News Research Proposal

Raphaël Troncy, CWI Amsterdam

January 13, 2008

1 Framework Package 7 - ICT

Call identifier: FP7-ICT-2007-3

- Funding scheme: Collaborative projects Small or medium-scale focused research projects (STREP)
- Challenge 4: Digital libraries and content
- Objectives: ICT-2007.4.4 Intelligent content and semantics
- Closure date: 8 April 2008 at 17:00, Brussels local time

Proposal details:

- Name of the proposal: **Making Sense of the News**
- Duration: 3 years

Contacts:

- Raphaël Troncy `<Raphael.Troncy@cwi.nl>` (CWI),
- Lynda Hardman `<Lynda.Hardman@cwi.nl>` (CWI),
- Laurent Le Meur `<Laurent.LeMeur@afp.com>` (AFP),
- Erik Mannens `<erik.mannens@ugent.be>` (IBBT),
- Rik van de Walle `<rik.vandewalle@ugent.be>` (IBBT),
- Marteen Verwaest `<Maarten.Verwaest@vrt.be>` (VRT),
- Charles Huot `<charles.huot@temis.com>` (Témis[1])

Participant no.	Participant organisation name	Part. short name	Country
1	Centrum voor Wiskunde en Informatica	CWI	Netherlands
2	Agence France Presse	AFP	France
3	IBBT/UGhent - Multimedia Lab	IBBT	Belgium
4	Vlaamse Radio en Televisie	VRT	Belgium
5	Témis	Témis	France

[1] `http://www.temis.com/`

2 Research Proposal

2.1 Vision

The web infrastructure provides a natural publishing platform for multimedia news content for multiple devices (e.g. PC, mobile phone, PDA). Traditional news providers (e.g. journalists, news agencies, press, broadcasters) make use of the Web for distributing news. More recently, non-traditional news actors, often called *citizen media* or *independent media*, make use of Web technologies to publish alternative views and opinions of events. Unsurprisingly, Web users use more and more the Web to get informed but are often overwhelmed by too many information without having all the time the keys to understand clearly the events.

We will build the technological infrastructure to enrich the user experience, giving the key to understand and have a critical view when searching for particular news or when browsing for being kept up-to-date. The technological infrastructure will allow the web user to aggregate multiple heterogeneous and distributed information sources (e.g. traditional media, blogs, podcasts, TV archives, photos and videos) available in multiple languages. News items come with a minimal set of metadata and can be further annotated with controlled vocabularies. They will be automatically enriched with more contextual information and linked with knowledge available on the Web. We will design and test new interfaces that use the semantic richness of the metadata for searching and browsing news content.

2.2 Use Cases

2.2.1 Michael: an individual news (Web) consumer

Michael has relatives living in the Republic of Kenya. Reading briefly the latest news, he starts getting worried this end of year 2007 and wants to better understand why Nairobi, the capital of the country, is under siege. The first articles he reads from various online newspapers contain bare facts about the police activity, and the number of people shot in the last two days, but Michael has trouble understanding the situation. After various searches, he understands that the latest controversial election is the root of the problem, since the opposition claims an important electoral fraud while the international community talks about evidence of a non transparent election. Michael wants to have visual information about the situation: maps depicting the region of conflict, photos and videos captured by reporters, graphics summarizing the parties of the conflict, usually prepared by news agencies. Finally, Michael is interested in reading both professional reports made by his favorites and trust media, and independent opinions reported by individuals living in the region.

2.2.2 Agnes: an independent journalist

Agnes is preparing an article summarizing the various IPR problems Google has encountered for aggregating news articles automatically from a number of online newspapers. She enters full text search queries in several news agencies databases she has access to, but she is completely overwhelmed by the number of results. She remembers that the AP news agency has signed a contract with Google to not be sued but has no clue how she could find the source of this information. Her search results are ranked chronologically, the latest news appearing first, but Agnes knows that the information she is looking for is at least one year old! Agnes is lucky and finds a press release from September 2006 about the verdict of a first hearing court in Belgium: Google loosing against Belgian newspapers and forced to no longer aggregate their online articles. This press release mentions the famous agreement with AP and dated it in August 2006. Agnes needs to do another search on the various databases, but she can now filters the query on the month *August 2006* – which still brings hundreds of results – and finally finds the official AP agreement press release after an hour.

