D3.2.1	Specification and Methodologies for multimedia news analysis
WP3.2.2	**Multimedia News Item Alignment ToolBox**
WP3.2.2	Multimedia News Item Alignment ToolBox
D3.2.2	Initial toolkit for multimedia news analysis
WP3.2.3	**Multimedia News Item Validation ToolBox**
WP3.2.3	Multimedia News Item Validation ToolBox
D3.2.3	Final toolkit for multimedia news analysis
WP3.3	**Event Detection and Annotation**
WP3.3	Event Detection and Annotation
D3.3.1	Initial report on event detection and annotation framework
D3.3.2	Final report on event detection and annotation framework
WP3.4	**Trend Detection from Microblogging Feeds**
WP3.4	Trend Detection from Microblogging Feeds
D3.4.1	Initial report on trend detection from microblogging feeds
D3.4.2	Final report on trend detection from microblogging feeds

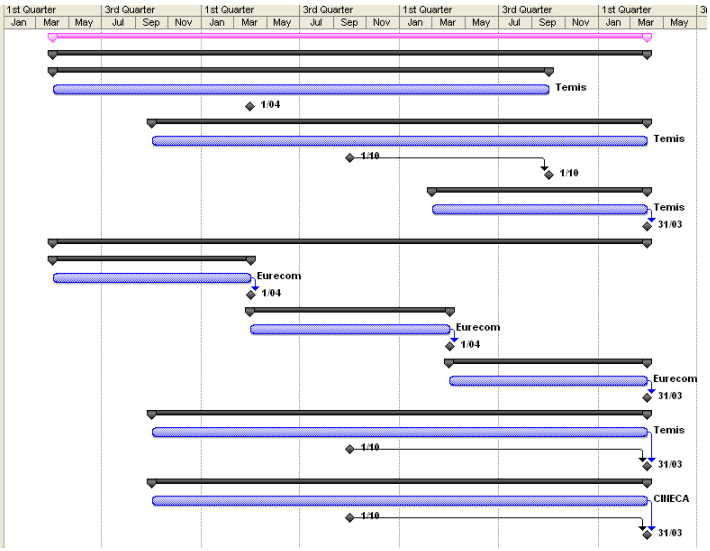


Figure 9: Gantt chart of WP3

Task Id	Task Name
WP4	**Semantic Multimedia News Interfaces**
WP4.1	**Analysis of user needs (in cooperation with WP1)**
WP4.1	Analysis of user needs (in cooperation with WP1)
D4.2.1	Report on user needs (complementary to D1.2.1 and D1.3.1)
WP4.2	**Map user needs to functionality provided by NEED (in cooperation with WP2, WP3 & WP5)**
WP4.2	Map user needs to functionality provided by NEED (in cooperation with WP2, WP3 & WP5)
D4.2.1	Report on functionality and expert panel feedback on the release of the first set of partial prototypes
D4.2.2	1st Public functional prototype ready for internal testing & demo/dissemination purposes
D4.2.3	2nd Public functional prototype ready for user testing
WP4.3	**Evaluation and user testing (in cooperation with WP6)**
WP4.3	Evaluation and user testing (in cooperation with WP6)
D4.3.1	Report on user evaluation tests for the "Paris" workshop
D4.3.2	Report on user evaluation tests for the "Sophia Antipolis" workshop
D4.3.3	Final public prototype

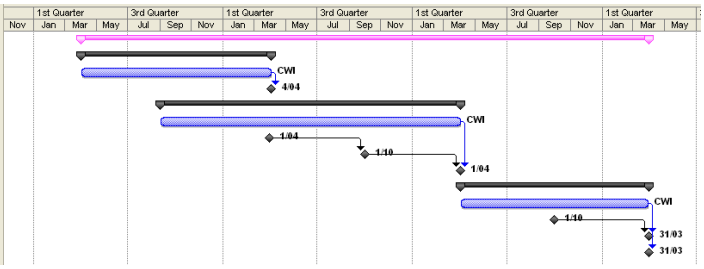


Figure 10: Gantt chart of WP4

Task Id	Task Name
WP5	**Framework Architecture Design**
WP5.1	**Infrastructure modeling and deployment**
WP5.1	Infrastructure modeling and deployment
D5.1	Report on the NEED infrastructure design and system architecture
WP5.2	**System architecture design**
WP5.2	System architecture design
D5.2	Report on the complete NEED implementation plan
WP5.3	**System Integration**
WP5.3	System Integration
D5.3.1	Framework for integrated NEED services v1
D5.3.2	Framework for integrated NEED services v2
D5.3.3	Framework for integrated NEED services v3

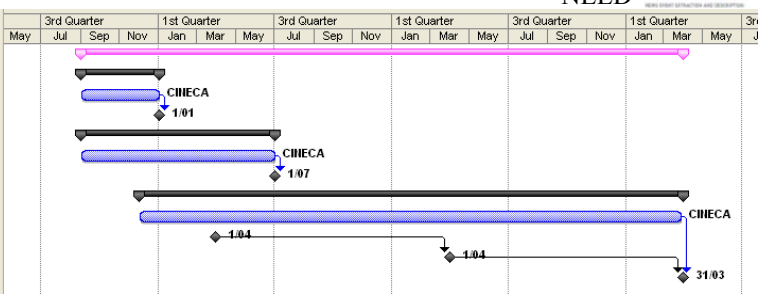


Figure 11: Gantt chart of WP5

Task Id	Task Name
WP6	**Evaluation**
WP6.1	**Workshop on list of user requirements**
WP6.1	Workshop on list of user requirements
D6.1	Strategic report of the results of workshop 1 on user needs
WP6.2	**Workshop on the success criteria of the NEED approach based on users scenarios**
WP6.2	Workshop on the success criteria of the NEED approach based on users scenarios
D6.2	Strategic report of the results of workshop 2 on use case scenarios
WP6.3	**Evaluation of NEED prototype**
WP6.3	Evaluation of NEED prototype
D6.3.1	First evaluation report of the prototypes developed for the use cases 1 and 2
D6.3.2	Second evaluation report of the prototypes developed for the use cases 1 and 2

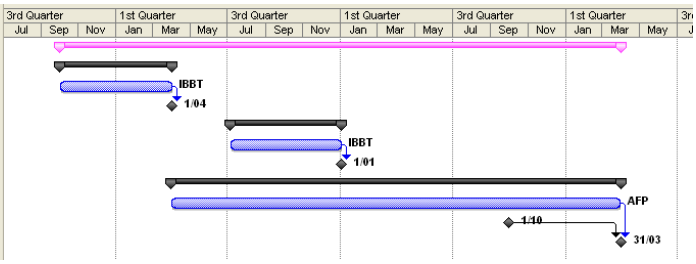


Figure 12: Gantt chart of WP6

Task Id	Task Name
WP7	**Standardization and Outreach**
WP7.1	**Standardization**
WP7.1	Standardization
D7.1.1	Working Groups Charter extension
D7.1.2	Working Groups Last Call Documents and Disposition of Comments
D7.1.3	Working Groups Candidate Recommendation Documents
D7.1.4	Working Groups Recommendations
WP7.2	**Outreach**
WP7.2	Outreach
D7.2.1	**Press Releases**
	Press Release 1
	Press Release 2
	Press Release 3
D7.2.2	Press Clippings
D7.2.3	Promotion Talks Report
D7.2.4	**Communication Material**
	Communication Material Dissemination 1
	Communication Material Dissemination 2
	Communication Material Dissemination 3

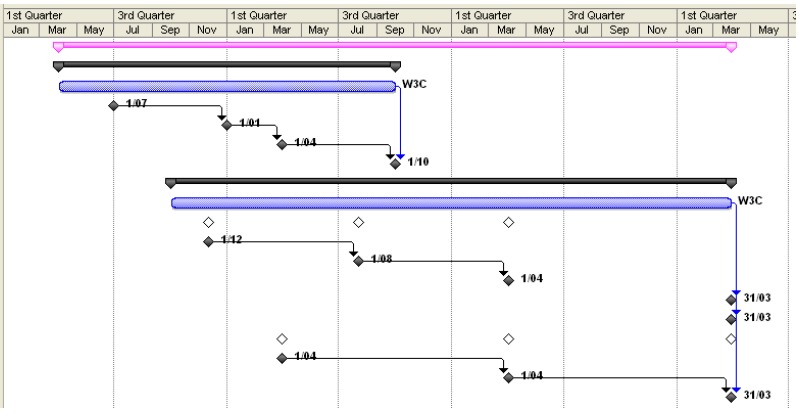


Figure 13: Gantt chart of WP7

Task Id	Task Name
WP8	**Exploitation and Dissemination**
WP8.1	**Project Dissemination**
WP8.1.1	The NEED Web site
D8.1.1	Project Web site fully functional
WP8.1.2	Six-monthly Electronic Newsletter
D8.1.2	**NEED Electronic Newsletter**
	NEED Electronic Newsletter 1
	NEED Electronic Newsletter 2
	NEED Electronic Newsletter 3
	NEED Electronic Newsletter 4
	NEED Electronic Newsletter 5
	NEED Electronic Newsletter 6
WP8.2	**Project Exploitation**
WP8.2.1	Definition of a Marketing Plan and Market analysis
D8.2.1	Marketing Plan
WP8.2.2	Package an "Events Detection" Skill Cartridge
D8.2.2	Skill Cartridge "Events Detection"
WP8.2.3	Package a wrapper of key NEED technologies on AFP news retrieval
D8.2.3	Contextualized visualization of news events product

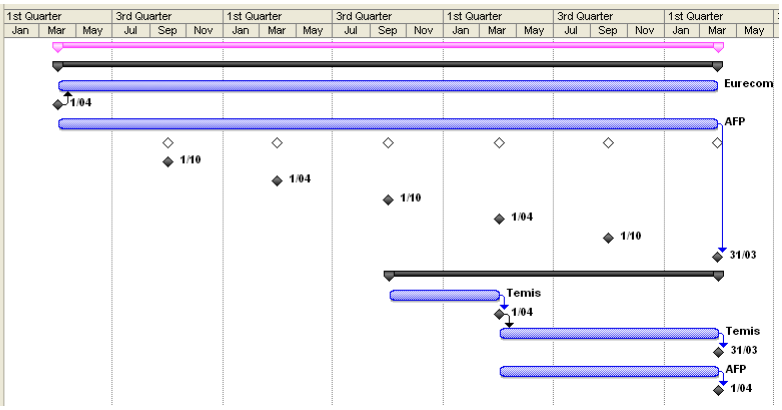


Figure 14: Gantt chart of WP8

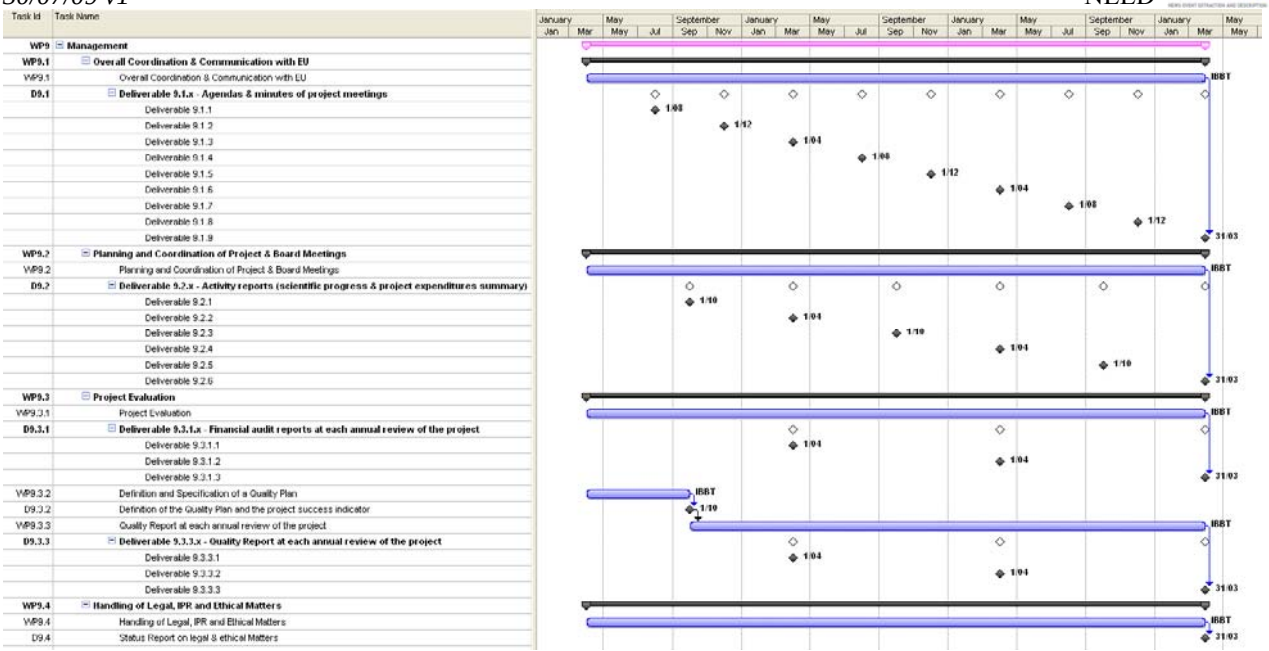


Figure 15: Gantt chart of WP9

1.3.4 List of Deliverables and Milestones

Del. no.	Deliverable name	WP no.	Nature	Dissemination level	Delivery date (proj. month)
D1.1.1	List of initial members of the expert user panel	WP1	R	PU	2
D1.1.2.x	Annual reports of the virtual (online) panel consultations: summary of the interim reports	WP1	R	PU	12x
D1.2.1	Report of the qualitative research on user needs for use cases 1 and 2	WP1	R	PU	6
D1.2.2	Report of the lead user survey for the use case 2	WP1	R	PU	6
D1.3.1	Initial report on functional user requirements for NEED technologies and services	WP1	R	PU	9
D1.3.2	Final report on functional user requirements for NEED technologies and services	WP1	R	PU	30
D1.3.3	Report on two extended use case scenarios (showing multiple alternative use cases and different user profiles and roles)	WP1	R	PU	18
D2.1	Report on existing information sources relevant for news items	WP2	R	PU	6
D2.2.1	Initial version of the news, event and provenance ontologies	WP2	R	PU	12
D2.2.2	Final version of the news, event and provenance ontologies	WP2	R	PU	24

D2.3.1	Initial version of the knowledge infrastructure, containing inter-linked ontologies	WP2	R	PU	12
D2.3.2	Final version of the knowledge infrastructure, including rules and domain-specific ontologies	WP2	R	PU	30
D3.1.1	Skill Cartridge™ of Named Entities	WP3	R	PU	12
D3.1.2.1	Specifications for the Event Skill Cartridge™	WP3	R	PU	18
D3.1.2.2	Event Skill Cartridge™	WP3	P	PU	30
D3.1.3	Temis cross web browser toolbar	WP3	P	PU	36
D3.2.1	Specification and Methodologies for multimedia news analysis	WP3	R	PU	12
D3.2.2	Initial toolkit for multimedia news analysis	WP3	R	PU	24
D3.2.3	Final toolkit for multimedia news analysis	WP3	R	PU	36
D3.3.1	Initial report on event detection and annotation framework	WP3	R	PU	18
D3.3.2	Final report on event detection and annotation framework	WP3	R	PU	36
D3.4.1	Initial report on trend detection from micro-blogging feeds	WP3	R	PU	18
D3.4.2	Final report on trend detection from micro-blogging feeds	WP3	R	PU	36
D4.1	Report on user needs (complementary to D1.2.1 and D1.3.1)	WP4	R	PU	12
D4.2.1	Report on functionality and expert panel feedback on the release of the first set of partial prototypes	WP4	R	PU	12
D4.2.2	First public functional prototype ready for internal testing & demo/dissemination purposes	WP4	R	PU	18
D4.2.3	Second public functional prototype ready for user testing	WP4	R	PU	24
D4.3.1	Report on user evaluation tests for the “Paris” workshop	WP4	R	PU	30
D4.3.2	Report on user evaluation tests for the “Sophia Antipolis” workshop	WP4	R	PU	36
D4.3.3	Final public prototype	WP4	R	PU	36
D5.1	Report on the NEED infrastructure design and system architecture	WP5	R	PU	9
D5.2	Report on the complete NEED implementation plan	WP5	R	PU	15

D5.3.1	Framework for integrated NEED services v1	WP5	R	PU	12
D5.3.2	Framework for integrated NEED services v2	WP5	R	PU	24
D5.3.3	Framework for integrated NEED services v3	WP5	R	PU	36
D6.1	Strategic report of the results of workshop 1 on user needs	WP6	R	PU	12
D6.2	Strategic report of the results of workshop 2 on use case scenarios	WP6	R	PU	20
D6.3.1	First evaluation report of the prototypes developed for the use cases 1 and 2	WP6	R	PU	30
D6.3.2	Second evaluation report of the prototypes developed for the use cases 1 and 2	WP6	R	PU	36
D7.1.1	Working Groups Charter extension	WP7	R	PU	3
D7.1.2	Working Groups Last Call Documents and Disposition of Comments	WP7	R	PU	9
D7.1.3	Working Groups Candidate Recommendation Documents	WP7	R	PU	12
D7.1.4	Working Groups Recommendations	WP7	R	PU	18
D7.2.1.x	W3C press releases	WP7	R	PU	8 / 16 / 24
D7.2.2	W3C press clippings	WP7	R	PU	36
D7.2.3	W3C promotion talks	WP7	R	PU	36
D7.2.4.x	NEED communication material	WP7	O	PU	12x
D8.1.1	Project web site fully functional	WP8	O	PU	1
D8.1.2.x	Six-monthly electronic newsletter	WP8	R	PU	6x
D8.2.1	Definition of the Marketing Plan	WP8	R	CO	12
D8.2.2	Skill Cartridge “Events Detection” product	WP8	P	PU	36
D8.2.3	Contextualized visualization of news events product	WP8	P	PU	36
D9.1.x	Agendas and official minutes of relevant project meetings	WP9	R	CO	4x
D9.2.x	Six-monthly activity reports (scientific progress and project expenditures summary)	WP9	R	CO	6x
D9.3.1.x	Financial audit reports at each annual review of the project	WP9	R	CO	12x
D9.3.2	Definition of the Quality Plan and the project success indicators	WP9	R	CO	6
D9.3.3.x	Quality Report at each annual	WP9	R	CO	12x

	review of the project				
D9.4	Status report on legal and ethical matters	WP9	R	PU	36

1.3.5 Work package description

Work package number	WP1	**Start date or starting event:**		M1		
Work package title	Use Case Scenarios					
Activity type	RTD					
Participant number	**1**	2	3	4	5	6	7
Participant short name	**IBBT-MICT**	EURECOM	CWI	AFP	ERCIM/ W3C	Temis	CINECA
Person-months per participant	**15**	0	6	10	0	0	0

Objectives

- To establish and maintain the expert user panel;
- To conduct two user studies at the beginning of the project, each focusing on one of the two use cases;
- To identify and understand the newsgathering practices and information needs of different users (journalists, informed decision makers, EU citizens interested in environmental issues, etc.) within both use cases (see Section 1.1.2);
- To gather functional user requirements for NEED technologies and services.

Description of work

WP1.1 Management of the expert user panel (*IBBT-MICT, month 1-30*)

From the beginning of the project an expert user panel will be constructed. This panel will act as a sounding board for the developments during the project. The expert user panel will consist of experts from various international business organizations, non-governmental bodies and research and government institutes. Membership numbers are targeted to be about 10 in the initial phase, ramping up to 15 towards the end of the project. A first list of interested organizations have been approached and have already committed to be part of the user panel (see Annex A).

Using a multi-methodological approach, this panel will be polled for their (expert) opinion on a regular basis, both through workshops and virtual panel consultations. For the virtual panel consultations, (selected parts of) the expert user panel will be consulted online on a regular basis (at least 4 times per year) to gather specific feedback on interim results and preliminary mockups. The virtual panel consultations will also be used as a means to ensure continuous contact with and involvement of the members of the expert user panel.

