Blog source validation report

<table>
<thead>
<tr>
<th>Project Reference No.</th>
<th>FP7 - 231854</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deliverable No.</td>
<td>D7.3.1</td>
</tr>
<tr>
<td>Workpackage No.</td>
<td>WP 7: User requirements, user evaluation and specifications</td>
</tr>
<tr>
<td>Nature:</td>
<td>R (Report)</td>
</tr>
<tr>
<td>Dissemination Level:</td>
<td>PU (Public)</td>
</tr>
<tr>
<td>Document version:</td>
<td>1.0</td>
</tr>
<tr>
<td>Date:</td>
<td>21/02/2011</td>
</tr>
<tr>
<td>Editors(s):</td>
<td>Liliana Bounegru, Eric Karstens (EJC)</td>
</tr>
<tr>
<td>Document description:</td>
<td>This document describes the goals and objectives of task 7.3: Mapping of pertinent blogs, specifically with original content, two blog source identification and validation exercises and their results, as well as the implications of these results on task 7.3 and core SYNC3 principles.</td>
</tr>
</tbody>
</table>
History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Reason</th>
<th>Revised by</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>06/02/2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>14/02/2011</td>
<td>Peer review</td>
<td>Google, ATC</td>
</tr>
<tr>
<td>0.3</td>
<td>15/02/2011</td>
<td>Peer review</td>
<td>NCSR’D’</td>
</tr>
<tr>
<td>0.4</td>
<td>18/02/2011</td>
<td>Prefinal version</td>
<td>EJC</td>
</tr>
<tr>
<td>1.0</td>
<td>21/02/2011</td>
<td>Final version</td>
<td>EJC, ATC</td>
</tr>
</tbody>
</table>

Authors List

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJC</td>
<td>Liliana Bounegru (author)</td>
</tr>
<tr>
<td>EJC</td>
<td>Eric Karstens (author and internal reviewer)</td>
</tr>
</tbody>
</table>
Executive Summary

This document is the final version of the first blog source validation report due M22 (January 2011). According to the Description of Work, the aim the Blog Source Validation Report is on the one hand to document the expert validation process for the relevance of blog sources indexed by the SYNC3 system, and on the other hand to map blog sources with original content that fits the SYNC3 definition of a news event, as a way to make sure that no relevant blogs are being overlooked by the SYNC3 system.

This report documents the slight shift in the purpose of task 7.3 in the context of the current stage of development of the SYNC3 system and the rapidly changing nature of the online environment, away from mapping sources with original content to expanding the list of blogs that serve as starting points in the blog crawling process. Expert validation of sources automatically indexed by the system has been shifted to later stages of development. The results of the two blog source identification and validation exercises are included in this report as well.

The validation of blog sources listed by the first SYNC3 prototype resulted in the identification of 73 relevant sources that have not been listed in the SYNC3 prototype and which have been added to the manually compiled list of sources that serves as the starting point in the blog crawling process. The difficulty to map blog sources with original content that fits the SYNC3 definition of a news event, confirmed a series of issues which showed up in other WP7 tasks as well. We recommend that future research and development efforts give a prominent place to the possibility to extend the news event definition by a topic definition and complement the organization of content around news events with an organization of content around stories or topics. As far as sentiment analysis is concerned, based on the observations included in this report and the User Evaluation Report (D7.4.1), we recommend that contextual elements such as blog or author profile, or the mere volume of circulation of a news event in the blogosphere be taken into account as indicative of sentiment, in response to the difficulties in identifying blogs which express sentiment towards news events in the content of the blog posts.
Table of Contents

Executive Summary ..........................................................................................................................3
List of Terms and Abbreviations ........................................................................................................5
1. Introduction..................................................................................................................................6
   1.1. Goals and objectives .............................................................................................................6
   1.2. Updated description and responsible contributors .................................................................6
2. Validation of proportion and relevance of blog sources in the first SYNC3 prototype ........8
3. Mapping of blog sources with original content ..........................................................................10
4. Conclusions ................................................................................................................................13
5. References ..................................................................................................................................14
6. Annexes ......................................................................................................................................15
   6.1. Annex 1: Validation of proportion and relevance of blog sources in the first SYNC3 prototype .................................................................................................................................15
   6.2. Annex 2: Occurrence of queries in a sample of blog sources .................................................15
### List of Terms and Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoW</td>
<td>Description of Work</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Goals and objectives

The Description of Work presents the purpose of T 7.3: Mapping of pertinent blogs, specifically with original content, in section B1.3.1.VII as follows: "to make sure that no relevant pertinent blogs are being excluded from the eventual SYNC3 system." [1]

The blogosphere recalls the concept of the public sphere as a sphere of online communication where public opinion emerges. Building on this conceptualization of the blogosphere, SYNC3 aims to render more accessible public debate on issues of public interest represented in news by means of structuring user-created content in the blogosphere around news events extracted from mainstream media. Task 7.3 aims to ensure that the concept behind SYNC3 is achieved by the ongoing development of a collection of representative, high quality and credible blog sources that cover all geographical regions and all news domains, with particular emphasis on sources with original content. Making visible the dynamics between the news sphere and the blogosphere, by linking together news articles with blog posts which discuss the same news event and thus syncing together the two conversations, is a way to enable fruitful and systematic interaction between journalists and informal online communication platforms, and thus to amplify events discussion in the news media sphere. In addition to this achievement, the emphasis of the initial formulation of task 7.3 in the DoW on mapping blog sources with original content was envisioned to enable journalists to identify citizen journalists who can act as sources for new stories.