2.2.3 Nicolas: CEO of a biotech company

Nicolas is the CEO of an awarded biotechnology company. His company is promised to a successful development but this activity sector is evolving fast and Nicolas needs to quickly know what products his identified and future new competitors are preparing. He would like to subscribe to an alert system that daily scan the press releases in a number of identified sources, and send him a digest that he could access everywhere at any time. Nicolas would appreciate to receive a multimedia presentation that structures the daily information, grouped by media, direct and new competitors, rumors and facts, etc.

2.3 Research Contributions

The technological infrastructure realizing this vision will be based on the NewsML model defined by the professional and semantic technologies (controlled vocabularies, domain ontologies). This research proposal will be articulated around three pillars:

- **News integration**:
  - Cross-media: News are multimedia! Textual stories come often with visual resources such as photos or videos, sometimes captured by anonymous witness participants of an event (cyber-journalism), and maps or graphics prepared by professionals. Our infrastructure will deal with all media.
  - Cross-languages: News are produced in various languages, reflecting sometimes small differences in the coverage of an event depending on a given culture. Our infrastructure will allow users to access to news content in multiple languages.
  - Cross-sources: News providers are sometimes biased (e.g. government view, lobby interests) thus motivating the user to access to multiple sources for a broader view of an event. Furthermore, both professional media and alternative media (portals and blogs) co-exist on the Web. Our infrastructure will allow the user to select and aggregate news information from several sources.
- **News enrichment**:
  - News items come with a minimal set of metadata. They can be further described by the provider or the consumer using controlled

vocabularies and domain ontologies. Our infrastructure will enhance automatically the metadata describing news in several ways: named entity recognition and natural language processing for semantic information extraction on textual information; visual processing on visual information; etc.

  - News items are not isolated pieces: often, several news items are necessary to understand fully the context of an event. Because journalists work under hard time pression, they often produce various versions of a story, sometimes correcting wrong information put in earlier versions. Our infrastructure will automatically propose links between *related* stories and maintain the versioning of each news items. Based on the metadata and the formal knowledge available on the web, and using semantic inferences, links between news items available in different languages, from different sources, on different media will be suggested, creating interconnected graphs of news items useful for the navigation.

- **News presentation**:
  - Pull mode: when searching for a particular event, or while browsing to be kept up-to-date, our infrastructure will propose new interfaces using the main dimensions of news: maps to go to geographic location (WHERE), timelines for going back in the history (WHEN), persons, organizations (WHO), event model (WHAT).
  - Push mode: user profiles, service subscription, alert and recommender systems.
  - Our infrastructure will propose a seamless integration of concept navigation and news stories navigation: user reads a story, switch to visual information, see the concepts that categorize the story, and find the (other) related stories attached to this concept.

2.4 Scrappy Notes

Questions to answer:

1. What is the problem we will solve?
2. Why is it important?
3. How will we solve it? What are we going to improve?
4. How will we know when we are done?
5. Why are we the right people to do it?

Expected features:

- News is a continuous stream, needs for a dynamic infrastructure;
- ...

3 Related Proposals

3.1 PEGASUS: eContentPlus targeted project for digital librairies

Abstract: The integration of innovative information technologies with traditional market transactional activities comprises an essential breakthrough for the evolution and prosperity of key business sectors in the current economical spectrum. The amazingly increasing quantities of digital news content originating from a plethora of sources, including the user generated content produced by 'citizen journalists', have brought irreversible changes to the business setting in the media sector. The public News Agencies no longer control the flow of news and struggle to survive the competition of freely available content. Although they still possess the most reliable and thus valuable news content, the Agencies have no longer the monopoly of the most efficient distribution channels. PEGASUS proposes a complete solution to help such news organisations face the challenges of this revolutionary era. The project combines world-leading commercial solutions with state-of-the-art technologies developed in recent R&D projects, to provide a standard-based interoperability access model which will allow for the single-point and rapid access of high quality digital content originating in distributed archives of reliable news organisations. The PEGASUS platform will virtually act as a Pan- European News Agency and, although operable on its own, will maintain the appropriate interfaces with the European Digital Library. Based on current business models, the project will target the definition of a necessary environment to allow for the sharing of digital news content through the establishment of concrete cooperation links governing the relationships in the European Media sector. The project is bringing together leading technology providers, academic partners and content owners, including major European News Agencies, originating in 5 EU Countries, in order to deliver an information service-oriented platform for managing and accessing a digital library of multimedia and multilingual news content.