The organization of the public workshops, for which the expert user panel will be invited, is part of WP6.

The activities conducted within this task can be summarized as follows:

- Establishment and maintenance of the expert user panel
  - Identifying user profiles of stakeholders for the two use cases
  - Contacting and recruiting representative experts for the user panel, ensuring a representative distribution over stakeholder categories and good European coverage
- Organization of iterative virtual panel consultations throughout the project (interim reports)

IBBT-MICT in cooperation with WP6 are involved in this task.

WP 1.2 Analysis of the user needs of different stakeholders (*IBBT-MICT, month 1-12*)

In the first phase, two user studies will be conducted to identify and understand the information and news gathering needs of different users in the two use cases.

For the use case 1, a user study, based on in-depth interviews and ethnographic methods, will analyze the

existing processes of knowledge management, information retrieval and news gathering within different organizations/stakeholders dealing with environmental issues, ranging from political parties over environmental NGO's and human organizations to news media and civil communities (both online and offline). The assumption is that task divisions, competencies, techniques, workflows, systems and information channels used to gather news and handle knowledge and information will widely differ from one organization/stakeholder to the other. The social research will help to identify commonalities and particularities of the various needs of different actors involved. This will result in a clear set of user requirements that will be used as input for the technical work packages 2, 3, 4 and 5.

In parallel, a similar study will be conducted for the use case 2. In-depth interviews will be held and ethnographic research will be carried out to analyze and evaluate the current newsgathering process in professional news organizations. Here, however, the focus will be on issues of verifying information, establishing provenance and trust index, and linking (video) news items to other related relevant information. Furthermore, using a combined methodology from Social Sciences and the lead user methodology[59] defined by Eric Von Hippel (MIT Sloan School of Management) in his book The Sources of Innovation[60], a survey will be conducted among key news addicts (professional journalists, news bloggers, news consumers, decision makers, etc.) to observe their strategies and behaviors to watch news, to verify its accuracy and reliability, and eventually to validate it in a professional media workflow.

The results from the user studies will be presented in the first WP6 workshop to the expert user panel to gather additional user feedback (see WP6). Identifying lead users will also help to identify market trends for exploitation activities in WP7.

The activities conducted within this task can be summarized as follows:

- In-depth interviews with different types of users combined with participant observation within organizations.
    - Task analysis: description of the different tasks and competencies in place in the organizations for retrieving and handling news and information
    - Workflow analysis: description of the routines, process and flow of information management within the organizations
    - System analysis: focus on systems and channels used for information and news management within the organizations in order to detect design and interface requirements
- Survey of lead users (media professionals, bloggers, ...) about their source verification strategies

Translation of user study results into functional requirements for NEED (in cooperation with WP2, WP3, WP4 and WP5)

IBBT-MICT, AFP and CWI are involved in this task.

WP1.3 Elaboration of the use case scenarios (*IBBT-MICT, month 6-30*)

Throughout the project, the use cases will be further refined, updated and elaborated on the basis of findings of the user studies, but also in function of the progress made in the different work packages. More particularly, the results of the user analyses will be presented using personas, scenarios and task hierarchies. Internal workshops will be organized to collect and integrate input from the different technical workpackages and translate this to the use case scenarios. The objective is to develop two broad user scenarios, one on handling information on environmental issue and another on verifying and contextualizing (video) news and information. These user scenarios will describe the roles and information needs of different actors (presented as 'personas') within the respective scenarios and will show the wide range of possible variants to the central use case. The elaborated use case scenarios will be presented in the second WP6 workshop to the expert user panel to gather additional user feedback (Task 6.1).

The activities conducted within this task can be summarized as follows:

[59] *Lead users* are defined by Eric Von Hippel as those users who face the needs that will be general in a marketplace months or years before the bulk of that marketplace encounters them and lead users are positioned to benefit significantly by obtaining a solution to those needs (see http://www.leaduser.com for more details)
[60] http://web.mit.edu/evhippel/www/sources.htm

Alignment of the use case scenarios to the user study findings

Identification of alternative use cases within each broader user scenario

Detailed descriptions of different archetypes of users (personas)

Organization of internal workshop to integrate intermediary results from the technical workpackages (WP2, WP3, WP4 and WP5)

IBBT-MICT in cooperation with the user groups are involved in this task.

Deliverables

- D1.1.1 (IBBT-MICT, month 2): List of initial members of the expert user panel
- D1.1.2.x (IBBT-MICT, month 12-24-36): Annual reports of the virtual (online) panel consultations: summary of the interim reports
- D1.2.1 (IBBT-MICT, month 6): Report of the qualitative research on user needs for use cases 1 and 2
- D1.2.2 (IBBT-MICT, month 6): Report of the lead user survey for the use case 2
- D1.3.1 (IBBT-MICT, month 9): Initial report on functional user requirements for NEED technologies and services
- D1.3.2 (IBBT-MICT, month 30): Final report on functional user requirements for NEED technologies and services
- D1.3.3 (IBBT-MICT, month 18): Report on two extended use case scenarios (showing multiple alternative use cases and different user profiles and roles)

Work package number	WP2	**Start date or starting event:**		M1		
Work package title	Knowledge infrastructure for news integration					
Activity type	RTD					
Participant number	**1**	2	3	4	5	6	7
Participant short name	**IBBT-MMLab**	EURECOM	CWI	AFP	ERCIM/ W3C	Temis	CINECA
Person-months per participant	**28**	18	0	5	5	0	0

Objectives

- To develop a formal ontology suitable for representing chains of news events;
- To develop a provenance ontology suitable for attaching reliable provenance metadata to news event and derive trust index;
- To expose controlled vocabularies and thesaurus on the web as formalized linked data;
- To provide an integrated knowledge infrastructure for multimedia news.

Description of work

The general objective of this work package is to create the knowledge infrastructure which models the news-related data and metadata. NEED will address metadata interoperability along the multimedia news workflow by designing, leveraging and integrating a number of multimedia and news ontologies. The resulting knowledge infrastructure will be thus ontology-mediated and expressed in knowledge representation web-based standards (OWL, RDF, SKOS). Further, an expressive provenance ontology will be developed in order to represent the provenance metadata extracted in WP3 and attach them to the news for deriving trust index. This ontology will be developed in collaboration with the Linked Data community and within the W3C Provenance Incubator Group in order to maximize its uptake. Finally, provisions will be made to link the multimedia and news ontologies to more generic context information (e.g. the description of a city that is subject of a news item) and specific vocabularies (e.g. fine-grained thesaurus used to describe some environmental issues). The goal here is to re-use as much as possible existing large datasets collected in the Linked Data cloud and add new ones when appropriate.

Several initiatives have proposed a way to represent news events in a structured way. The media industry has developed within IPTC the EventsML-G2 standard while we have compared and aligned many event ontologies [Shaw et al., 2009]. EventsML-G2 relies on the News Architecture (NAR), an IPTC framework for the management and description of news and associated information. NewsML-G2 is also based on this architecture and can be further linked with an event model. Both models make use of multiple (and sometimes overlapping) controlled vocabularies, maintained by different authorities and we have already proposed a first OWL ontology of this model that needs to be refined and embraced by the media industry [Troncy, 2008]. Finally, other work on bridging W3C and MPEG-7 related standards for the description of multimedia content has been proposed, and has resulted in the core ontology for multimedia (COMM) [Arndt et al., 2007]. There is now an urgent need to integrate these various knowledge models in order to optimize the global workflow of the multimedia news production and consumption chain.

This Work Package will not only use Semantic Web Technologies, it will also actively contribute to current standardization processes on the creation of a Media Ontology (in the W3C Media Annotations Working Group[61]), a Provenance Ontology (in the W3C Provenance Incubator Group[62]), and the refinement and reference implementation of the Media Fragments URI specification (in the W3C Media Fragments Working Group[63]). It is composed of three tasks and runs for the first 30 months of the project.

[61] http://www.w3.org/2008/WebVideo/Annotations/
[62] http://www.w3.org/2005/Incubator/prov/
[63] http://www.w3.org/2008/WebVideo/Fragments/

WP2.1 Identification of News Context Sources (*IBBT-MMLab, month 1-6*)

This task will focus on listing, analyzing and classifying the different sources and type of information that compose a news item. These sources can describe specific news items; include multimedia content, information on the context or related events, information on the origin of the news (meaning the creator or publisher) or simply explaining in more details some important concepts for understanding the event context. Provenance and trust information should be attached to all metadata. The goal is to see which information sources are available and how these can be inter-linked to provide rich context regarding a particular news item. Additionally, the results of the user interviews in WP1 will rank these sources based on their importance for the various targeted users. The different categories of data that will be exploited are:

- General knowledge bases for describing news items (e.g. the IPTC News Codes, general thesaurus indexing all companies such as TechCrunch, etc.);
- Existing metadata of multimedia content (possibly included in the news item);
- Specific or general domain knowledge, not necessarily in the context of news (Geonames for information about geographical locations, IMDB for information about movies, WordNet for general terms, DBPedia for a selected part of Wikipedia, etc.);
- Specific vocabularies for the environmental issues use case
- Social media to determine what the public response is on a news item, or to obtain more information on (the network of) the creator;
- Provenance information.

IBBT-MMLab, EURECOM and AFP are involved in this task.

WP 2.2 Ontology Engineering (*IBBT-MMLab, month 3-24*)

Once relevant information sources within each category have been determined, the goal is to create a general knowledge infrastructure. We will use the Resource Description Framework (RDF) data model and the linked data principles to integrate various ontologies developed in this task.

Several schemas for representing the notion of a (news) event have been proposed and IPTC is now working on the future version of EventsML, but it has not yet defined a formal model for the representation of chains of events. This task aims to produce a formal OWL ontology while being compatible with existing event-centric models. The design of this ontology will start from the core questions such as "what makes a news event?", "what is the signature of an event?", "what are its specific characteristics?" and "how are events chained?". This knowledge infrastructure will be based on our pre-existing work regarding a news ontology [Troncy, 2008] and an event ontology [Shaw et al., 2009].

Finally, it is important to determine which information is relevant, redundant, and trustworthy to contextualize a news item. Actually, linking to external information without the appropriate means that allow contrasting its provenance can be harmful, especially in sensitive domains. For example, in September 2009, Germany was shocked by a scheme that managed to have the Deutsche Presse Agentur (DPA[64]) report an attempted suicide attack supposedly perpetrated by a non-existing terrorist group called Berlin Boys[65] in a city called Bluewater, California. In fact, this turned to be a hoax organized by a group of German filmmakers trying to call the attention of the media on their next work. The elaborated hoax involved at least two faked websites, a faked Wikipedia entry and California phone numbers for public safety officials that were actually being answered by hoaxers in Germany using Skype. Eventually, the hoax prompted a 1000-word tome on the website of Frankfurter Allgemeine Zeitung and forced DPA to publicly apologize for the incident. In a provenance-aware world, DPA would have had means to automatically reason on the provenance of the information and would have realized e.g. that the city itself did not really exist despite it appeared in Wikipedia, used by DPA as a trustworthy information source.

We will propose a Provenance ontology compatible with Semantic Web knowledge representation formalisms and based on Provenance and Linked Data-specific vocabularies such as the Open Provenance Model (OPM), VoID, SIOC, and FOAF in order to describe the provenance of information in terms of

[64] http://www.dpa.de/
[65] Read http://www.wired.com/threatlevel/2009/09/bluewater

agents, artefacts, processes, and the relationships between them. Reasoning about the provenance of information will allow assessing and certifying its quality and reliability.

The following sub-tasks will be carried out:

- WP2.2.1: Specification of the News ontology (EURECOM, month 3-24)
- WP2.2.2: Specification of the Events Ontology (EURECOM, month 3-24)
- WP2.2.3: Specification of the Media Ontology to integrate different multimedia metadata standards and to represent information extracted from the media resources (in cooperation withWP3) (IBBT, month 3-24)
- WP2.2.4: Specification of the Provenance Ontology (IBBT, month 3-24)
- WP2.2.5: Integration of the various models with social media ontologies (e.g. FOAF, SIOC) (IBBT, month 3-24)

IBBT-MMLab, EURECOM, W3C and AFP are involved in this task.

WP2.3 Ontology Alignment, Integration and Reasoning over Multimedia News (*IBBT-MMLab, month 18-30*)

News items are more and more multimedia-based and make use sometimes of user generated content provided by citizen journalists. Hence, AFP plans to double its photo activity and multimedia products during the next 5 years. Describing both the multimedia aspects of the news items and its content requires an integrated knowledge infrastructure aligning ontologies such as NewsML, EventsML, COMM and other multimedia and events-based models, and dealing with lightweight annotations such as the tags accompanying user generated content. This task will provide this integrated knowledge infrastructure by:

- Creating relations between concepts in the different ontologies
    - Use standardized constructs (RDF, OWL, SKOS) for inter-linking
    - Introduce rules when custom relations are not sufficient
- Integrating the formal representations into a knowledge infrastructure
    - Create rules to identify and deal with redundant or contradictory information
    - Create ways to prioritize and summarize the different information sources

IBBT-MMLab, EURECOM and AFP are involved in this task.

Deliverables

- D2.1 (IBBT, month 6): Report on existing information sources relevant for news items
- D2.2.1 (EURECOM, month 12): Initial version of the news, event and provenance ontologies
- D2.2.2 (EURECOM, month 24): Final version of the news, event and provenance ontologies
- D2.3.1 (IBBT, month 12): Initial version of the knowledge infrastructure, containing inter-linked ontologies
- D2.3.2 (IBBT, month 30: Final version of the knowledge infrastructure, including rules and domain-specific ontologies

Work package number	WP3	**Start date or starting event:**		M1		
Work package title	Event Detection and News Enrichment					
Activity type	RTD					
Participant number	1	**2**	3	4	5	6	7
Participant short name	IBBT-MMLab	**EURECOM**	CWI	AFP	ERCIM/ W3C	Temis	CINECA
Person-months per participant	4	**30**	0	6	0	30	8

Objectives

- To extract named entities and structured data from textual document;
- To automatically align quotes to video sequences
- To extract concepts from video content and to assess the provenance of news content
- To perform trend detection on large amount of micro-blogging feeds
- To extract events using statistic and linguistic technologies and to provide semantic web description of these events

Description of work

The objective of this work package is to design and implement an efficient workflow dedicated to:

- Entities and events extraction: named Entities (like persons, organizations etc.) will be extracted in order to efficiently index news documents. Events (press events and events related to the scenarios) will be detected in documents.
- Documents categorization: news documents will be automatically categorized in a set of predefined themes. Extracted named entities will enrich the document contents and then enhance the categorization process.
- Text content analysis and enrichment: a set of analysis and visualisation tools will allow users navigating into knowledge in an efficient and personalized way. They will allow locating relevant information quickly, automatically identifying concepts in texts and adding dynamically computed contextual links, or accessing dynamically a customized knowledge base. This functionality permit to users to navigate dynamically through heterogeneous documents.
- Provide a suite of contextual and content analysis techniques for multimedia news items.

This work package runs for the entire duration of the project and consists of four tasks.

WP3.1 Text Processing for Entity Extraction and Categorisation (*Temis, month 1-36*)

This task aims to provide tools for extracting and categorising semantically entities from textual documents being news items, blogs or event microblogging feeds.

- WP3.1.1 Named Entity Extraction (Temis, month 1-30)

Identifying entities such as the names of companies, associations, organizations, products figures, dates and places as well as the relationships between these entities (e.g. company-company, person-company, company-product, etc.) is key to indexing and presenting news stories. We will build on the knowledge extraction engine based on Temis Insight Discoverer™ Extractor (IDE) powered by Skill Cartridges™, a hierarchy of knowledge components describing the information to extract for a given business, specific field or topic. We will customize this extraction engine and the sequence of the three linguistic analysis steps it performs to the news domain: corpus recognition (language identification), morpho-syntactic analysis (lemmatization) and knowledge extraction using dedicated rules.

- WP3.1.2 Entity Categorisation and Evaluation (Temis, month 6-36)

The extraction process provides as a result a set of entities structured according to the ontologies and controlled vocabularies defined in the knowledge infrastructure (see WP2). Next, the sets of extractions will be exploited in order to automatically categorize documents by assigning one or several predefined categories to a document. Categorization classically used a representation of documents like a set of words with their frequencies ("bag of words approach"). Entities extraction permit to provide a semantic representation to documents (concepts and not only words) and then enhance efficiently the categorization process.

The categorization process provided by Temis Insight Discoverer™ Categorizer (IDK) is based on an artificial learning algorithm that learns from a tagged corpus the most representative descriptors for each category. The categorization model will then assign automatically one or more categories to all new documents.

The entities extracted are semantically categorized following a semi-supervised learning approach. Entities already categorized according to the ontologies and controlled vocabularies developed in the knowledge infrastructure (WP2) undergo morpho-syntactic processing by the extraction engine, which associates a semantic descriptor to them (such as frequency of nouns or verbs, noun phrases). These entities are used as a basis for learning, and enable the categorisation engine to create the categorization model using an algorithm that combines the various semantic descriptors assigned to the same category. The categorizer engine can then assign all new entities to the various pre-defined categories based on similarity with the semantic descriptor. We will finally assess the correctness of this automatic categorisation.

- WP3.1.3 Toolbar for Enriching Documents(Temis, month 24-36)

This sub-task is dedicated to provide a cross web browser toolbar packaging the technologies described above. This Temis Tool Bar will allow users locating relevant information quickly in documents, dynamically computed contextual links, or accessing dynamically different data sources. This functionality will permit to users to simply navigate through heterogeneous documents in a customized way. Further, we will thoroughly compare the performances of our extraction and disambiguation technologies with other services such as OpenCalais or Zemanta in order to bring to the market more competitive alternatives.

Temis and EURECOM are involved in this task.

WP 3.2 Content and Context Analysis for Multimedia News Items (*EURECOM, month 1-36*)

In order to further enrich news items with relevant pieces of information gathered across the web, a number of multimedia analysis steps have to be performed on the accompanying and related media items. NEED will create tools to address the detection of concepts in multimedia items, to link automatically news items that are related to each other and to assess the provenance of news items. Rather than attempting to solve the never-ending multimedia semantic gap problem, NEED will focus on further developing known multimedia analysis techniques providing already good enough results in the news domain.

- WP3.2.1 Concept Detection Toolbox

First of all, and similarly to the state-of-the-art, candidate media items will undergo the low-level content analysis step necessary for further inspection. In case the media item corresponds to a video, the identification of the temporal structure of a video is an essential task for video indexing, retrieval and presentation. The steps commonly taken for video analysis are to detect shot boundaries and to extract key frames that best represent the semantic content of each shot. In NEED we will address video content as moving objects rather than representative keyframes in order to extrac richer low-level semantic information using feature extraction and machine learning techniques.

The high-level concepts (i.e., semantic entities) occurring in media items are detected based on both context information (i.e., existing metadata attached to the media items) and information coming from the low-level image, video, and audio analysis. More specifically, media news items will be enriched based on the ontologies developed in WP2 (top-down approach) and based on feedback techniques coming from the low-level feature extraction algorithms (bottom-up approach). Therefore, interaction will be necessary between the low-level content analysis algorithms and the high-level concept detection techniques, thus finding semantic entities by an iterative process.

The high-level concept detection technique will be based on a smart decision-taking engine that is able to build up a plan for the detection of a requested high-level concept. Therefore, a first step in developing this

high-level concept detection technique, the low-level feature extraction techniques will be formally described in terms of their capabilities, requirements, and output. Next, the decision-taking engine decides which low-level feature extraction techniques to use, in order to detect a requested high-level concept. Based on the outcome of these feature extraction techniques, the decision-taking engine can decide to adapt his plan, or to return the found high-level concepts. Note that this decision-taking engine will be both steered by formally described domain-specific rules and real-world knowledge available in the Semantic Web.

- WP3.2.2 Multimedia News Item Alignment ToolBox

One of the key services provided by NEED is the alignment between news quotes and multimedia document gathered through web search. Hence, a user of the NEED services reading a quote in a article would be able to automatically watch a video of this person saying these sentences if available. In order to achieve level of performance yet unattained, we will investiguate joint spatio-temporal multimodal mining and search methods. The task consists in searching predefined web services (YouTube, Flickr, CNN, Wikipedia) for relevant information according to a given "news quote" or eventually multiple quotes. As in every search situation, the resulting list of multimedia items will contain both relevant and irrelevant documents. A semi-automated filtering stage should be applied in order to prune totally irrelevant documents, and rank the remaining ones according to some automatically computed confidence level.

- WP3.2.3 Multimedia News Item Provenance Validation ToolBox

Multimedia news item validation consists in identifying within relevant related news items those originating from the exact same event and pruning others out. Here, multimodal algorithms will be researched and developed in order to detect visual clues allowing to verify with a high level of confidence the temporal and spatial origin of a document in order to ensure their validity. For video documents, this will consist in searching for visual cues strongly correlating the documents originating from "trusted" sources with the ones from socially contributed repositories. The resulting toolkit will present the user with relevant news items and highlight likely invalid items.

EURECOM and IBBT-MMLab are involved in this task.

WP3.3 Event Detection and Annotation (*Temis, month 6-36*)

Event detection is at the heart of NEED. It is a very complex task since some events are scheduled (e.g. Tour de France, UEFA Cup) while others are definitively not (e.g. natural disaster like a typhoon, protests in Iran, USA mortgage crisis). Several parameters will be used to qualify an event: time, knowledge base of personalities, past and recurrent events, industry background, etc. Our approach for detecting events and semantically annotating them will rely on the entities extracted and categorized from textual stories as well as on the visual clues detected by our multimedia analysis tool-kit. These will also be complemented by the statistical tools developed within WP3.4. We will develop semantic algorithms that gather individual pieces of news from different media and group them into contextualised (breaking or scheduled) events. Identified scenarii will help validating the approach.

The Events skill cartridge™ will be implemented to allow the extraction of entities and relationships of the journalistic domain. NEED services will also be used to analyze any kind of news information related to some environmental issues (e.g. climate change and global warming, pollution, energy conservation and renewable energy, etc.) in liaison with the use cases developed within WP1.

The Events Skill Cartridge™ will be designed so as to respond to the main questions which arise when an event comes, namely:

- What is the nature of the event: sport, environment, politics, technology, etc.?
- When and where did the event occur?
- Who are the actors involved in the events?
- Which are the operation theatres for such events?
- What and who is the target for such events (if any)?
- What are the causal and consequence relationships between events?

Temis, AFP, EURECOM, IBBT-MMLab and CINECA are involved in this task.

WP3.4 Trend Detection from Micro-blogging Feeds (*CINECA, month 6-36*)

Most of the existing Emerging Trend Detection (ETD) systems employ some text mining techniques for detecting topics and then monitor these topics over the time and define whether these topics are emerging or not (either applying machine learning techniques or relying on visualization techniques). Topic detection is a key point in the ETD process and is usually addressed by association analysis, cluster analysis or semantic proximity regions identification.

In NEED, we intend to leverage the knowledge infrastructure developed within WP2 to identify relevant topics. The rationale is that there are still few evidences in the literature for trend detection system using Semantic Web technology. In our approach, the semantic tags associated to each media item (texts and video) by the Event detection and annotation tool developed in WP3, together with the item timestamp, will provide the basis for trend detection. This approach will automatically filter the resulting trends from noise. Techniques for identifying real emerging trends among all discovered trends, will also be investigated. These techniques range from statistical analysis of relevance indicators, such as TF-IDF, over time, to applying time series trend analysis using both parametric and nonparametric methods, to time series clustering and classification. Statistical data processing in large scale data set, based on the Hadoop MapReduce framework, will be investigated, exploiting the distributed application architecture.

CINECA andTemis are involved in this task.

Deliverables

- D3.1.1 (Temis, month 12): Skill Cartridge™ of Named Entities
- D3.1.2.1: (Temis, month 18): Specifications for the Event Skill Cartridge™
- D3.1.2.2: (Temis, month 30): Event Skill Cartridge™
- D3.1.3 (Temis, month 36): Temis cross web browser toolbar
- D3.2.1 (EURECOM, month 12): Specification and Methodologies for multimedia news analysis
- D3.2.2 (EURECOM, month 24): Initial toolkit for multimedia news analysis
- D3.2.3 (EURECOM, month 36): Final toolkit for multimedia news analysis
- D3.3.1 (Temis, month 18): Initial report on event detection and annotation framework
- D3.3.2 (Temis, month 36): Final report on event detection and annotation framework
- D3.4.1 (CINECA, month 18): Initial report on trend detection from micro-blogging feeds
- D3.4.2 (CINECA, month 36): Final report on trend detection from micro-blogging feeds

Work package number	WP4	**Start date or starting event:**		M1		
Work package title	Semantic Multimedia News Interfaces					
Activity type	RTD					
Participant number	1	2	**3**	4	5	6	7
Participant short name	IBBT	EURECOM	**CWI**	AFP	ERCIM/ W3C	Temis	CINECA
Person-months per participant	0	0	**33**	6	0	0	3

Objectives

- To analyse user needs and extract requirements for designing innovative user interfaces;
- To design and develop multiple event-based interfaces for searching and exploring multimedia news content;
- To evaluate the usability of the user interfaces developed.

Description of work

The work will be carried out in three phases. Following a user-centered design approach, in the beginning of the project WP4 personal will assist WP1 in getting the user requirements explicit. In the second part of the project, the focus will shift to rapid incremental prototyping of a series of interfaces that deploy the functionality developed in the other technical work packages. Every prototype will be made available online ("release soon, release often") and exposed to selected users (from WP1's expert user panel) for informal testing and fast feedback loops. Prototyping will be based on the open source ClioPatria[66] platform, leveraging development work carried out in other European projects (e.g. Europeana, NoTube and PrestoPrime). Results from user experiences with the prototypes will be fed back as requests for improvements in functionality to the other technical WPs. In the final year, the focus will shift to more extensive and formal user testing and using the test results and other feedback to improve the overall user experience of complete system.

WP4.1 Analysis of user needs (*CWI, month 1-12*)

WP4 personnel will work jointly with WP1 on task WP1.2: Analysis of the user needs. The two main goals of this cooperation is (1) to support the analysis phase by providing concrete interface mock-ups, and (2) to bring designers and developers in direct contact with end-users as early as possible in the process. Confronting the expert user panel with concrete scenarios, paper-based and interactive mock-ups helps both to get more detailed and focused feedback from the users, and to test many design decisions before the actual development process. We also want to confront interaction designers with the ways users solve their problems with the current state of the art, i.e. our designers need to understand the techniques and workarounds users deploy to overcome the limited support of current tools.

- Support WP1 user interview phase by developing concrete scenarios and mock-ups;
- Support WP1 system analysis by describing currently used techniques and workarounds;
- Cooperate on translation of user study results into functional requirements for NEED.

CWI in cooperation with WP1 are involved in this task.

WP 4.2 Map user needs to functionality provided by NEED (*CWI, month 6-24*)

Based on the results of WP4.1, we will use rapid prototyping to develop interfaces that deploy the techniques developed in the other technical WPs, convey the detected contexts to the user and allow the user to interact with the information on a high, semantic level. We will develop short develop/release cycles and will request frequent feedback from WP1's expert user panel on intermediate designs and prototypes. While the final design decisions will depend on the outcome of the user requirements analysis, we anticipate to design and

[66] http://e-culture.multimedian.nl/software/ClioPatria.shtml

develop web-based user interface prototypes for various user personas/task combinations, including:

- Event-driven visualizations and map/time web mashups of news articles using the event models developed in WP2 and the events, locations and actors detected in WP3
- Interfaces that convey provenance, trust and other social network information using models developed in WP2
- Interfaces that deploy state of the art open web standards for "deep" linking into media fragments (WP3), generate charts and other visualizations of related quantitative background information

CWI in cooperation with WP2, WP3 and WP5 are involved in this task.

WP4.3 Evaluation and user testing (*CWI, month 24-36*)

Goal of the project is to use the results of 4.1 and 4.2 to develop a prototype that is sufficiently mature to subject to formal user evaluation experiments before the end of the second year. WP4 personnel will assist WP6 in the preparation of the evaluation workshops in Paris and Sophia Antipolis by checking the user surveys, experimental designs and test protocols for completeness, e.g. to ensure they contain sufficient detail to expect useful feedback for the developers (note that developers will explicitly **not** be directly involved with testing their own software).

- Paris evaluation (Month 24-30)
    - Provide feedback on testing materials and protocols developed by WP6
    - Cooperate with WP6 to prepare experimental website with the NEED prototype to be tested, implement extra logging facilities to gather more test data, develop online questionnaires if needed etc.
    - Translate test results into concrete change request and prioritization of the changes (with W1 and WP6)
    - Implementation of as many requests as possible before end of M30
- Sophia evaluation (Month 30-36)
    - Provide feedback on testing materials and protocols developed by WP6
    - Setup experimental system for testing second version
    - Translate test results into concrete change request and prioritization of the changes (with W1 and WP6)
    - Implement change requests for version that will be delivered at the end of the project
- Release end-of-project prototypes on NEED website (M36)

Independent of the WP8 exploitation plan and the deployment of NEED technology by the industrial partners, we expect to keep the NEED website with the open source UI prototypes and public event detection web services available for demonstration purposes after the project. NEED's UI development approach is to build on and extend existing open source software, especially for the more domain-independent software components. Just as NEED will build on the results of previous projects, we expect that the technological development of the UIs and other components will continue in future projects, both in the news and other application domains.

CWI in cooperation with WP6 are involved in this task.

Deliverables

- D4.1 (CWI, month 12): Report on user needs (complementary to D1.2.1 and D1.3.1)
- D4.2.1 (CWI, month 12): Report on functionality and expert panel feedback on the release of the first set of partial prototypes
- D4.2.2 (CWI, month 18): First public functional prototype ready for internal testing &

demo/dissemination purposes

- D4.2.3 (CWI, month 24): Second public functional prototype ready for user testing
- D4.3.1 (CWI, month 30): Report on user evaluation tests for the “Paris” workshop
- D4.3.2 (CWI, month 36): Report on user evaluation tests for the “Sophia Antipolis” workshop
- D4.3.3 (CWI, month 36): Final public prototype

Work package number	WP5	**Start date or starting event:**		M6		
Work package title	Framework Architecture Design					
Activity type	RTD					
Participant number	1	2	3	4	5	6	**7**
Participant short name	IBBT-MMLab	EURECOM	CWI	AFP	ERCIM/ W3C	Temis	**CINECA**
Person-months per participant	4	4	4	0	0	4	**28**

Objectives

- To implement and integrate the NEED services
- To design and provide the overall architecture taking advantage of input coming from layers WP2, WP3 and WP4 and requirements from WP1

Description of work

This work package will provide the core integration activities of the project. It will design and provide the overall architecture of NEED services based on the user and technical requirements coming from WP1. Once the components have been developed in WP2, WP3 and WP4, the integration phase will begin and allow for the completion of a working prototype of the NEED services.

The work to be carried out in this work package will evolve in phases as follows:

1. The first phase has the purpose of collecting the scope, goals, objectives and technical requirements of the different development work packages to characterize the different components of the system.
2. The second phase includes the design of the overall system's architecture. It will involve activities such as the selection and definition of schemas and ontologies needed for the communication among the components, interfaces and coordination patterns.
3. The third phase includes the definition and the development of integration test cases, and the coordination of the testing activities in order to ensure that the integration meets the requirements documented.
4. The last phase involves preparing the integrated NEED solution comprising the components that will be developed in the context of WP2, WP3 and WP4.

This work package is structured in three tasks and runs from the month 6 to the end of the project.

WP5.1 Infrastructure modeling and deployment (*CINECA, month 6-9*)

This task will model the test-bed infrastructure for supporting the deployment of the NEED services and their effective execution. The model will consider the need of the project to deal with a large amount of data distributed across different sites in Europe. The implementation of infrastructure will permit the execution of data-intensive applications and be reliable, efficient and able to scale in order to support larger data repository, if needed. The CINECA experience in implementing high performance infrastructure will support the modelling process and part of the existing infrastructure will be made available during the project.

CINECA is involved in this task.

WP 5.2 System architecture design (*CINECA, month 6-15*)

This task will analyse the system requirements in order to design the overall NEED architecture. In addition, a detailed implementation plan is developed, which will ease the development and integration phase. This implementation plan will include all system component definitions and interfaces along with the integration guidelines for component developers. As this task does not aim to build yet another platform, but providing a framework to integrate components, a survey on available technologies is essential for the design of the system architecture. This survey has to identify prominent market leaders, prominent used technologies in commercial and open source environment, communities and companies behind the tools and open source

projects, as well as formats and technologies. It will be performed by input from the use case, provided expertise and market surveys of the tool provider as well as Internet research in open source communities.

CINECA is involved in this task.

WP5.3 System integration (*CINECA, month 9-36*)

This task will tackle all aspects related with the framework and components overall integration.

CINECA will integrate the design of the system architecture based on its experience on large scale data storage, management, mining and processing, while all research partners will participate for the integration of technical modules.

To evaluate the approach of the implemented components as well as their usability and applicability to the use cases, this task will verify that the requirements are satisfied in correspondence to the project evaluation criteria and measure the performance obtained executing the applications.

CINECA in collaboration with all research partners will participate in this task.

Deliverables

- D5.1 (CINECA, month 9): Report on the NEED infrastructure design and system architecture
- D5.2 (CINECA, month 15): Report on the complete NEED implementation plan
- D5.3.1 (CINECA, month 12): Framework for integrated NEED services v1
- D5.3.2 (CINECA, month 24): Framework for integrated NEED services v2
- D5.3.3 (CINECA, month 36): Framework for integrated NEED services v3

Work package number	WP6	**Start date or starting event:**		M6		
Work package title	Evaluation					
Activity type	RTD					
Participant number	1	2	3	**4**	5	6	7
Participant short name	IBBT-MICT	EURECOM	CWI	**AFP**	ERCIM/ W3C	Temis	CINECA
Person-months per participant	6	2	6	**10**	0	0	0

Objectives

- To organize four public evaluation workshops (with participation of the expert user panel) to evaluate and test the NEED prototypes;
- To organize several iterative online consultations of the expert user panel to gather feedback on mock-ups and intermediary results;
- To establish metrics to measure the success criteria of the NEED technologies and interfaces to solve the problems of the end-users;
- To provide feedback from the end-users workshops to the partners in order to enhance the developments iteratively through the duration of the project, taking into account evolving needs of lead users as well as new services available on the web.

Description of work

To evaluate interim project results, four workshops will be organized to gather feedback from the expert user panel and especially the lead users identified in WP1.2 (see D1.3). The public workshops will be open for a broad range of stakeholders.

Each workshop will focus on a specific project milestone:

- The list of user requirements for NEED technologies and services (D.1.4 and D1.5);
- On the success criteria of the NEED approach based on user scenarios;
- On first evaluation of NEED prototype (both usability and success criteria);
- On second evaluation of NEED prototype (both usability and qualitative success criteria on NEED approach).

Consequently, the public workshops with the expert user panel will be held on M10, M18, M28 and M34. The user feedback received during the evaluation workshops will be summarized in strategic reports, that will be used as input to guide and refine further research activities within the different work packages

WP6.1 Workshop on list of user requirements (*IBBT-MICT, month 6-12*)

A workshop will be organized by IBBT-MICT in Ghent at the beginning of the project to gather feedback on the user needs and requirements identified through the user studies in WP1. Apart from representatives of the expert user panel, a selection of lead users, ranging from news experts and heavy news consumers over social scientists and researchers in the media field to communication students, will be invited to discuss on how NEED technologies can improve their experience on researching and exploring news on the internet. This workshop will allow to prioritize and refine the list of user requirements (as we assume that user needs and requirements can and will evolve over time).

IBBT-MICT, AFP and CWI are involved in this task.

WP6.2 Workshop on the success criteria of the NEED approach based on users scenarios (*IBBT-MICT, month 15-21*)

Intermediary project results will be discussed with the expert user panel during the second public workshop in Ghent. First, the elaborated use case scenarios (D.1.6) will be presented to the expert users to gather

feedback and extra ideas. In addition, the technical project partners will release first NEED mock-ups and prototypes to the expert user panel for feedback. The user feedback received during this workshop will be summarized in a strategic report that will serve as input for the technical work in the second half of the project. (Note: This feedback session complements the regular and more informal virtual/online panel consultations throughout the project, as foreseen in WP1).

IBBT-MICT, AFP and CWI are involved in this task.

WP6.3 Evaluation of NEED prototype (*AFP, month 24-36*)

The evaluation will be split in two phases:

- Phase 1 evaluation (M24-M30): first evaluation of NEED technologies on three month of multimedia news; this first evaluation will be conducted through a workshop in Paris (gathering in one or two sessions at least 50 end users, involving professional journalists, researchers and students in social sciences, journalism, as well as specialist of environmental issues) on a selection of subjects. A report on this evaluation will be published and will provide a first feedback to the partners about the acceptance of NEED proposal by end users.
- Phase 2 evaluation (M30-M36): second evaluation of NEED technologies on six month of multimedia news. This second evaluation will be conducted through a second (double) workshop held in Paris (France) and in Sophia Antipolis (France) (gathering in one or two session at least 50 end users, involving professional journalists, researchers and students in social sciences, journalism, as well as specialist of environmental issues) on a new selection of subjects. A report on this evaluation will be published and will provide a definitive feedback to the partners about the acceptance of NEED proposal by end users in order to refine the final NEED prototype.

AFP in cooperation with the user groups are involved in this task.

Deliverables

- D6.1 (IBBT-MICT, month 12): Strategic report of the results of workshop 1 on user needs
- D6.2 (IBBT-MICT, month 20): Strategic report of the results of workshop 2 on use case scenarios
- D6.3.1 (AFP, month 30): First evaluation report of the prototypes developed for the two use cases
- D6.3.2 (AFP, month 36): Second evaluation report of the prototypes developed for the two use cases

Work package number	WP7	**Start date or starting event:**		M1		
Work package title	Standardisation and Outreach					
Activity type	RTD					
Participant number	1	2	3	4	**5**	6	7
Participant short name	IBBT-MMLab	EURECOM	CWI	AFP	**ERCIM/ W3C**	Temis	CINECA
Person-months per participant	4	4	0	2	**16**	0	0

Objectives

- To disseminate the technical developments of the project across the broad research community and the professional media industry;
- To support the W3C in administering its *Video in the Web* related standard activities, their wide adoption in the European industry, and their evolution toward achieving efficient and interoperable communications within distributed software.

Description of work

This work package aims at fostering the development of new W3C Video Web standards with significant European participation and includes European outreach activities on future W3C Recommendations. The targeted audiences for this activity are:

- Individuals and organizations seeking to put their own video content in the Web and have become part of a linked Web of information in a variety of content types (text, images, audio, video);
- Media producers who want to ensure that their users are getting the best experience and that their content can be found on the Web;
- End-users who would like to watch and interact with online video content;
- Content aggregators dealing heavily with video content (e.g. news sites, some educational sites, etc) and who need formats that reliably support various requirements that they may be subject to, including captioning and video description.

This work package runs for the entire duration of the project and consists of two tasks:

- Technical activities, done through its European staff, will provide European research and industry with competent partners within the W3C community, providing guidance with the standardization processes currently in place at W3C through its working groups, studying the state of the art and monitoring the other standardization bodies in the domain, and helping to raise the level of European participation.
- Outreach activities to various communities (media, developers, industry). This outreach task serves to disseminate the results of the NEED project standardization work in W3C to a European audience. The dissemination plans include press releases relating to the "W3C Video in the Web" and NEED, talks by W3C and NEED team members at events as well as the development of outreach material such as information brochures and posters. The effectiveness of the press outreach efforts will be tracked by collecting press clippings that mention results of W3C Video work and NEED. The effectiveness of the talks will be measured by tracking and analyzing the type of attendees at NEED/W3C talks.