EU Reviews: PEGASUS aims to provide a standards-based interoperability access model which will allow for the single-point and rapid access of high quality digital content originating from the distributed archives of news organisations. The proposal presents mainly a set of R&D activities aimed at optimising information retrieval of multi-media content. In fact, technical platform development and the integration of content-based algorithms are the main focus of the proposal while content objectives are limited in that they serve only to test algorithms. The proposal provides very little information on content-related deliverables in regards to interoperability with the European Digital Library and to the usability of the content by target user groups. Content accessibility via the European Digital Library is not detailed and sustainability via a dissemination and exploitation plan does not consider target users or partners from European countries other than those represented in the consortium. Moreover, dissemination and exploitation plans focus only on the technological results of the project and it is not clear how the sustainability of the results will be achieved.

3.2 MESH - Multimedia sEmantic Syndication for enHanced news services: FP6 IP Project

MESH is a running EU IST FP6 IP project, from March 2006 till February 2009, `http://www.mesh-ip.eu/?Page=Project` composed of 12 partners (Telefonica, ITI, ATC, Motorola, QMUL, INA, Noterik, University of Twente, Deutsche Welle, DFKI, Universidad Autonoma de Madrid, DIAS Publishing Ltd.).

MESH aims to extract, compare and combine content from multiple multimedia news sources, automatically create advanced personalised multimedia summaries, syndicate summaries and content based on the extracted semantic information, and provide end users with a "multimedia mesh" news navigation system.

Abstract: In our days we are confronted with vast amounts of information commonly referred to as "news". News about all aspects of our everyday lives are nowadays accessible to all corners of the world. But how easy is it for anyone to navigate this flood of information and how possible is it to get an objective view of controversial events, at national or international level?

- *Was the latest war an invasion or a liberation?*
- *Were the latest elections a grand victory or the result of an unfair election system?*

Our era of knowledge should provide for methods of understanding the meaning of "news". Contemporary methods should be able to organise news in a semantic way that would allow the reader to have a complete overview of all similar and conflicting views, being also able to filter information according to personal preferences and interests.

3.3 NEWS - News Engine Web Services: FP6 Project

NEWS is a completed EU IST FP6 project, from April 2004 till March 2006, `http://www.news-project.com/` composed of 5 partners (DFKI, Ontology Ltd., Agenzia ANSA SCARL, Agencia EFE S.A. and Universidad Carlos III de Madrid).

Abstract: The goal of the NEWS project is to develop News Intelligence Technology for the Semantic Web. Its main purpose is to extend the reach and delivery capabilities of online content provision and syndication services by supporting advanced personalized news discovery, analysis and presentation, and fostering interoperability across the news content provision and fruition lifecycle.

News content provision and syndication services have started to adopt basic information filtering tools, but go little beyond basic search and categorization. There is still an abyssal gap between the need of users to personalise content selection and presentation, and what the news industry can offer. In keeping with the "Semantic-based Knowledge Systems" strategic objective, NEWS addresses these shortcomings through a program of research and development aimed at:

- Using Semantic Web standards to define ontologies for the news industry;
- Implementing an ontological annotation component which automatically applies Semantic Web standards for the news industry to newswires;

- Developing news intelligence components with multilingual and multimedia capabilities, which use automatic ontological annotation to support semantic-based analysis, personalisation and delivery of knowledge from newswires, and

- Integrating ontological annotation and news intelligence components as Web services into a standard interoperable platform that enables end users and applications alike to find and utilise service components dynamically.

With the achievement of these objectives, NEWS endeavours to exert leverage on the European news industry to make available to users a host of cutting edge semantic-based Web services that are composable, interoperable and able to negotiate with each other. Such an orientation attends to the key IST objectives in FP6, by promoting European leadership in innovative news intelligence technologies and offering EU citizens wider and deeper access to news content.