WP7.1 Standardization (*W3C, month 1-24*)

Video on the Web (and this includes audio, as the two are typically used together) has seen explosive growth, improving the richness of the user experience but leading to challenges in content discovery, searching, indexing and accessibility. Enabling users (from individuals to large organizations) to put video in the Web requires that we build a solid architectural foundation that enables people to create, navigate, search, link and distribute video, effectively making video part of the Web instead of an extension that doesn't take

full advantage of the Web architecture. The W3C Video in the Web Activity was created in August 2008 and is now composed of 3 working groups: Media Annotations, Media Fragments and Timed Text.

Within NEED, the W3C Media Annotations Working Group and the W3C Media Fragment Working Group are relevant to help standardizing the naming and description aspects of multimedia news content. Those W3C Working Groups will see their charter ending around mid 2010. The majority of the members of those groups are European. The support of the NEED project will increase the independence of W3C management and help continuing the work in those areas without fearing member pressure to divert that effort to other work items, and guarantee that W3C can stay committed to the development of standards pertaining to video technologies. The standards expected to be produced by this activity are under a Royalty-Free Patent Policy[67].

Furthermore, NEED will also actively participate in and contribute to other various standardization bodies such as the International Press Telecommunications Council (IPTC) and the European Broadcaster Union (EBU) with the expectation of having significant impact on their development. In the former, AFP is leading the Metadata Working Group and is an early adopter of standards such as NewsML-G2 and EventsML-G2 while the later has already decided to use these standards in the forthcoming EBU Core format. NEED aims thus to further develop multimedia news standards and provide and distribute reference implementations of them.

The activities conducted within this task can be summarized as follows:

- Plan exact milestones for the W3C Working Group tasks from the beginning of the "Last Call" period until the Recommendation phase;
- Support the work of a W3C Team contact that helps the group with W3C technology and process and provides publication support;
- Provide reference implementations of NewsML-G2 and EventsML-G2 within IPTC and EBU.

W3C, EURECOM and AFP are involved in this task.

WP 7.2 Outreach (*W3C, month 1-36*)

W3C is an international standardization body and most of its communication efforts are worldwide. However, as most of the team involved in that effort is based in Europe, there is a natural match in doing dissemination and outreach to European stakeholders. The central objective of this outreach task is to disseminate the results of the W3C Video in the Web Working Groups in European Industry: this is both to increase awareness and to invite EU industry participation in the Web standardization work dedicated to Video in the Web.

We plan to outreach to different types of audiences:

- To the EU industry: by the way of promotion/technical talks at industry track in conferences and/or workshops, in industry shows such as the (NAB Show, IBC, MipTV, MipCom, Gartner events, News Linked Data events, etc.)
- To the Web developers and content (video) producers: talking and/or participating in various foocamps, and developer-oriented EU conferences such as Paris Web[68], Over the air[69], etc.
- To the research community: talks at academic conferences (such as WWW conference series where W3C is used to have its own track)
- To the press: by distributing W3C press releases to the IT and large public publications, and by reaching out to analysts specialized in this domain
- To the European Commission stakeholders: attending the ICT coordination meetings and liaising

[67] W3C's new royalty-free patent policy, which strives to ensure that the technology contained in W3C specifications can be implemented without paying a license fee. Based on experience, royalty-free standards reach a high level of acceptance on the market far quicker than standards that bear royalties. The latter usually requires a lengthy negotiation process to determine 'reasonable' license fees, and may include long administrative delays due to the necessity of setting up patent pools between the many partners involved in the standardization effort.
[68] http://www.paris-web.fr/
[69] http://overtheair.org/

with other EU projects thematically identical.

W3C is involved in this task

Deliverables

- D7.1.1 (W3C, month 3): Working Groups Charter extension
- D7.1.2 (W3C, month 9): Working Groups Last Call Documents and Disposition of Comments
- D7.1.3 (W3C, month 12): Working Groups Candidate Recommendation Documents
- D7.1.4 (W3C, month 18): Working Groups Recommendations
- D7.2.1.x (W3C, month 8, 16, 24): Press releases (3): The W3C issues press releases on a regular basis. These press releases generally obtain a high level of attention and coverage in the European and international IST trade press. Several press releases will be part of the standardization activities in the NEED project. Three press releases are planned.
- D7.2.2 (W3C, month 36): 6 Press Clippings concerning the "W3C Video on the Web" and "NEED" results will be collected to measure the success of the outreach activities in this WP and be gathered within a report due at the end of the project.
- D7.2.3 (W3C, month 36): 10 Promotion Talks reporting on the results of the “W3C Video on the Web” to conferences (e.g. IST conference) and seminars on this topic that reach a European audience. These events tend to be attended by an audience that is not necessarily already involved in W3C activities. Therefore, these types of presentations are an excellent means to increase the awareness of European research and industry of W3C’s video standardization work. Ten of these talks are planned as part of the task WP7.2 and will be gathered within a report due at the end of the project.
- D7.2.4.x (W3C, month 12x): Communication Material, that includes material such as brochures, posters, branding, goodies and other similar material that help to increase the awareness of the Video in the Web work within the NEED project.

Work package number	WP8	**Start date or starting event:**		M1		
Work package title	Exploitation and Dissemination					
Activity type	RTD					
Participant number	1	2	3	4	5	**6**	7
Participant short name	IBBT	EURECOM	CWI	AFP	ERCIM/ W3C	**Temis**	CINECA
Person-months per participant	0	3	0	8	0	**8**	0

Objectives

- To build and maintain the NEED project web site and to push the project results to a very wide audience using all sort of social media tools;
- To bring some key results of the project to the market and to exploit them commercially beyond the end of the project following a precise marketing plan;
- To package and deploy an "Events Detection" Skill Cartridge;
- To package and deploy a "Contextualized visualization of news events" product.

Description of work

This work package deals with the dissemination and promotion of the project results and on the exploitation of the project resulting in a marketing plan and several products ready to enter the market. This work package runs for the entire duration of the project and consists of two tasks.

WP8.1 Project Dissemination (*AFP, month 1-36*)

An important tool of dissemination and promotion will be the NEED web site. Six-monthly Electronic Newsletters will act as an instrument for regular exchange of information with other related projects, the news industry and the research community at large. Finally, joint publications in international conferences (WWW, ISWC, CHI, ACM MM, SAMT, etc.) and journals will help to spread the scientific and technical results of the project among both the research community and media professionals.

All partners are involved in this task.

- WP8.1.1 The NEED web site (EURECOM, month 1-36)

The NEED web site will support delivery and spread of project promotional material such as information leaflets and electronic newsletter. The Web site will work as a portal, electronic archive or digital library of the community and store different types of documents: researchers' papers, public demonstrators, software, ontologies, etc. It will also include a wiki to support the collaborative work within the project. It will offer an RSS feed and use all social media tools to increase awareness of the project results.

- WP8.1.2 Six-monthly Electronic Newsletter (AFP, month 1-36)

The newsletter will report the main activities promoted and undertaken within the project and will be distributed to both the research community and the professional media industry.

WP8.2 Project Exploitation (*Temis, month 18-36*)

Metadata creation, especially in the area of "events detection" is a key element for content provider and many large organizations dealing with external and internal information. One of many potential derivative products that will came out of the NEED project is a semantic component "Events detection" Skill Cartridges. This valuable component will be very attractive for the market at large that needs to extract and wrap structured data out of an ever-growing amount of digital information published using web technologies.

Visualizing contextualized and personalized information is another derivative product that will come out of

the NEED project. Some of the key technologies developed will be packaged based on a Lucene[70] / SolR[71] implementation for enhancing the user experience on AFP news retrieval. This package will be used internally by AFP and, in case of success, could be marketed to the media industry as well as Digital libraries and documentary services.

Temis and AFP are involved in this task.

- WP8.2.1 Definition of a Marketing Plan and Market analysis (Temis, month 18-24)

We will focus on the creation of an operational marketing plan as well as a specific packaging of this knowledge component in order to serve the market. Among several benefits the marketing plan will strengthen the following:

  - Productivity gains: Computer-aided indexing and categorization, Creation and management of business knowledge bases based on text mining and ontology management system, Assistance for Knowledge Management data maintenance.
  - Flexibility: Assisting content selection for the publishing process, Monitoring content repurposing and repackaging, Retrieving information easily using metadata, thesaurus, subject, association, etc. Creating better connections between related contents in publication.
  - Deployment: Non-intrusive solution, Open, multilingual and standard based technology, Interoperability: XML, J2EE, API, Web Services, Knowledge representation: OWL, RDF, SKOS, Integration with Content Management Systems, Search Engines, Authoring tools.

- WP8.2.2 Package an "Events Detection" Skill Cartridge (Temis, month 24-36)
- WP8.2.3 Package a wrapper of key NEED technologies to enhance the user experience on AFP news retrieval (AFP, month 24-36)

Deliverables

- D8.1.1 (EURECOM, month 1 and continuously updated): Project web site fully functional
- D8.1.2.x (AFP, month 6x): Six-monthly electronic newsletter
- D8.2.1 (Temis, month 24): Definition of the Marketing Plan
- D8.2.2 (Temis, month 36): Skill Cartridge "Events Detection" product
- D8.2.3 (AFP, month 36): Contextualized visualization of news events product

[70] http://en.wikipedia.org/wiki/Lucene
[71] http://en.wikipedia.org/wiki/Solr

Work package number	WP9	**Start date or starting event:**		M1		
Work package title	Project Management					
Activity type	RTD					
Participant number	**1**	2	3	4	5	6	7
Participant short name	**IBBT-MMLab**	EURECOM	CWI	AFP	ERCIM/ W3C	Temis	CINECA
Person-months per participant	**18**	3	1	1	1	1	1

Objectives

- To oversee the coordination of the project;
- To assign responsibilities clearly at activity and sub-activity level;
- To define clear lines of communication among the participants;
- To specify a framework for self-assessment of the project and to monitor its activities;
- To handle all sort of legal, IPR and ethical issues.

Description of work

The management activities are designed to guarantee that the project runs smoothly by ensuring that the goals are clearly defined and understood. The proposed management structure of the project will be functional immediately after the project kick-off meeting and the activities will be undertaken during the whole duration of the project. Erik Mannens (IBBT-MMLab) will act as the Administrative Coordinator of the project while Raphaël Troncy (EURECOM) will be the Scientific Coordinator.

WP9.1 Overall Coordination and Communication with the EU Commission (*IBBT, month 1-36*)

It is the main activity and main task of the Management WP.

- Scientific quality: The progress of individual actions and achievements in the corresponding activities will be supervised by the WP leaders who report to the Project Steering Board (PSB). It will check and approve all deliverables before delivery to the EC.
- Financial: It will ensure that financial audits as required by the EU are being implemented in the project. Based on the accounting figures, future budgets will be planned and overspending will be limited.
- Project coordination and decision: The administration executive will ensure that relevant information on project progress and status is being exchanged among the members of the PSB. Using this information the PSB can quickly identify if changes are needed and efficiently decide what corrections or new actions should be implemented.

This activity will deliver six-monthly activity reports to the EC, containing scientific and technological progress, standardisation activities and financial expenditures for each partner.

All partners are involved in this task.

WP9.2 Planning and Coordination of Project and Boards Meetings (*IBBT, month 1-36*)

It offers logistic support for the planning and coordination of all project and boards meetings, including the kick-off meeting for a successful project start. The project aims to have three technical meetings per year in order to ensure the scientific and technical progress of the project.

It prepares and provides the means for successful annual project reviews as required by the EC. The administration executive will ensure that the major annual deliverables are submitted on time to the reviewer. Logistic organization of the review will be provided.

IBBT and EURECOM are involved in this task.

WP9.3 Project Evaluation (*IBBT, month 1-36*)

NEED will conduct internal and external project assessment based on effectiveness studies and peer-reviews.

- WP9.3.1 Definition and Specification of a Quality Plan (IBBT, month 1-6)

In order to evaluate the quality of the project, a Quality Plan will be defined. This plan will identify measures for assessment of the eventual level and degree of success in achieving the project objectives. This will form a framework for self-assessment of the project.

- WP9.3.2 Quality Control (IBBT, month 6-36)

This task will monitor the project objectives based on the Quality Plan defined previously. One specific activity will be devoted to periodically gathering evidence of the success (or otherwise) of the project activities. This evidence will be delivered to the Project Steering Board (PSB), so that it can feed back recommendations to shape the evolution of the network's activities. The quality assurance manager in charge of the validation is not part of the technical development team.

All partners are involved in this task.

WP9.4 Handling of Legal, IPR and Ethical Matters (*IBBT, month 1-36*)

It determines the way that partners behave with respect to each other according to the terms of the Consortium Agreement (CA). It will deal with the liability of partners, partner withdrawal procedures, the settlement of disputes, the responsibilities of partners regarding accurate and timely reporting of difficulties, confidentiality, including the difference between foreground and background information, IPR, including arrangements for licensing.

All partners are involved in this task.

Deliverables

- D9.1.x (IBBT, month 4x): Agendas and official minutes of relevant project meetings;
- D9.2.x (IBBT, month 6x): Six-monthly activity reports (scientific progress and project expenditures summary);
- D9.3.1.x (IBBT, month 12x): Financial audit reports at each annual review of the project;
- D9.3.2 (IBBT, month 6): Definition of the Quality Plan and the project success indicators;
- D9.3.3.x (IBBT, month 12x): Quality Report at each annual review of the project;
- D9.4 (IBBT, month 36): Status report on legal and ethical matters.

Summary of effort

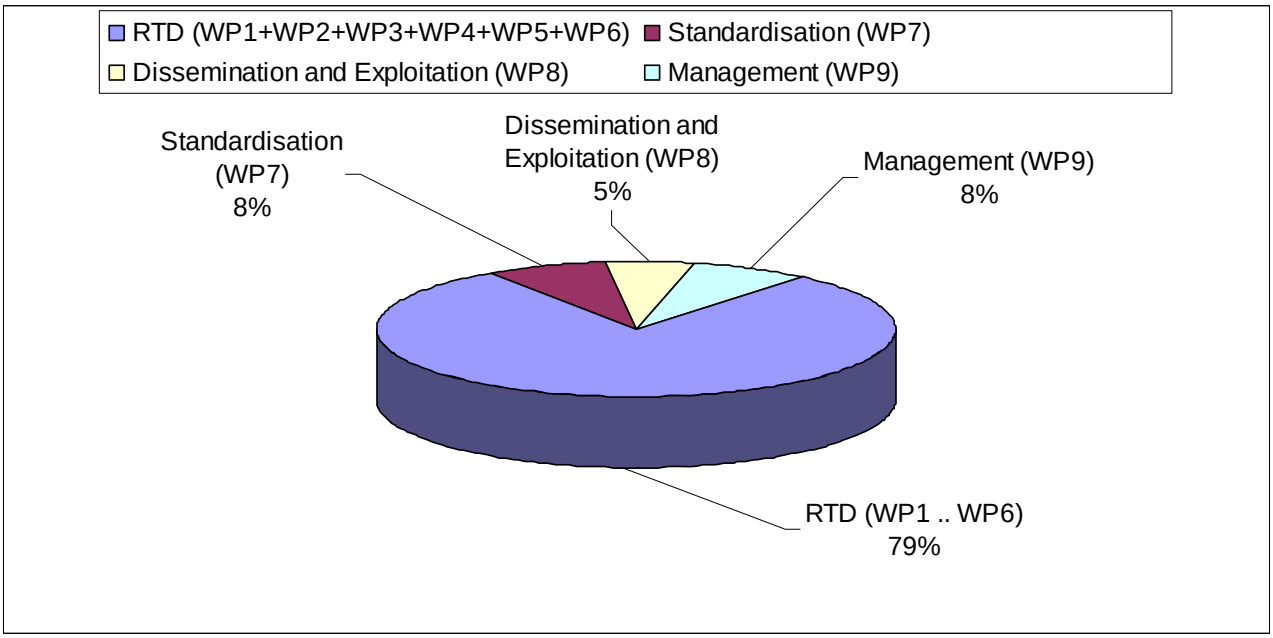
The **NEED** project duration is **36 months**, with a total effort of **346 PMs**.

The four core technical work packages (WP2, WP3, WP4 and WP5) are well balanced in effort, while the two user-oriented work packages (WP1 and WP6) highlight the user-centred approach advocated by the project to develop the NEED services. In particular, the project will pay a lot of attention to evaluate the various technologies developed during the project with various user groups and integrate their feedback into multiple virtuous loops of user requirements, development and evaluation. AFP will provide the core bulk of news data which will then be completed by the treasure trove of user generated content crawled on the web.

The important effort allocated in the standardisation work package (WP7) highlights the unique skills of this consortium and its determination to push forward some technologies for the benefit of all.

The overall allocation per partner is also well-balanced, while the WP leaders have been chosen on their knowledge and expertise in the field.

Partic. no.	Partic. short name	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	Total person months
1	**IBBT**	15	28	4	0	4	6	4	0	18	79
2	**EURECOM**	0	18	30	0	4	2	4	3	1	64
3	**CWI**	6	0	0	33	4	6	0	0	1	50
4	**AFP**	10	5	6	6	0	10	2	8	1	48
5	**ERCIM/W3C**	0	5	0	0	0	0	16	0	1	22
6	**Temis**	0	0	30	0	4	0	0	8	1	43
7	**CINECA**	0	0	8	3	28	0	0	0	1	40
Total		31	56	78	42	44	24	26	19	26	346



Section 2: Implementation

2.1 Management structure and procedures

The management structure of the project is classical and kept relatively simple while ensuring timely delivery of the scientific and technical results. The overall project organisation encompasses administrative, scientific, technical, evaluation and standardisation issues.

The consortium management of **NEED** is packaged as *WP9 - Project Management* with the following objectives:

- To guarantee the successful project completion within the agreed time, costs and quality requirements.
- To ensure compliance with EC standards and procedures for project management and tracking.
- To create and maintain effective channels of communication among the consortium partners, and to co-ordinate with other EU funded projects and other interested parties.
- To provide administrative and technical coordination, including financial, legal, contractual, and ethical management of the consortium.
- To provide quality assurance within the lifecycle of the project by monitoring all activities progress.
- To maintain regular contact with the European Commission.

All of the consortium partners are fully committed and agree to work together with the utmost co-operation for the timely fulfilment of their responsibilities. **We will encourage the mobility of scientific staffs** (exchange of researchers and PhD students, and visiting scientists) in order to maximize the scientific cooperation and the technological integration among the consortium partners.

2.1.1 Management structure

The management of the project is structured to address emerging issues swiftly and effectively. The key structures and roles in the project management structure are the following:

- The Project Steering Board (PSB)
- The User Groups (UG)
- The Standardisation Advisory Board (SAB)
- The Work Package Leaders (WPL)
- The Project Coordinator (PC)
- The Scientific Coordinator (SC)
- The Project Office (PO)

The interaction between them is shown in the figure below:

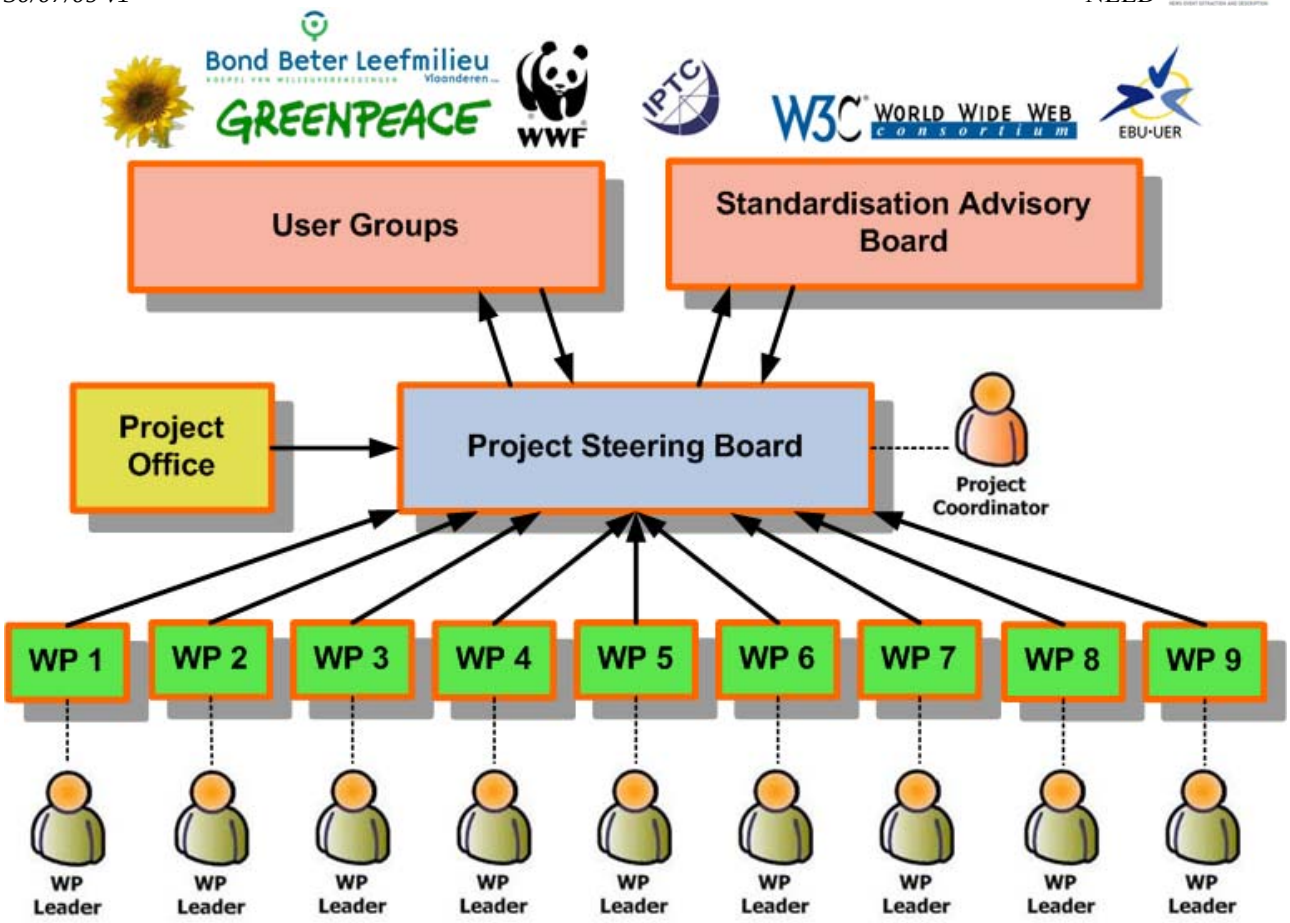


Figure 16: Project management structure

NEED will be coordinated by a *Project Coordinator* (PC) responsible for the administrative management and by a *Scientific Coordinator* (SC) responsible for the scientific progress of the project. The PC will be the reference and unique interface with the European Commission. The PC, **Erik Mannens** is an experienced project manager having successfully coordinated more than 10 National and International research projects within IBBT, while the SC, **Dr. Raphaël Troncy** has gained a lot of experience leading tasks and work packages of EU projects within the various institutes he has visited. Erik Mannens and Raphaël Troncy are used to work together: in particular, they co-chair together the W3C Media Fragments WG since September 2008.

Each work package is managed by one of the partners who will appoint a *Work Package Leader* (WPL). The *Project Steering Board* (PSB) consists of one representative per partner and is chaired by the Project Coordinator. Each member of the PSB will be able to make decisions as to the partner's particular technical interests and how to use the resources allocated for the project. They will have also the authority to make decisions on behalf of his or her company in terms of overall strategy and resources allocated to the project.

The PSB and the Project Coordinator will be assisted by the *Project Office* (PO) that is located at the coordinator, IBBT. The Project Office will be primarily responsible for the administrative management, such as: taking care of payment delivery to the partners; controlling the financial reports from the individual groups; and obtaining audit certificates from each participant. It will also provide the infrastructure and maintain the collaborative working tools for internal communication, such as: mailing list, wiki, subversion and project document repositories.

Given the strong standardisation activity targeted by the project, a *Standardisation Advisory Board* (SAB) has been set up. This advisory board, external to the Project core members, is strongly linked to all research, development, integration and dissemination activities of the project. The mission of the SAB is to ensure that the output of the research becomes a valuable resource for industrial innovation. The SAB will give advices on strategic positioning of the project activities. One of its main roles is to actively encourage timely feedback from the industrial members belonging to each standardisation body.

Finally, specific *user groups* will be set up in order to evaluate the technologies developed in NEED. Several organizations have already committed to participate in the project including the EU Green Party, the Belgium Green Party and the Flemish *Bond Beter Leefmilieu* organization (see **Annex A**) while the participation of other non governmental organizations is pending at the time of writing of this project proposal. It is part of a task within WP1 to collect more user groups that will provide continuous guidance in the project. Specific budget for user groups gatherings have been planned.

2.1.2 Roles and Decision-making Bodies

The following roles will be assigned in the project:

- ***Project Coordinator*** (PC) is responsible for the whole project operation and communication with external bodies. The PC represents the project and the consortium, reports to the Commission, monitors overall consortium performance, administers project resources, promotes project visibility and promotes dissemination of project results. The PC is also responsible for auditing technical performance of the project, ensuring that the technical objectives are being met, and that required information is exchanged among the different work packages.
- ***Scientific Coordinator*** (SC) is responsible for the overall scientific coordination and direction taken by the project.
- ***Work Package Leaders*** (WP Leaders) are responsible for each work package, including the effective coordination and cooperation between tasks and across work packages. The WP Leaders will be appointed by the lead participant for the corresponding work package. Each lead participant has been carefully chosen given its strong interest and expertise in the work to be conducted:
    - WP1 (Use Case Scenarios): IBBT-MICT,
    - WP2 (Knowledge Infrastructure for News Integration): IBBT-MMLab,
    - WP3 (Events Detection and News Enrichment): EURECOM,
    - WP4 (Semantic Multimedia News Interface): CWI,
    - WP5 (Framework Architecture Design): CINECA,
    - WP6 (Evaluation): AFP,
    - WP7 (Standardization and Outreach): ERCIM/W3C,
    - WP8 (Exploitation and Dissemination): Temis and
    - WP9 (Project Management): IBBT-MMLab.
- ***Task Leaders*** are in charge of the realization of specific tasks. They will activate the partners assigned to the task, report to WP Leaders regarding the progress of specific work, and provide the necessary input to other inter-dependent tasks.

Additionally, the following bodies will be involved in the management of the project:

- ***Project Steering Board*** (PSB) is the high-level representation and management body of the project, co-chaired by PC and SC, and composed of one representative from each partner. This body is responsible for all formal decisions regarding the strategic directions of the project, such as relations with the European Commission, policies for promotion and exploitation of results, administrative arrangements. The PSB will meet at least three times per year.
- ***User Group*** (UG) is already composed of the following members:
    - Danny Jacobs, director of *Bond Beter LeefmilieuVlaanderen vzw (BBL)*
    - Bart Staes, for the European Green Party
    - Tinne van der Straeten, for the Belgium Green Party

  and will be completed with more users during the project.
- ***Standardisation Advisory Board*** (SAB) has been already set and is being consulted from the very outset of proposal building. The current members of the board are Ivan Herman (W3C), Michael Steidl (IPTC) and Jean Pierre Evain (EBU). Members of the project consortium are already involved in these standardisation bodies where its participants co-chair specific working groups.

Ivan Herman is the Semantic Web Activity[72] Lead at W3C. He is a member of IW3C2 (International World Wide Web Conference Committee), the committee coordinating the yearly WWW conference series where he serves as a liaison for W3C, and of SWSA (Semantic Web Science Association), the committee responsible for the International Semantic Web Conferences series.

Michael Steidl is the Managing Director of IPTC, the consortium of the world's major news agencies, news publishers and news industry vendors. He is chief executive at the international association of news providers and news industry system vendors and he is co-chairing the special interest group regarding all photo metadata issues of IPTC, and responsible for the organisation of the yearly International Photo Metadata Conference[73].

Jean-Pierre Evain is a Senior Engineer at EBU. He is the secretary of the EBU Project Group P/MAG – Metadata Advisory Group and the official liaison with IPTC. He was co-chair of the working group in charge of developing the TV-Anytime[74] Metadata Specification.

2.1.3 Procedures

Reporting and Deliverables:

Internal Bi-Monthly Activity Reports (MAR) will be produced every two months by the WP Leaders, based on the information collected from the Task Leaders. These reports will describe the present status of developments, and indicate the specific technical work undertaken in the corresponding period. The PSB will analyse these reports and take the necessary actions. They will finally be used for producing the annual review reports, including possible recommendations made by the PSB.

External Six-Monthly Management Reports (MMR) will be produced every six months by each of the consortium partners, and delivered to the PC who will combine them into a unified MMR delivered to the European Commission. The PC will revise the reports in order to ensure their consistency and completeness. The reports will encompass all project activities executed by the partners in the corresponding period, such as research and development activities, internal meetings attendance, conference and workshops participation, deliverables, expenses, potential deviations from the project plan.

All the reports will be uploaded to a collaborative space allowing all partners to read and comment them. Project deliverables will be provided by the Task Leaders, revised by the WP Leaders, inspected by the PC to ensure their consistency and completeness, and delivered to the European Commission by the PC in the agreed format.

Meetings:

Even though partners have already established a collaborative and cohesive working relationship, they will meet regularly and at least three times per year to present and discuss the scientific and technical progress of the project. These consortium plenary meetings will also include a scientific talk, selected by the PSB and highlighting the research developed in the project, and will host the Project Steering Board meetings.

Additional technical meetings, requested by the Project Coordinator or the WP Leaders and covering particular technical aspects of the project can be organised on demand. **NEED** will encourage the mobility of scientific staff (exchange of researchers and PhD students, visiting scientists) and virtual meetings through videoconferencing to ensure maximum interaction and exchange between the partners.

[72] http://www.w3.org/2001/sw
[73] http://www.phmdc.org/
[74] http://www.tv-anytime.org/

Tools and Instruments:

The management structure will make extensive use of the following tools and instruments in order to facilitate the coordination of the project and enhance communication between partners:

- *Email reflectors* and *archived mailing lists* for communications within the whole project and within each work package or subsequent ad-hoc interest groups. Regular audio-conference and video-conference will be organised among the project partners.
- A *semantic wiki* coupled with a *tracker* will provide the collaborative working space for monitoring the project activities and action points and for storing the accumulated knowledge and key decisions. It will allow consortium partners to publish and share information, meetings agenda and minutes.
- A *version control software* will be used to maintain current and historical versions of files such as source code, web pages, and documentation developed in the project.
- The project *web site* will be the main dissemination tool for the project. It will act as a document repository and will be regularly updated to disseminate the findings, publications and software products of the project.
- The project will finally create various accounts on different *social media sites* in order to better disseminate its results.

Consortium Agreement:

A Consortium Agreement (CA) following the template and best practices from existing projects will be agreed between all consortium participants.

Conflict Resolution and Decision Making Process:

Attempts will be made to resolve conflicts as soon as possible after they are identified. The WPL in cooperation with the PC and the SC will be responsible for resolving conflicts occurring under the work package s/he is leading. They will employ a problem solving approach in order to achieve consensus, striving for a win-win outcome for the conflicting parties.

If conflicts cannot be resolved at that level, the PSB will be asked to intervene and possibly vote. The consensus rule will predominate but a majority vote might occur if a decision cannot be reached. Veto rights are in any case granted to every consortium partner, in which case resolution is deliberated in consultation with the commission.

Conflict resolution, voting and decision making processes will be clearly defined in the Consortium Agreement to be agreed and signed off by the PSB that will safeguard the smooth execution of the project.

Risk Assessment:

The **NEED** risk management plan will be produced on the basis of existing risk management practices. The plan will report risk identification, analysis and mitigation strategies for the project. Along these lines, the following table outlines the main risks identified in the context of the project, their possible impact and some proposed solutions and mitigation strategy. This contingency plan will be further elaborated during the first months of the project lifetime, and reviewed at the end of the first reporting period.

Risk	Measure	Corrective action
Risks stemming from the multidisciplinary nature of partners	Most of the partners of the project are experienced leaders in their field of research and industrial sector and have worked together before.	The project management will revise its structure if necessary and ensure smooth communication between technology providers, academia and users.
Co-ordination problems and disputes among partners	Communication flow strategy clearly established: main vehicle for information exchange within the project are working papers, project meetings, telephone conferences, e-mail and tool suites recording decisions and tracking issues. The decision and conflict resolution procedures are specified in the quality plan and agreed upon.	The PSB will analyse the situation, possibly with the help of external experts, and decide how to proceed after having reached an agreement through discussion, or democratic voting.

Risk	Measure	Corrective action
Legal controversies among partners	Usage of individual foreground technology and knowledge will be regulated at the beginning of the project, and included in the Consortium Agreement.	Confront the involved partners with the established procedures. If necessary, produce separate non-disclosure agreements.
Underestimation of effort needed to produce deliverables and complete activities	To ensure the successful completion of the activities and the validity of their results, each work package contains planning of work, validation and quality assurance activities. WP leaders are responsible for timely completion of activities – project and technical management ensure timely submission of deliverables.	The management structure will closely monitor resource and budget consumption in order to take corrective actions wherever necessary.
Technical risks related to shifts in standardization efforts or the appearance of a disruptive technology	The PSB reviews technical and research aspects of the project and controls technical activities and directions. It is in constantly liaison with the SAB for the development of new standards in the field.	The project consortium will be pro-active in standardizing and promoting new technologies.
Uptake of results	To ensure that the results of the project are accepted by the research and development community and deployed in the media industry, NEED includes strong industrial exploitation partners – AFP, Temis and ERCIM/W3C. It is in constant liaison with the SAB that reflects the current need of the market.	An exploitation plan will be set up at the beginning of the project and will be revised during the lifetime of the project.
Focusing on technical challenges instead of addressing real user needs	The project consortium combines technical and user-oriented partners that will work in close collaboration in each work package. We will establish a close loop of the resulting applications to test users (AFP journalists and clients, user groups).	Revise the scientific and technological objectives of the project using the results and lessons learned after the first two user evaluations scheduled at months 12, and 24.

Figure 17: Risks and corrective measures

It should be noted that in addition to this plan, the Project Management team will implement a comprehensive Risk Management procedure (including technical and non technical issues) to address any unexpected risk in the project. The latter will be clearly defined in the technical annex of the project.

2.2 Individual participants

2.2.1 Interdisciplinary Institute for Broadband Technology (IBBT), Belgium

Expertise:
IBBT, the Interdisciplinary institute for BroadBand Technology, is a research institute founded by the Flemish Government, focusing on information & communication technology (ICT) in general, and applications of broadband technology in particular. The IBBT was founded as a virtual research centre (over 500 researchers), based on research teams from existing knowledge centers. Two research groups from IBBT will participate in NEED, namely the *Multimedia Lab* and the *Media and ICT* research group.

Multimedia Lab[75] (MMLab) is a research group within Ghent University (Department of Electronics and Information Systems) founded in 2001. This lab accounts for about 25 researchers and has a wide range of activities, including fundamental/basic research, applied research, and contract-based research with industrial partners. Besides, MMLab is doing scientific consultancy for both industrial and governmental partners and is one of the founding research groups of IBBT. MMLab is very active within MPEG and W3C standardization, via the submission of technical contributions, by chairing several ad-hoc groups, and through the editorship of several specifications. The main research topics that are dealt with by MMLab can be summarized as follows: Video coding and compression; Image/video processing and analysis; Multimedia content adaptation; Multimedia technology; Gaming technology; Ontology engineering; Standardization in the domain of multimedia applications and systems. MMLab has been a partner in more than 50 research projects (FP5, FP6, IWT, FWO, IBBT and bilateral).

Media and ICT (MICT) is a research group within Ghent University (Communication Department) established in 2004. MICT counts 17 members and conducts fundamental social research as well as applied and governmental research in the field of new information and communication technologies. MICT's research is not just focussed on the end user, but also incorporates the vision and strategies of the other actors involved in the innovation process: hard- and software developers, service providers, content providers, media institutions, the publicity sector, policy makers, etc. Hence MICT approaches its research from a three-way perspective: user, policy and business. Central to the work of MICT is its social-scientific research experience. MICT uses its expertise in both qualitative and quantitative research methodologies in all phases of the innovation development process. The MICT research activities mainly focus on five clusters whereby the strong methodological expertise functions as a base: Media Production & Distribution; User Experience & Behaviour; Profiling & Targeting; ICT & Society; Gaming. The combination of these fields of expertise results in a mix of applied and fundamental research funded by Flemish, Federal and European institutions. Being one of the core research groups of the Interdisciplinary Institute for Broadband Technology (IBBT), the research activities of MICT have a strong interdisciplinary character.

Role in the project:
IBBT will bring expertise in social communication between Media and ICT and will therefore lead **WP1: Use Case Scenarios**. It will also bring expertise in ontology engineering and lead **WP2: Knowledge Infrastructure for News Integration**. Its expertise in applying user studies and semantic web technologies in the news domain within the PISA project[76] will provide solid foundations for NEED. Finally, IBBT will be the administrative coordinator of the project, thus be responsible of **WP9: Project Management**.

Key Person:
Ing. Erik Mannens (MMLab) is a project manager at IBBT since 4 years where he has successfully manages more than 10 projects. He received his Master's degree in electro-mechanical engineering (1992) at KAHO Ghent and his Master's degree in computer science (1995) at K.U. Leuven University. His major expertise is centred on broadcasting, iDTV and web development. He is involved in several projects as a senior researcher; he is co-chair of the W3C Media Fragments Working Group and actively participating in other W3C's semantic web standardization activities. He is also member of the technical committee of ACM Multimedia, and SAMT.

[75] http://multimedialab.elis.ugent.be
[76] https://projects.ibbt.be/pisa

Dr. Davy Van Deursen (MMLab) received his M.Sc. and Ph.D. degrees in computer science from Ghent University, Belgium in 2005 and 2009, respectively. He is currently a post-doctoral researcher at Ghent University - IBBT (Ghent, Belgium), where he is working for the Multimedia Lab research group. Between 2007 and 2009, he was involved in Siruna[77], which is a spin-off of IBBT and Ghent University and offers a platform that serves as a thin client mobile application gateway. Since 2008, he is also actively participating within the W3C Media Fragments Working Group, whose mission is to specify a URI scheme for the addressing of media fragments. His research interests and areas of publication include video coding, media content adaptation and delivery, media content presentation, Semantic Web technologies, standardization, and multichannel publishing.

Dr. Ing. Chris Poppe (MMLab) received a Master degree in Industrial Sciences from KaHo Sint-Lieven, Belgium, in 2002 and received a Master degree in Computer Science from Ghent University, Belgium, in 2004. In 2004, he joined the Multimedia Lab research group within IBBT, where he obtained the Ph.D. degree in 2009. His research interests include video coding technologies, video analysis, and multimedia metadata extraction, processing and representation, with a strong focus on standardization processes.

Prof. Rik Van de Walle (MMLab) received his M.Sc. and Ph.D. degrees in Engineering from Ghent University, Belgium in 1994 and 1998, respectively. After a visiting scholarship at the University of Arizona (Tucson, USA), he returned to Ghent University. In 2001 he became a professor at the Department of Electronics and Information Systems (Ghent University-IMEC, Belgium) and founded the Multimedia Lab. Rik Van de Walle has been/is editor of the following MPEG specifications: MPEG-21 Digital Item Declaration Language; MPEG-21 Digital Item Processing; MPEG-21 Digital Item Processing - Technologies under Consideration; and MPEG-21 Reference Software. Rik Van de Walle has been involved in the organization of and/or review of papers for several international conferences (e.g., IEEE ICME, WIAMIS, ISAS-SCI, ACIVS, Mirage, EUROMEDIA-Mediatec). His current research interests include multimedia content delivery, presentation and archiving, coding and description of multimedia data, content adaptation, interactive (mobile) multimedia applications, and interactive digital TV.

Dr. Laurence Hauttekeete (MICT) received her Ph.D in Communication Sciences from Ghent University, Belgium, in 2004. In her Ph.D she looks at tabloidisation tendencies in the Flemish press, the development of a measurement model. Her research interests include the printed and audio-visual media, media economics, new media and qualitative and quantitative research methods. She joined MICT (media and ICT) in 2005 as a senior researcher. Most of the projects she works upon are situated in a few domains: ‘digitisation’, ‘culture and media’, ‘technology and youngsters/education’, and ‘ICT and government authorities’.

Prof. Steve Paulussen (MICT) is senior researcher of the IBBT research group for Media & ICT (MICT) at Ghent University, Belgium. In 2004, he obtained his PhD with a thesis on the impact of the Internet on journalistic newsgathering and news production. Since then, he is working as a senior researcher at MICT on different research projects that relate to issues of media production, media experience and media use in today's 'convergence culture' . His main research interests are in the field of journalism studies and (online) news production. From 2000 to 2006, Steve was involved in the European COST A20 project on 'The Impact of the Internet on the Mass Media'. Further, he was a member of the research consortium of the EU FP6 project 'Adequate Information Management in Europe' (AIM) (2005-2008). Currently, Steve is part of an international team of 8 researchers who investigate participatory journalism practices in 10 different European countries and the US, and on a national level, he participates in an inter-university multidisciplinary strategic research project on 'Flemish E-publishing Trends' (FLEET[78]). Finally, Steve Paulussen is a member of the Center for Journalism Studies (CJS) at Ghent University and since 2008 a part-time lecturer in journalism theory at the Vrije Universiteit Brussel (VUB).

[77] http://www.siruna.org/
[78] http://www.fleetproject.be

2.2.2 EURECOM (EURECOM), France

Expertise:
Eurecom, Sophia Antipolis, France, is a graduate education and research center, funded by two schools: Telecom ParisTech (France) and EPFL (Lausanne, Switzerland), with several academic and industrial members. The research activity is organized in three themes: mobile, corporate and multimedia communications. The Multimedia Communications Department research topics include multimedia analysis, signal processing, information theory, speech processing, biometry, and multimedia semantics for the social semantic web. We have a very active collaboration program, and participate in many projects at the national (Argos, RPM2) and European (STATION, GM4iTV, PorTiVity) level.
We have extensive expertise in research related to multimedia processing and retrieval, and focusing on the challenges of automatic indexing, multimodal analysis and semantic description for exploring large collections of media content. EURECOM has participated in the EU Network of Excellence K-Space[79], which aims at narrowing the large disparity between the low-level descriptors that can be computed automatically from multimedia content and the richness and subjectivity of semantics in user queries and human interpretations of audiovisual media – the so-called Semantic Gap. Our group is a regular participant in the TRECVID evaluation campaigns and has recently organized the 15th Multimedia Modeling international conference (MMM 2009) in January 2009.

Role in the project:
EURECOM will act both as scientific coordinator and as an expert in multimedia semantics in the **NEED** project. EURECOM will bring expertise in low-level and high-level multimedia analysis and will more generally lead the main technical WP of the project, **WP3: Event Detection and News Enrichment**. It will also contribute to the project through its expertise in modelling multimedia ontologies (within WP2). Finally, it will play an active role in WP7, co-chairing already a relevant W3C Working Group.

Key Person:
Dr. Raphaël Troncy is an assistant professor at EURECOM since 2009. He obtained with honours his Master's thesis in Computer Science at the University Joseph Fourier of Grenoble (France), after one year spent in the University of Montreal (Canada). He benefited from a PhD fellowship at the National Audio-Visual Institute (INA) of Paris where he received with honours his PhD from the University of Grenoble (INRIA/INA) in 2004. He selected as an ERCIM Post-Doctorate Research Associate 2004-2006 where he visited the National Research Council (CNR) in Pisa (Italy) and the National Research Institute for Mathematics and Computer Science (CWI) in Amsterdam (The Netherlands). He was a senior researcher for CWI from 2006 till 2009. Raphaël Troncy is co-chair of the W3C Incubator Group on Multimedia Semantics and the W3C Media Fragments Working Group, contributes to the W3C Media Annotations Working Group and actively participates in the EU K-Space Network of Excellence. He is an expert in audio-visual metadata and in combining existing metadata standards (such as MPEG-7) with current Semantic Web technologies. He works closely with the IPTC standardisation body on the relationship between the NewsML language family and Semantic Web technologies.

Dr. Benoit Huet received his BSc degree in computer science and engineering from the Ecole Superieure de Technologie Electrique (Groupe ESIEE, France) in 1992. In 1993, he was awarded the MSc degree in Artificial Intelligence from the University of Westminster (UK) with distinction, where he then spent two years working as a research and teaching assistant. He received his DPhil degree in Computer Science from the University of York (UK) for his research on the topic of object recognition from large databases. He is currently Assistant Professor in the multimedia information processing group of Eurecom (France). His research interests include computer vision, content-based retrieval, multimedia data mining and indexing (still and/or moving images) and pattern recognition. He has published over 80 papers in journals, edited books and refereed conferences. He is a member of IEEE, ACM and ISIF. He has served in many international conference organization and technical program committee. He is regularly invited to serves as reviewer for prestigious scientific journals as well as expert for project proposal at national, European and International level. He is the conference chair of the International Conference on Multimedia Modelling (MMM'2009) which took place in Sophia-Antipolis (France) in January 2009.

[79] http://www.k-space.eu

2.2.3 Stichting Centrum voor Wiskunde en Informatica

Expertise:
The Interactive Information Access[80] group at the Centre for Mathematics and Computer science (CWI) carries out research on improving models and tools for presenting multimedia information to end-users on a variety of platforms. CWI is the research institute for mathematics and computer science research in the Netherlands. CWI's mission is twofold: to perform frontier research in mathematics and computer science, and to transfer new knowledge in these fields to society in general and trade and industry in particular. CWI has always been very successful in securing considerable participation in European research programs and has extensive experience in managing these international collaborative research efforts. CWI is also strongly embedded in Dutch university research: about twenty of its senior researchers hold part-time positions as university professors and several projects are carried out in cooperation with university research groups. In addition, CWI has strong links to the World Wide Web consortium, and houses the Benelux office. CWI has a staff of 210 fte (full time equivalent), 160 of whom are scientific staff. CWI operates on an annual budget of EURO 13M.

Role in the project:
CWI will bring its expertise in developing and evaluating end-user interface technology for semantic multimedia applications. In particular, it will build upon the successful open source ClioPatria platform, leveraging development work carried out in other European projects such as Europeana, NoTube and PrestoPrime. CWI will focus on demonstrating added value of semantic technology to end-users by developing and evaluating prototype interfaces to access, query, and explore large and heterogeneous multimedia news repositories, leading **WP4: Semantic Multimedia News Interfaces**.

Key Person:
Prof. Lynda Hardman received her PhD from the University of Amsterdam in 1998. From the eighties, Hardman has been working on user interfaces for hypertext, multimedia and hypermedia browsing and authoring systems. Her current research efforts are focused on improving design methods for human interaction for emerging technologies, with specific projects in annotated media repositories. She is professor of Multimedia Interaction at the University of Amsterdam. She recently co-edited a special issue of the Multimedia Systems Journal on the canonical processes of media production, and a special issue for IEEE Intelligent Systems on AI and Cultural Heritage.

Dr. Jacco van Ossenbruggen received his PhD from VU University Amsterdam in 2001. He has been working on structured hypermedia documents on the Web and intelligent user interfaces for heterogeneously annotated media repositories. He is an expert in integrating large cultural heritage data sets and played a key role in developing the award winning MultimediaN E-Culture Demonstrator. He is currently active in the EuropeanaConnect project, where he works on the semantic layer for the Europeana cultural heritage search engine, and on the PrestoPrime project, working on sustainable end-user access to large audiovisual archives. He co-edited a special issue for IEEE Intelligent Systems on AI and Cultural Heritage and is an assistant professor with the Web & Media research group at the VU University. Jacco was an active member of the W3C SYMM WG during the development of SMIL 1.0 and 2.0, of the W3C Semantic Web Best Practices and Deployment WG and a founding member of the W3C incubator on Media Semantics.

[80] http://www.cwi.nl/en/research-groups/Interactive-Information-Access

2.2.4 Agence France-Presse (AFP), France

Expertise:
Agence France-Presse (AFP) is a global news agency, delivering fast, accurate, in-depth coverage of the events shaping our world from wars and conflicts to politics, sports, entertainment and the latest breakthroughs in health, science and technology. With 2,900 staff and stringers of 80 nationalities, spread across 165 countries, AFP covers the world 24 hours a day in six languages, delivering the news in video, text, pictures, multimedia and graphics. AFP produces roughly 5.000 text dispatches, 2.000 photos, 80 graphics and 50 videos per day. AFP's photo database, called ImageForum, contains more than 9 millions pictures, including partnerships.

AFP participates in R&D projects through its Medialab unit which is currently involved in the Papyrus European project as well as two French funded R&D projects, Scribo and RMM2. AFP's Medialab, created in 2000, is a small unit of 5-6 persons (2-3 engineers, 3 journalists) dedicated to research and development (both in products and processes) aimed at boosting innovation inside the company, especially in the areas of semantic web technologies, image retrieval, mashups, content structuring and management.

Role in the project:
AFP is the news content provider of the project and will provide large corpus of textual news stories in French, English, Dutch and other languages, info-graphics, pictures and videos. AFP will also contribute their major experience in modelling the domain of news, as they are a key member and chair of the IPTC. AFP will finally be responsible for the evaluations conducted within the project, providing the technologies developed in **NEED** to AFP journalists and bringing back their feedback, and therefore leading **WP6: Evaluation**.

Key Person:
Denis Teyssou is the head of Agence France-Presse Medialab, AFP research and development unit. Denis has been a journalist covering all aspects of news for more than twenty years at AFP, mainly in Madrid, Spain, where he was Deputy Director as well as news and online editor during his latest stay from 2000 till mid-2004. He has been involved in multimedia content managing system and taxonomy implementation in several languages at the Technical editor in chief department in Paris from 2004 till 2007, before joining the Medialab. Denis, an open source computer geek, has earned a webmaster degree at the CNAM institute in Paris in 2006 where he has been studying also innovation management. He previously attended post-graduate studies in Information sciences in Bordeaux 3 University and earned two Paris XI University degrees in Astronomy and Astrophysics.

2.2.5 GEIE ERCIM (ERCIM/W3C), France

Expertise:
ERCIM, the European Research Consortium for Informatics and Mathematics, aims to foster collaborative work within the European research community and to increase cooperation with European industry. The members of ERCIM include leading research establishments from 18 European countries. Encompassing over 10,000 researchers and engineers, ERCIM is able to undertake consultancy, development and educational projects on any subject related to its field of activity. ERCIM was founded in 1989 and is a European Economic Interest Grouping (EEIG). ERCIM is the parent organization for the European part of the W3C (World Wide Web Consortium).

The World Wide Web Consortium (W3C) is an international consortium of over 400 members worldwide from research and industry where Web standards and guidelines are developed to ensure a universal access to One Web. Founded in 1994 by the inventor of the Web, Tim Berners-Lee, W3C has successfully overseen processes of issue raising, design, consensus building and testing resulting in over 120 technical standards that make the Web work. Success stories include standards for HTML, XML, XHTML, Cascading Style Sheets, VoiceXML, Web Accessibility, Web Services, Semantic Web, mobile Web, graphics, video and many other areas of technology that are designed to enable all Web users, from the elderly to the young generations, to share knowledge, expand commerce, and innovate.

Role in the project:
The World Wide Web Consortium's main role is to standardize Web technologies by creating and managing Working Groups that produce specifications (called "Recommendations" or "Web Standards") that describe the building blocks of the Web, and produce them freely available to all, as part of the Web open platform.

The project will be supported out of the Video in the Web Activity of the W3C. This W3C activity was launched in August 2008 to make video a "first class citizen" of the Web. Video on the Web (and multimedia in general) has seen explosive growth, improving the richness of the user experience but leading to challenges in content discovery, searching, indexing and accessibility. Enabling users (from individuals to large organizations) to put video in the Web requires that we build a solid architectural foundation that enables people to create, navigate, search, link and distribute video, effectively making video part of the Web instead of an extension that doesn't take full advantage of the Web architecture. In this project, the W3C will be looking for opportunities to develop Video related standards and/or guidelines from ideas that are discussed during the project. It will also commit to development of practical work items associated with the project. ERCIM/W3C will lead **WP7: Standardization and Outreach**.

Key Person:
Dr. Marie-Claire Forgue joined W3C at INRIA Sophia-Antipolis in January 2001, as Head of W3C European Communications. Marie-Claire received a Ph.D. degree in Computer Science from the University of Nice and INRIA, France. After a year as a postdoctoral fellow at the Dynamic Graphics Project Lab at the University of Toronto, Canada, she worked in NTT's Human Interface Lab, Japan, for two years. Her research interests were focused on illumination algorithms and scene modelling. After that, she studied filmmaking in Vancouver, Canada. She has directed several short films and documentaries, and got interested in interactive multimedia back in 1993.

Yves Lafon studied Mathematics and computer science at ENSEEIHT in Toulouse, France, and at Ecole Polytechnique de Montreal in Montreal, Canada. His field of study was signal recognition and processing. He discovered Internet Relay Chat and the Web in Montreal in 1993 and has been making robots and games for both. He joined the W3C in October 1995 to work on W3C's experimental browser, Arena. Then he worked on Jigsaw, W3C's Java-based server, on HTTP/1.1 and started the work on SOAP 1.2. Yves is now the Web Services Activity leader, team contact of the W3C Media Fragment Working Group, editor of the HTTPbis specification.

2.2.6 TEMIS SA (Temis), France

Expertise:
TEMIS is a leading provider of Text Analytics and Text Mining solutions, using concepts and meaning extraction, automatic classification and relationships representation to address the unstructured data management needs of corporations and governments in Europe and the United States. Worldwide, 1000 companies and governments have chosen to implement TEMIS solutions in various environments where information processing is critical such as Competitive Intelligence, Customer Relationship Management, Scientific Intelligence, IP management or Human Resources. TEMIS technology provides superior results, using its award-winning and patent protected linguistic technology as well its packaged Skill CartridgesTM for domain-specific analysis. TEMIS linguistic technology is available today in 20 languages, including Chinese, Japanese, Korean and Arabic.

TEMIS, recently introduced its first industry-specific edition of Luxid®, serving the global information needs of Governments and Corporations, which brings answers to the challenge of information discovery and knowledge extraction from unstructured data. Luxid® is a break-through solution that supports demanding activities such as Competitive Intelligence, Scientific Intelligence, Customer Sentiment Analysis, Reputation Management, and Publishing. In 2007, TEMIS won the European Information and Communications Technology Prize with Luxid®, TEMIS Information Intelligence solution.

Role in the project:
In the NEED project TEMIS will contribute on named entities extraction, events detection and automatic categorization of documents with WP3. Temis will also develop Web services to access dynamically other web pages or even external knowledge bases. Being part of the market, Temis will naturally lead the general exploitation of the project and be responsible of **WP7: Exploitation and Dissemination**.

Key Person:
Charles Huot, co-founder and Chief Operation Officer, leads the group's expansion worldwide and provides guidance in achieving a balance between applying business and technology expertise to clients' needs. He is responsible for TEMIS' strategic development, which focuses on the development of long-term relationships with corporate customers. Before co-founding TEMIS, Charles spent 10 years with IBM, where he was instrumental in developing international sales for their pioneer Text Mining software. While at the University of Marseille, his academic studies focused on Competitive Intelligence strategies. Charles is a leading specialist in this field, holding regular seminars on competitive intelligence across Europe and the United States

Christophe Aubry, Vice-President Professional Services & Co-founder of TEMIS S.A., is responsible for project management and delivery of applications toward TEMIS customers. Prior to joining TEMIS, Christophe Aubry was the project coordinator for the IBM solution "Technology Watch" where he was responsible for bringing this new product from the development stage to commercialization. With strong experience in using applied mathematics in technology, Christophe was able to develop this innovative Text Mining solution based on text analysis and computational statistics. Christophe holds a Master's degree in mathematical science from the University of Orléans, France.

Sylvie Guillemin-Lanne, innovative projects manager, is responsible for project management of Innovative EU and French Research Projects, including Homeland security. Sylvie started working at TEMIS in 2000 as a Senior Linguistic Consultant. She developed the first version of the Competitive Intelligence Skill Cartridge™. She was project manager for several customers' projects before taking the management of innovative projects team. Sylvie has many years of experience as a computational linguist, previously for IBM where she took an active part in the specifications and development of a grammar checker and developed the French transfer components of an English/French translation machine. Sylvie holds a DEA en Linguistique Informatique (Paris VII, 1991).

Dr. Amanda Bouffier joined TEMIS as a Project Manager towards several customers. Beside private customers in domains like bank and publishing, she's especially in charge of innovative projects, including European and French Research projects. Amanda Bouffier has obtained with honours a PhD in computer science (Natural Language processing) in 2008 at the University of Paris 13.

2.2.7 CINECA Consorzio Interuniversitario (CINECA), Italy

Expertise:
CINECA[81] is the high technology bridge between the academic world, research and the world of industry and public administration. CINECA is the largest Italian High Performance computing center and one of the largest at European level. It offers support to scientific research activities for academia and research institutions, providing HPC facilities and specialized applications. CINECA implements a leading edge computing environment and advanced data center architectures and technologies. Among its institutional missions is the development of IS and services for universities and the Ministry of University and Research (MUR). In this role, the Consortium is engaged in the constant search for solutions capable of supporting universities in their governance processes.

Over time CINECA has been appointed a specific technical role by the MUR in order to develop, deploy and maintain the infrastructure and systems to allow the interaction between all the components of the academic world and the central administration, guaranteeing the Ministry constant monitoring of the governance processes and coordination of activities. CINECA is also actively involved in technology transfer aimed at SMEs and local government bodies. CINECA has been very active in the EU projects area in the FP4, FP5, FP6 and FP7 programs. It is presently participating in several projects in the research infrastructure and e-infrastructure domain (PRACE, DEISA-2, HPC-Europa 2) as well as project in the ICT area: software and services, digital libraries and technology enhanced learning (SmartLM, Beingrid, Papyrus, PlugIT, Arrow).

Role in the project:
Within the NEED project, CINECA will provide the infrastructure, the architecture design and implementation and the integration of the different components (ontology, analysis tools and interfaces) developed inside the project, leading **WP5: Framework Architecture Design**. The project will benefit of CINECA experience in large scale data storage and management as well as in data mining and text mining applications. In this context, CINECA developed NLP techniques and text mining applications specifically for the ontology learning task and the automatic classification of news in the Italian Research National Portal. Inside the ASTREA National funded research project, CINECA developed Information Extraction techniques for the monitoring of the Italian judicial system through the analysis of the texts of the Court decisions. Inside the PAPYRUS European Project, CINECA developed the audio component analysis tools for the targeted multimedia content analysis and a language processing tool to identify relevant concepts in news items, add new semantic metadata and link to the News Ontology. CINECA will benefit of its participation to the project by gaining experience on emerging trend detection, which can be exploited in a new service for the Italian Universities that will focus on trend detection in scientific publications.

Key Person:
Roberta Turra leads the Knowledge Discovery and Management team within the Information and Knowledge Management Services Dept. She graduated in statistics and joined CINECA in 1994, where she has been developing data and text mining applications and services. Her research interests include Predictive Modelling, Natural Language Processing, Semantic metadata generation and Multimedia Mining.

Dr. Giorgio Pedrazzi is a technologist at CINECA. He has received his PhD in Statistical Methodology for Scientific Research from the University of Bologna. His research interests include Data Mining, Text Mining and Information Extraction. Presently is involved in the European project Papyrus (FP7-ICT-2007-1) and in the development of Data and Text Mining applications for CINECA customers.

Sergio Bernardi has been involved for many years in the area of IT infrastructure management and data center operations as Head of Department. He's presently responsible for the Business Development area of the System and Technology Department (DSET) of the Consortium CINECA.

Rosita Bacchelli leads the Technology and Application Development and Standards team within the Information and Knowledge Management Services Dept. She studied computer science at University of Modena and joined CINECA in 1995, where she has been developing large web portals and web-based applications.

[81] http://www.cineca.it

2.3 Consortium as a whole

2.3.1 Consortium Setup

The **NEED** Consortium is an interdisciplinary team of engineers and researchers in Computing Science, Social Science and Web Science, journalists and professionals in the media industry. It comprises four leading research institutes (IBBT, EURECOM, CWI and CINECA), a worldwide leader in the production of news (AFP), a key technology provider (Temis) and a standardisation body in direct liaison with the market (ERCIM/W3C). This mixture of research and industry ensures a secure provisioning of news content, a concentration of the required know-how with regard to research and development and the best-possible options for uptake of the results throughout Europe and in the World.

Realizing the ambitious vision of **NEED** requires the participation of a major news agency provider (AFP) and the skills and infrastructure for analyzing large scale social media content (CINECA) in order to cover the entire news production workflow. Automatic processing of the multimedia news content requires an expert in image and video analysis (EURECOM/IBBT) and an international expert in textual news content analysis (Temis). Best practices in ontology engineering need to be mastered for integrating this large amount of data (IBBT/EURECOM). Social media use needs to be analysed from a sociologist point of view (IBBT) and powerful but usable interfaces needs to be designed by someone that understands user needs while being an expert in using semantic web technologies to describe multimedia news content (CWI).

NEED aims to have a clear impact in various standardisation bodies. The consortium partners (respectively CWI, AFP and IBBT) are already co-chairing specific technical working groups within respectively W3C, IPTC, EBU and ISO and will thus be able to disseminate largely the project results. Furthermore, ERCIM/W3C is directly involved in the consortium. Each WP leaders has been chosen given its knowledge and expertise in the field while at least **three** partners will be involved in each task, emphasizing the cooperation and integration among the partners.

The following table summarises the skills contributed by each organisation along with the primary role every partner will undertake in the project:

Partner No.	Short Name	Country	Partner skills	Role in the project
1	IBBT	Belgium	Research institute expert in **knowledge-based image and video analysis**, interactive television and mobile applications, and in media communications	Project **administrative coordinator**, expert in understanding **user needs**, leading the **knowledge infrastructure** for news integration.
2	EURECOM	France	Research institute and school, expert in **multimedia analysis** and **multimedia semantics**	Project **scientific coordinator**, providing research and implementation of multimedia analysis toolkits, leading the **event detection and news enrichment** work package.
3	CWI	The Netherlands	National research institute with long-record in **multimedia semantics** and housing the W3C Benelux office.	Expert in deploying semantic web-based middleware and interfaces for multimedia, leading the research and development of **semantic multimedia news interfaces**.
4	AFP	France	**One of the three leading worldwide news agencies** with content in six languages.	**User** and multimedia **news content provider**, leading the **evaluation** of NEED technologies.

Partner No.	Short Name	Country	Partner skills	Role in the project
5	ERCIM/W3C	France	Standardisation body for the **Web standards**, willing to make video a first class citizen on the Web.	Leading the **standardisation activities**.
6	Temis	France	**European leader** in **language processing**, dealing with more than 20 languages.	Technology provider for **knowledge extraction** from text, event detection and annotation, leading the **project exploitation and dissemination**.
7	CINECA	Italy	**Largest** Italian **High Performance computing center** and one of the largest at European level.	Expert in infrastructure for deploying **large scale** processing and management of semantic metadata, provide **data mining** technology for trend detection.

Figure 18: Consortium as a whole

2.3.2 Sub-contracting

Some partners foresee an amount of 3.000 Euros for subcontracting in the management costs. This is simply for the required audit certificates and obviously fully inline with common practice. Apart from that, no sub-contracting is foreseen in NEED.

2.3.3 Involvement of Other Countries

There is no partner involved in the NEED proposal that is based outside of the EU Member states or associated countries.

2.4 Resources to be committed

Mobilisation of Resources:
The **NEED** consortium will mobilize the critical mass of resources (personnel, equipment, data, users and finance) necessary for the successful completion of all the objectives of the project. IBBT will constantly, as the project co-ordinator, monitor and report the utilisation of the project's resources, providing the required support and contingency actions in cases of variations, following the procedures described in Section 2.1. Own funded resources from all partners have been secured to ensure the realisation of the project goals as planned. Academic and Research partners (IBBT, EURECOM, CWI, CINECA) will contribute complementary effort from permanent staff employed, while commercial partners (AFP, Temis) and non profit organization (ERCIM/W3C) have committed to complement personnel and activities from their own expenses as **NEED** is in line with their internal R&D and commercial objectives. Furthermore, AFP has committed to deliver to the project consortium both a very large corpus of multimedia news data and users for testing and evaluating the technologies developed during the project. Finally, the participation of user groups has been secured with special provision in the project budget.

Personnel:
As it has been described in Sections 2.2 and 2.3, the **NEED** consortium has all the necessary expertise required in the project. The Consortium consists of four leading academic institutions and research centres (IBBT, EURECOM, CWI and CINECA), a world renowned news provider (AFP), an experienced technology provider (Temis) and a standardization body (ERCIM/W3C). Highly qualified and experienced personnel from the participating organisations will be involved in the project and contribute to its successful completion. Indicative descriptions of the key people who will be involved in the project have been included with the partner profiles in Section 2.2. Finally, three key representatives of standardisation bodies (W3C, IPTC and EBU) and several user groups have expressed their commitment to support the project and participate in the **NEED** evaluation and dissemination activities.

Dissemination:
A significant part of the budget is dedicated to dissemination activities. This includes the cost of 16 PM and a communication budget of around 20.000€ for the project. The communication material includes the cost for 3 press releases, of press clippings, of the design and branding of NEED, of the design and realization of posters, brochures, and multimedia communication materials and the production of some goodies. This activity being refunded at 100% per EU rule, the overall budget becomes important in volume while being still reasonable with respect to the overall project budget given the high ambition of the consortium to promote the technologies developed and make real breakthrough in the market at mid-term.

Other Costs:
The Consortium partners have to a large extent the required equipment to perform the intended tasks. However, an amount of 5.300€ has been calculated to cover expense related to acquisition of equipment and consumables to support the research activities, the system development and the conducting of the trials.
A provision of 27.000€ has been made to cover the expenses of the user groups that will gather at multiple times along the lifetime of the project. This budget will cover the expenses of the participants (travel, catering and hotel costs) on the basis of 10 to 15 participants per workshop.
The travel expenses foreseen for participation in project and reviews meetings, dissemination and standardisation activities are 137.700€ given the worldwide scope of the standardisation bodies. The detailed explanations are as followed:

- For each partner, 3 project meetings per year per person that is 7.700 € per person for the project;
- For the research partner, 2 international conferences per year that is 13.200 € for the project;
- For the standardisation partner, 2 international meetings per year that is 4.400 € for the project;
- Specific barcamps targeted by the project (e.g. ParisWeb developer events), 3.600€ for the project.

Financial Resources:
The precise financial information for the project is given in the A3 Forms. We provide below a number of graphical representations to better illustrate the main financial aspects of the project which are representative of the resources required for the realization of the **NEED** objectives and vision.

The graphic below shows the total budget for each activity type: RTD activities (including the travel costs), Demonstration activities (including the equipment costs), Dissemination activities and Management activities (including audit costs).

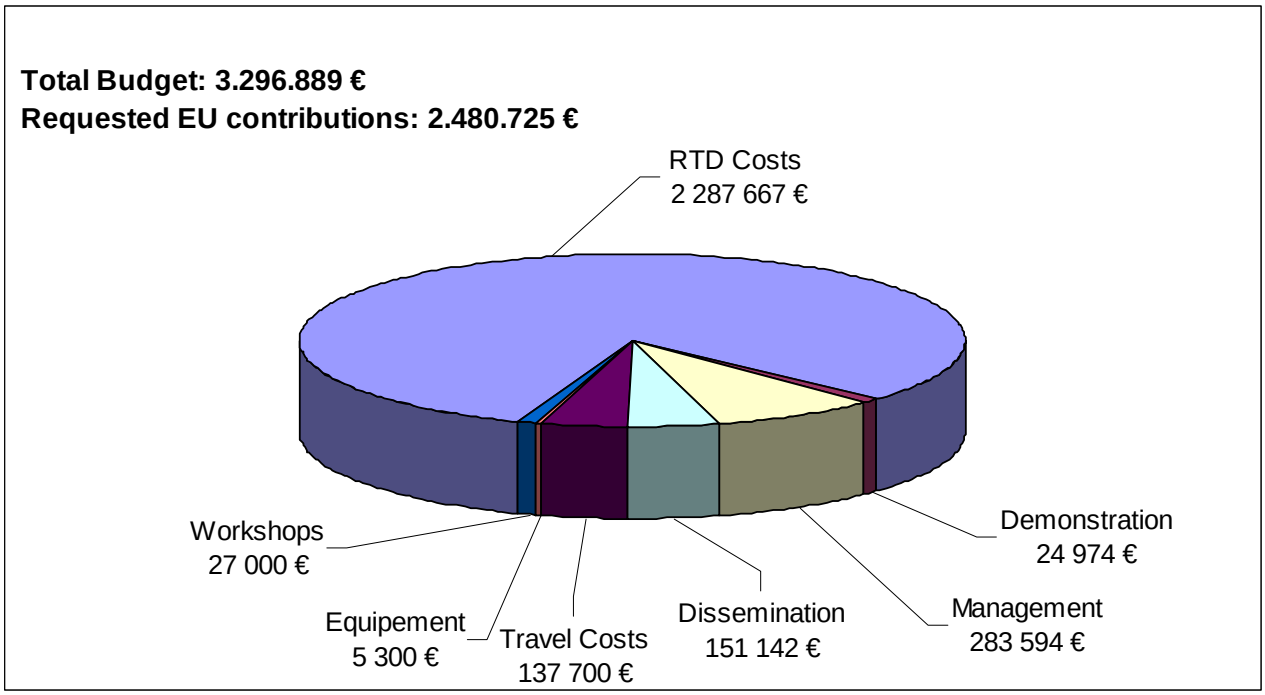


Figure 19: Total budget per activity type

The graphic below represents the distribution of the total budget and the requested EU funding for each partner.

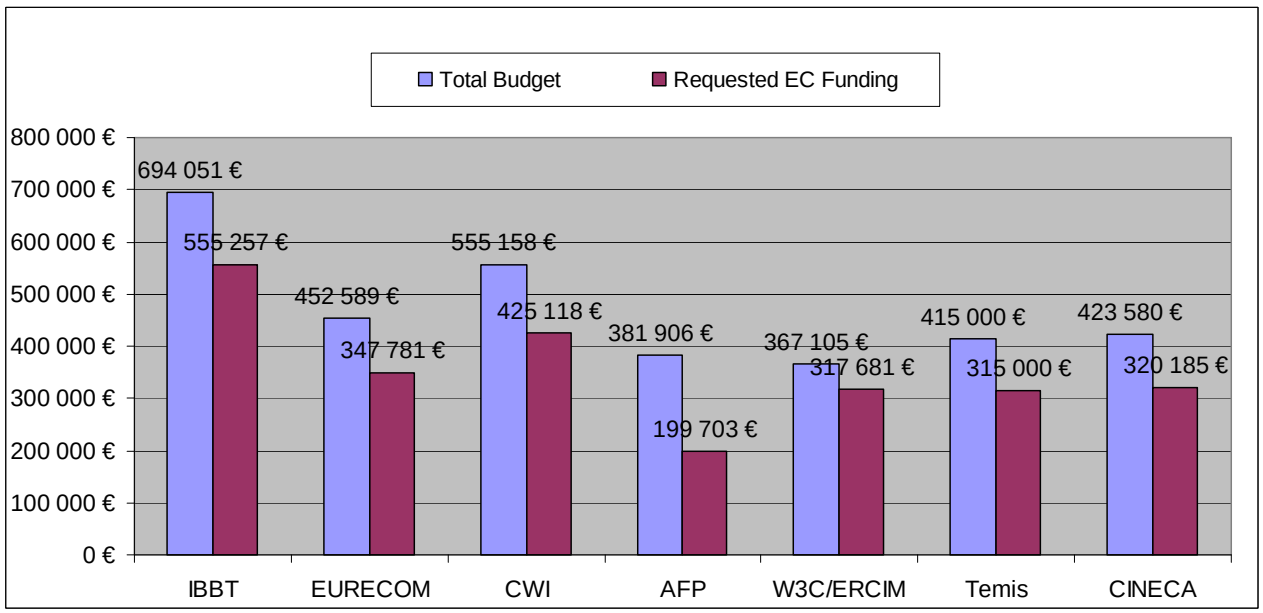


Figure 20: Total budget and requested EU contribution per partner

Section 3: Impact

3.1 Expected impacts listed in the work programme

Nearly every European citizen reads, watches or listens to the news. As voting citizens, we need to understand local, national and international politics to allow us to cast our vote. As company employees, we need to understand the state and development of local, national and international economies to enable us to understand our markets. As part of our leisure time, we want to know about our favourite sports teams, the lives of our soap idols or the most recent books available. In summary, today's globalised world demands well-informed citizens. Furthermore, **NEED** aims to process multilingual news content. It is thus a perfect project specifically tailored for a European Dimension.

The project is a consortium of world leading research groups, top news organisations with multinational coverage, and a company with multinational clients that provides the best technologies in Europe.

3.1.1 Relevance to the Objectives of ICT-2009.4.3: *Intelligent Information Management*

NEED will be particularly relevant to the objectives a), b) and d) of ICT-2009.4.3.

NEED will focus on worldwide news events, which is an extremely large, rapidly evolving and potentially conflicting and incomplete amount of information. Apart from the news coming from the partner AFP, **NEED** will provide mashups with existing news feeds on the internet, such as Twitter, YouTube, Social Networks such as Facebook or LinkedIn and encyclopaedia such as Wikipedia in order to contextualize the presentation of news events. This will result ultimately in a better experience and satisfaction of the end user, by delivering more trustable, more pertinent and usable information, evaluated and validated by lead users and hot news heavy consumers.

NEED activities will make digital resources that embody creativity and semantics easier and more cost-effective to produce, organize, search, personalise, distribute and (re)use, across the value chain.

Expected ICT-2009.4.3 impacts	**How NEED addresses these expected impacts**
Better leveraging of human skills, improved quality and quantity of output and reduced time and cost allowing users to concentrate on more creative and innovative activities.	**NEED** will provide technologies for searching any type of news from multiple sources linked with appropriate knowledge thus easing the creation and repurposing of news content. **NEED** will provide multiple semantic toolbox such as automatic alignment quotes in videos, provenance verification, etc. optimizing journalists' search and annotation tasks and thus their productivity, so that they can focus on more creative tasks.
Increased ability to identify and respond appropriately to evolving conditions (e.g. in finance, epidemiology, environmental crises …) faster and more effectively. Reinforced ability to collaboratively evolve large-scale, multi-dimensional models from the integration of independently developed datasets.	**NEED** will provide a real-time information portal on environmental issues that will survive the project providing thus the means to alert EU citizens and more largely the entire world on the upcoming scientific and political development of what is often considered as the next biggest human challenge.
Higher levels of information portability and reuse by creating an ecology of systems and services that are dynamic, interoperable, trustworthy and	**NEED** will redistribute all the metadata created as linked semantic web datasets that could be further re-used by any third party applications. **NEED** will thus contribute to the open linked dataset community.

accountable by design.	
Increased EU competitiveness in the global knowledge economy by fostering standards-based integration and exploitation of information resources and services across domains and organisational boundaries.	**NEED** will be pro-active in standardization thanks to the presence of the ERCIM/W3C partner and the various consortium participants. **NEED** will propose a competitive alternative for extracting structured knowledge from textual content.
Strengthened EU leadership at every step of the computer-aided information and knowledge management lifecycle, creating the conditions for the rapid deployment of innovative products and applications based on high quality content	**NEED** will carefully associate the users for all the technologies developed, by gathering requirements at preliminary stage and by constantly evaluating the models and tools developed with user studies.

Specific expected impacts for Objective ICT-2009.4.3: Intelligent Information Management

NEED is consistent to the Big Challenges of the Strategic Research Agenda of the **NEM European Technology Platform**[82]: It directly addresses Challenge 2 "to empower end-users by putting the user first", providing the end-user with personalized access to several sources of information and selected background knowledge.

Through **NEED**, people will become better informed citizenship and will develop enhanced critical thinking.

3.1.2 Scientific Impact

Apart from the undisputable advantages the **NEED** technologies will offer to the news industry, there are also several innovations in the domain of web and computer science that will be produced as an outcome of the project. The development of **NEED** will require significant innovation, models and algorithms, to be used as a basis for further research as well as for the creation of new tools.

Firstly, the modelling of various ontologies (news, event, provenance), within a general news architecture, and combined with general multimedia ontologies for the representation of news information will constitute a major accomplishment. Ontologies have proven to be a very useful tool to provide semantics both in the context of the Semantic Web and for personal information management. They are, however, complex structures, poorly reused, rarely linked to existing knowledge already formalized and available on the web, and almost never evaluated. **NEED** will provide best practices guidelines in the knowledge engineering field for designing ontological models, bearing in mind a constant evaluation of them in end user interfaces.

Secondly, the research on context and content analysis on texts and multimedia resources, as well as knowledge extraction will contribute significantly not only for the news domain but for other domains. It is expected that the technologies developed for detecting events can be similarly applied for extracting and annotating patterns in technical fields. Some multimedia analysis techniques (e.g. quote alignment in videos, face detection) will finally be tested in real world scenario in order to benefit to the whole market.

Finally, all the metadata generated will be compatible with current semantic web formats and immediately exposed on the web, thus increasing the number of existing linked semantic web datasets for the benefit of all. Therefore, we will make use but also contribute to the ever growing linked data cloud, in particular bringing new bubbles and datasets describing environmental issues.

[82] http://www.nem-initiative.org/

3.1.3 Impact on the Media Industry

NEED novel interfaces will be used by the industrial partners AFP and Temis in order to provide their clients with innovative environment for navigating through interlinked news events and for enhancing the user experience.

NEED interfaces will not only benefit from professional journalists, but also students, professors, with new ways to search and to find relevant information on the web, to validate in a more trustable way the sources of information. It will finally help the media industry to link its own production with relevant information on the web, either with semantically related information, or with producers linked information gathered on social networks.

3.1.4 Contribution to Standards

NEED will be beneficial to the whole media sector: news agencies, independent journalists and broadcasters. As already stated, the technologies will be developed in close collaboration with the International Press Telecommunication Council[83] (IPTC) gathering all news agencies in the world, and with the European Broadcaster Union[84] (EBU). The later has set up a particular programme in the context of news production automation and **NEED** will deliver the reference software for this programme, and will act as consultant for the future Electronic Program Guides development.

NEED will finally be beneficial to the web at large, by contributing within the W3C[85] consortium to make video a first class citizen in the future web.

[83] http://www.iptc.org/
[84] http://www.ebu.ch/
[85] http://www.w3.org/

3.2 Dissemination and/or exploitation of project results, and management of intellectual property

3.2.1 Dissemination of Project Results

A structured dissemination plan will be followed during the project in order to support an effective exploitation of the project results. Dissemination activities will be conducted using the following four instruments:

Project web site, newsletters and social media communication:
A number of deliverables and milestones has been planned to assure the effectiveness of the general dissemination activities:

- The project web site will contain researchers' papers, public demonstrators, software modules, ontologies, etc.
- Six-monthly electronic newsletters will report the main activities promoted and undertaken within the project and will be distributed to both the research community and the professional media industry.
- Accounts on social media sites will be created in order to communicate quickly project results and to federate user communities and early adopters of **NEED** technologies.

Leading conferences and journals:
The scientific results of the project will result in articles to be submitted to international, high quality journals, conferences and workshops. Tutorials, workshops or special sessions will also be organised by the academic partners of the project. Relevant events targeted by **NEED** include but are not limited to: World Wide Web Conference (WWW), International Semantic Web Conference (ISWC), ACM Multimedia, ACM International Conference on Computer-Human Interaction (CHI), ACM International Conference on Information Retrieval (SIGIR), International Conference on Semantic and Digital Media Technologies (SAMT), International Conference on Knowledge Engineering and Knowledge Management (EKAW), International Conference on Image Processing (ICIP), International Workshop on Semantic Web User Interaction (SWUI), International Workshop on Content based Multimedia Indexing (CBMI), etc.

Standardisation bodies:
As already largely stated, **NEED** will invest significant efforts to contribute to existing and emerging standards. Consortium members are already participants and often co-chairing specific technical working groups within W3C, IPTC, EBU or ISO. **NEED** aims thus to further develop multimedia news standards and provide and distribute reference implementations of them.

Industrial exhibitions:
AFP will intend to disseminate practical implementation of **NEED** results within international media networks such as the international newspaper industry research association[86] (IFRA) which recently merged with the World Association of Newspapers[87] (WAN) or the American and European network of news agencies[88] (Minds International).

3.2.2 Exploitation of Project Results

Exploitation at AFP:
AFP will exploit the results of the project in its daily business activity. The metadata integration will benefit to the overall value chain: broadcasters will produce TV news more easily based on the intelligent processing of news stories, while documentalists and archivists in media and in Digital libraries will get richer descriptions and contextualization of news assets and thus will be able to search and reuse easily news content material.

[86] http://www.ifra.com/
[87] http://www.wan-press.org/
[88] http://www.minds-international.com/

News content production is characterized by a dynamic and flexible environment. Players on this market such as AFP have a prominent role since they are both early adopters of technologies that automate and accelerate parts of their production processes. As opposed to traditional workflows, where a news agency would provide raw material and a news editor would take care of the packaging and distribution to the consumer, we are now in a situation where these roles have become interchangeable. Using internet technology, news agencies have direct access to the end user and through the European Broadcasting Union (EBU), traditional broadcasters have in fact become part of a news agency network. The clear result of this is that news agencies, broadcasters or news publishers in general, as well as the end user are confronted with a massive increase of available information.

The designed infrastructure of **NEED** will provide journalists with extraction of concepts and metadata, suggestion of related or similar news, selection of multimedia objects (e.g. photos, videos or sounds) related to the news they are currently working on, in order to produce richer and more relevant content. It will also provide tools to link different media objects between them in a semi-automatic way, based on the metadata. Journalists will be thus guided for producing faster and better multimedia news content by optimizing their tasks such as searching for particular facts, gathering information on a given topic, finding quotes in a video, etc.

Exploitation at Temis:
Temis plans to have a new knowledge extraction module that will be market to several business such as Fortune 1000, Press, Information aggregator. The foreseen market for this component is in the range of several millions of Euros in the coming 4 years.

Metadata creation and especially in the area of “events detection” is a key element for content provider and many large organisations dealing with external and internal information. One out of many potential derivative products that will came out of the **NEED** project is a semantic component named “Events detection” Skill Cartridges. This valuable component will be very attractive for the market. We will focus on the creation of an operational marketing plan as well as a specific packaging of this knowledge component in order to serve the market. Among several benefits the marketing plan will strengthen gains in productivity (creation and management of dedicated knowledge bases using text mining and ontology management systems), flexibility (monitor content repurposing and repackaging, create associations between contents, ontologies and thesaurus-based information retrieval) and deployment (multilingual and standard-based technologies improving interoperability in knowledge aware environment).

In addition to the development of the skill cartridge market, TEMIS intend to push a new set of offering based on the availability of web services for application developers. This model is based on the same approach developed by OpenCalais. A development kit will be available for download together with an API key including a free evaluation period of the event annotation web service. Contrary to OpenCalais, Temis has no motivation for copying locally the content being analyzed that will therefore remain the entire property of the customers. We believe that this complementary service will reach another portion of the market and serve as the basis for the raw and fair comparison of various human language technologies for extracting knowledge and structure data from textual content.

3.2.3 Management of Knowledge and Intellectual Property

The project consortium clearly recognises that management of knowledge and IPR securing are fundamental for effective cooperation in RTD activities, avoiding information bottlenecks related to confidentiality or competitiveness and enhancing the exploitation potential of project results. Management of knowledge and IPR issues will be carefully integrated within the framework of the Consortium Agreement (CA). The Consortium agreement covers technological and commercial collaboration between partners, patenting of the technology developed (where applicable), and licensing of the technology to companies outside the consortium after an initial period of confidentiality. Background knowledge and existing intellectual property of the partners will therefore be identified at the start of the project, and its protection will be ensured according to the mutually agreed procedures. Foreground rights will be defined and regulated in the CA, whose preparation and planning will be discussed, and decided by all consortium partners.

The project consortium plans also to develop and promote several technologies within dedicated working groups set up by standardisation bodies such as W3C or IPTC. These have generally open patent policy and disclosure agreement from their members securing a prompt discovery of any IPR claims and a widespread use of the technologies developed.

Section 4: Ethical Issues

The **NEED** partners are aware of ethical and societal issues concerning privacy of users, trustworthy relationships in the whole value chain of multimedia news content creation, management, processing, distribution, and consumption, as well as accessibility.

4.1 Privacy and User Studies

In order to facilitate content processing according to user preferences and context of use, usage research teams will need to keep user data with the related metadata. For testing of developed techniques in realistic scenarios, the user model will keep a history of use for individual users and automatically learn the personal preferences from this. This could raise privacy issues. Individual data stored in digital format, even if anonymous, are subject to privacy regulations. In principle, this implies that data storage is to be restricted to what is necessary, that users have to be notified of this fact, that users have a right to inspect what is stored about them, and that their data is not to be transmitted to third parties and used for purposes other than those covered by the relationship the user has entered to the owner of the data.

Users who take part in user studies will be made aware that they are part of an experimental setting. They will be informed of this before the study, or afterwards if the experiment would be influenced by prior knowledge of the goals. When the same users are asked to take part in future studies their experimental results will be kept independently of their personal data to ensure anonymity.

Overall, **NEED** will observe European legal regulations concerning privacy and will particularly comply with Data Protection legislation[89] in the Member State where the research will be carried out regarding ICT research data that relates to volunteers.

European Legislation Framework on Project Related Topics	**NEED compliance**
Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)	✓
Directive 95/46/EC of the European Parliament and Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data	✓

4.2 Trustworthiness

Trustworthiness is another important issue in the news industry. **NEED** will take the necessary measures to foster trust of users in the technology developed. The project will deal with the policy aspects of trust, and it will recommend and apply existing technology to achieve and ensure security and trustworthiness. Observing a line of prudent concern, **NEED** will make sure that by means of operational guidelines and procedures all user-specific information will reside in the users' terminal equipment and will only be exchanged to configure, perhaps dynamically, the communication link to be established. From a research point of view, **NEED** will also deal with provenance and trust issues by performing research on multimedia data authentication and copyright protection as described in the work plan (see the Activity WP3.2).

[89] National legislation transposing Directive 95/46/EC - http://ec.europa.eu/justice_home/fsj/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf

4.3 Accessibility

NEED partners are aware of accessibility issues generated when new technologies are introduced and therefore special care must be taken to avoid the creation of barriers that put people off the new technologies. For this reason, the project aims to benefit both the global media industry and the European non specialist citizen from accessing multimedia news information using the technology that is being developed. The reason for this is to establish at an early stage what is interesting to the user, what is helpful to the user, e.g. with different literacy skills and languages, and conversely what may be technologically interesting but either confusing or not addressing features that the user, i.e. viewer, listener, or traveller needs. This will provide feedback on the targeted user centred research and on the tools and techniques that are given attention. If this can be achieved, it will contribute to the success of the research and have wide societal implications.

We do not see, however, accessibility issues as addressing the needs of a small part of the population, but rather view them in the wider sense of providing appropriate interfaces for users in "interface challenged" situations, such as getting informed of the latest news while driving a car ("visually impaired") or while being frequently interrupted by children ("cognitively impaired"). Including knowledge about the expressivity of different modalities in which information is and can be expressed will contribute to enabling creating output presentations appropriate for different users. Many proposals have been made for minimum standards for access, and many organisations encourage content creators to follow these accessibility standards. One of the opportunities of **NEED** is to express the underlying semantics of these explicitly in terms of standard Semantic Web languages.

We will finally conform to W3C Accessibility guidelines[90] while publicising **NEED** research and development activities on the Web. The presence of ERCIM/W3C in the consortium will guarantee that accessibility issues are always taken into account when communicating project results.

4.4 Gender Issues

NEED partners are aware of the gender issues in technical projects and will do their utmost to ensure adequate female participation at all levels of the project, given the scarcity of the resource and while maintaining quality norms. We will also ensure that the interfaces created in the project are tested by equal numbers of male and female users and will consider the potentially different communication needs of these users in creating the applications.

[90] http://www.w3.org/WAI/

ETHICAL ISSUES TABLE

	YES	PAGE
Informed Consent		
• Does the proposal involve children?		
• Does the proposal involve patients or persons not able to give consent?		
• Does the proposal involve adult healthy volunteers?		
• Does the proposal involve Human Genetic Material?		
• Does the proposal involve Human biological samples?		
• Does the proposal involve Human data collection?		
Research on Human embryo/foetus		
• Does the proposal involve Human Embryos?		
• Does the proposal involve Human Foetal Tissue / Cells?		
• Does the proposal involve Human Embryonic Stem Cells?		
Privacy		
• Does the proposal involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)		
• Does the proposal involve tracking the location or observation of people?		
Research on Animals		
• Does the proposal involve research on animals?		
• Are those animals transgenic small laboratory animals?		
• Are those animals transgenic farm animals?		
• Are those animals cloned farm animals?		
• Are those animals non-human primates?		
Research Involving Developing Countries		
• Use of local resources (genetic, animal, plant etc)		
• Impact on local community		
Dual Use		
• Research having direct military application		
• Research having the potential for terrorist abuse		
ICT Implants		
• Does the proposal involve clinical trials of ICT implants?		
I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	**YES**	

Section 5: References

[Arndt *et al.*, 2007]
R. Arndt, R. Troncy, S. Staab, L. Hardman and M. Vacura. *COMM: Designing a Well-Founded Multimedia Ontology for the Web*. In 6th International Semantic Web Conference (ISWC'2007), vol. LNCS 4825, pages 30-43, Busan, Korea, 2007.

[Aroyo *et al.*, 2007]
Lora Aroyo, Natalia Stash, Yiwen Wang, Peter Gorgels, Lloyd Rutledge: CHIP Demonstrator: Semantics-Driven Recommendations and Museum Tour Generation. ISWC/ASWC, pages 879-886, 2007.

[Faloutsos *et al.*, 1994]
C. Faloutsos, W. Equitz, M. Flickner, W. Niblack, D. Perkovic, and R. Barber. Efficient and effective querying by image content. Journal of Intelligent Information Systems, 3:231–262, 1994

[Fernández *et al.*, 2007]
N. Fernández, L. Sánchez, J. M. Blázquez and J. Villamor. *The NEWS Ontology for Professional Journalism Applications*. In Ontologies - A Handbook of Principles, Concepts and Applications in Information Systems, Integrated Series in Information Systems, Vol. 14, Springer editor, 2007.

[Gamon et al., 2008]
Michael Gamon, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, Arnd Christian König: BLEWS - Using Blogs to Provide Context for News Articles. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2008.

[Garcia and Celma, 2005]
R. Garcia and O. Celma. Semantic Integration and Retrieval of Multimedia Metadata. In 5th International Workshop on Knowledge Markup and Semantic Annotation, pages 69–80, 2005.

[Haralick, 1996]
Haralick, RM, Statistical and structural approaches for textures, Proc IEEE, 67, pages. 786-804, 1979.

[Hildebrand *et al.*, 2006]
Michiel Hildebrand, Jacco van Ossenbruggen, and Lynda Hardman. /facet: A Browser for Heterogeneous Semantic Web Repositories. In: The Semantic Web - ISWC 2006, pages 272-285, 2006.

[Hunter, 2001]
J. Hunter. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. In 1st International Semantic Web Working Symposium (ISWC), pages 261–281, 2001.

[Hyvönen *et al.*, 2005]
E. Hyvönen, M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen. MuseumFinland — Finnish museums on the semantic web. Journal of Web Semantics 3(2-3) pages 224–241, 2005.

[Kim *et al.*, 2005]
S.-M. Kim, J. Byun, C. Won. A scene change detection in H.264/AVC compression domain, Lecture Notes in Computer Science vol. 3768, pages 1072–1082, 2005.

[Li et al., 2005]
Zhiwei Li, Bin Wang, Mingjing Li, Wei-Ying Ma. A Probabilistic Model for Retrospective News Event.. In the 28th Annual International ACM SIGIR Conference (SIGIR'2005), 2005.

[Liu *et al.*, 2004]
Y. Liu, W. Wang, W. Gao, W. Zeng. A novel compressed domain shot segmentation algorithm on H.264/AVC. In Proceedings of the IEEE International Conference on Image Processing, Vol. 4, pages 2235–2238, 2004.

[Lowe, 1999]
Lowe, D. G. "Object recognition from local scale-invariant features". In Proceedings of International Conference on Computer Vision, pages 1150-1157, 1999.

[Martens *et al.*, 2007]
Martens G., Poppe C., Van de Walle R.. Enhanced Grating Cell Features for Unsupervised Texture Segmentation. In Performance Evaluation for Computer Vision: 31ste AAPR/OAGM Workshop 2007, pp. 9-16, 2007

[Mehrotra *et al.*, 1997]
S. Mehrotra and K. Chakrabarti and M. Ortega and Y. Rui and T. Huang, Multimedia analysis and retrieval system, Multimedia analysis and retrieval system. In Proceedings of the 3rd International Workshop on Information Retrieval Systems, 1997.

[Nack *et al.*, 2005]
F. Nack, J. van Ossenbruggen and L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part II). IEEE Multimedia, 12(1), 2005.

[van Ossenbruggen *et al.*, 2004]
J. van Ossenbruggen, F. Nack and L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). IEEE Multimedia, 11(4), 2004.

[Ringel et al., 2003]
M. Ringel, E. Cutrell, S. Dumais, E. Horvitz. Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores, INTERACT 2003, September 2003.

[Rui *et al.*, 1998]
Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: A power tool in interactive content based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Segmentation, Description, and Retrieval of Video Content, 8(5):644–655, September 1998

[Smith and Chang, 1996]
John R. Smith and Shih-Fu Chang. VisualSEEk: A Fully Automated Content-Based Image Query System. ACM Multimedia, pp. 87-98, 1996

[Troncy and Carrive, 2004]
R. Troncy and J. Carrive. A Reduced Yet Extensible Audio-Visual Description Language: How to Escape From the MPEG-7 Bottleneck. In 4th ACM Symposium on Document Engineering (DocEng'04), Milwaukee, Wisconsin, USA, 2004.

[Troncy *et al.*, 2006]
R. Troncy, W. Bailer, M. Hausenblas, P. Hofmair, and R. Schlatte. Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7 Profiles. In 1st International Conference on Semantics And digital Media Technology (SAMT'06), pages 41–55, Athens, Greece, 2006.

[Troncy *et al.*, 2007]
R. Troncy, O. Celma, S. Little, R. Garcia and C. Tsinaraki. MPEG-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In 1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies (MAReSO'07), Genova, Italy, 2007.

[Tsinaraki *et al.*, 2004]
C. Tsinaraki, P. Polydoros and S. Christodoulakis. Interoperability support for Ontology-based Video Retrieval Applications. In 3rd International Conference on Image and Video Retrieval (CIVR), pages 582–591, 2004.

[Viola and Jones, 2002]
Paul Viola and Michael Jones. Robust Real-time Object Detection. International Journal of Computer Vision, 2002

[Wolf, 1996]
W. Wolf. Key frame selection by motion analysis. In Proceedings of ICASSP 96, vol. II, pages 1228–1231, 1996.

[Zang *et al.*, 1993]
H.-J. Zhang, A. Kankanhalli, S. Smoliar, Automatic partitioning of full-motion video, Multimedia Systems 1 (1) (1993) 10–28.

[Zap *et al.*, 2005]
L. Zapf, N. Fernández and L. Sánchez. *The NEWS Project - Semantic Web Technologies for the news domain*. In 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT'05), pages 455-460, London, UK, 2005.

[Zeng and Gao, 2005]
W. Zeng, W. Gao. Shot change detection on H.264/AVC compressed video. In Proceedings of the IEEE International Symposium on Circuits and Systems, Vol. 4, pages 3459–3462, 2005.

Annex A: Letters of Endorsement

EUROPEES PARLEMENT

LID VAN HET EUROPEES PARLEMENT

Brussels, Thursday the 22nd of October 2009

To whom it may concern,

As a Member of European Parliament I wish to endorse the NEED-project as its objectives and research work proposes a highly innovative approach to exploitation of all types of aggregated, real-time multimedia content put into different perspectives (e.g. provenance, historical context, ...) with a vision of revolutionizing the user experience with a great impact on our own dissemination and awareness activities.

Yours sincerely,

Bart Staes
Greens/EFA

Tinne VAN DER STRAETEN
FEDERAAL PARLEMENTSLID
KAMER VAN VOLKSVERTEGENWOORDIGERS

SECRETARIS

20 October, 2009

To whom it may concern,

Groen! wishes to endorse the NEED-project as its objectives and research work proposes a highly innovative approach to exploitation of all types of aggregated, real-time multimedia content put into different perspectives (e.g. provenance, historical context, ...) with a vision of revolutionizing the user experience with a great impact on our own dissemination and awareness activities and we wish to extend our collaboration and support during the project life-time to make it succeed by being part of their user group.

Yours sincerely,

T Van der Straeten

Tinne Van der Straeten

Federaal Parlementslid Groen!

Leuvenseweg 21 - 1000 Brussel - Tel.: +32 2 549 88 15 - Fax: +32 2 549 87 98 - e-mail: tinne.vanderstraeten@groen.be

Bond Beter Leefmilieu
KOEPEL VAN MILIEUVERENIGINGEN Vlaanderen vzw

Brussels, Wednesday the 21st of October 2009

Our ref.: AS/DJ/09258

To whom it may concern,

Bond Beter Leefmilieu, the federation of the Flemish environmental organizations, wishes to endorse the NEED-project as its objectives and research work proposes a highly innovative approach to exploitation of all types of aggregated, real-time multimedia content put into different perspectives (e.g. provenance, historical context, ...) with a vision of revolutionizing the user experience with a great impact on our own dissemination and awareness activities and we wish to extend our collaboration and support during the project life-time to make it succeed by being part of their user group.

Yours sincerely,

Danny Jacobs

Danny Jacobs / Director
Bond Beter Leefmilieu Vlaanderen vzw (BBL)
Address: Tweekerkenstraat 47 | 1000 Brussel
Office: +32(0)2 282.17.26 | Fax +32 (0)2 230.53.89 | danny.jacobs@bblv.be
Mobile: +32 (0)475 619.666
Website: www.bblv.be