1.2. Updated description and responsible contributors

Task 7.3 started in M12 (March 2010) and is planned to run until the end of the project, specifically M35 (February 2012). Three deliverables are planned in M22 (January 2011), M29 (August 2011) and M35 (February 2012). The task generally focuses on the ongoing activity of collecting and validating pertinent blogs necessary for testing and demonstrating the system throughout the entire development process. An initial list of 500 manually collected blog sources was delivered as Annex 8.2 to "D7.2.1: Content Package with Simulations" in M11 (February 2010). The list was used for crawling in order to get an initial set of news items for the first implementation period. EJC is responsible for the ongoing expansion of this initial list based on the needs of the consortium as part of task 7.3.

The Description of Work specifies two ongoing subtasks for task 7.3 [1]:

On the one hand, blog sources selected by the system will be run through an expert validation process in order to establish their degree of relevance. On the other hand – and more importantly – the specialists’ panel will actively search for and monitor such blogs which have a standing in their own right as independent news sources, particularly in European Neighbourhood countries with a public sphere that is developed to a lesser extent and where blogs not only contribute to the formation of opinion, but rather take over the function of original news organisations.

Whereas the goal of task 7.3, to ensure that SYNC3 covers as large a proportion as possible of relevant blogs that comments on news, has remained unchanged, due to certain developments during the first year of implementation of the project, the objectives and subtasks that address
them have been slightly reformulated. To avoid taking up too many resources and producing an enormous amount of irrelevant and unsuitable results, it was decided that early SYNC3 prototypes would crawl a manually compiled list of 500 blog sources. Since the SYNC3 system to date works with a list of hand-picked blog sources, the objective of the first subtask of task 7.3 has been reformulated to expert validation of the degree of relevance of blog results generated by the SYNC3 system at various stages of development. For the first blog source validation report, we have thus taken on the task to further refine the quality of SYNC3 results by systematically expanding the scope of blogs taken into account and thus improving the algorithms. The rapidly changing nature of the online environment and the proliferation of the variety of social media platforms such as micro-blogging platforms and social network sites and their increased prominence in the networked news environment calls into question the meaningfulness of the second subtask, mapping of blog sources with original content, in the context of the SYNC3 definition of news events. The organization of content in SYNC3 around the extremely granular news event definition is another issue towards which future research and development efforts need to be directed as it will be explained in section 3.

This deliverable is structured around documenting the method and results of the two subtasks: the validation of proportion and relevance of blog sources in the first SYNC3 prototype, and the mapping of blog sources with original content. The conclusions present the implications of the findings of this first blog validation report on task 7.3 and on a series of SYNC3 core principles, along with suggestions on how they could be approached.
2. Validation of proportion and relevance of blog sources in the first SYNC3 prototype

Following the actual implementation work, the objective of the first subtask is to validate the proportion and relevance of the blog sources listed in the first SYNC3 prototype (http://sync3.atc.gr/de.l3s.sync3.gui/). This objective is achieved by means of a small exercise whereby the results returned by the first SYNC3 prototype for a series of general queries are compared with the results generated by two other dominant blog search engines, namely Google Search (http://blogsearch.google.com/) and Technorati (http://technorati.com/). The question guiding this process is: "Does SYNC3, at various stages of development, correctly identify a sufficient number of relevant blog posts per news event in comparison with two dominant blog search engines?"

Given the fact that, at this stage of development, SYNC3 indexes a limited set of blog sources and that the comparison is done against dominant blog search engines, the results of the quantitative comparison are expected to be unfavourable to SYNC3. However, given the early prototype stage of the system and the limited set of indexed sources, performance measurement is not the main goal of this exercise. The main purpose of the comparison of results with dominant blog search engines is to identify an as large a number as possible of relevant blog sources from the results returned by major blog search engines, to be added to the SYNC3 list of blog sources that serve as starting points for the crawling process in order to expand it.

To achieve this, we devised the following exercise:

- Type a query related to a news event in the SYNC3 search box. The queries were designed to cover some of the most common news categories as identified in the SYNC3 Deliverable “D7.2.1: Content Package with Simulations,” namely: world, business, politics, sport, science, life, art, technology, and money.
- Type the same query in two blog search engines namely Google Blog Search and/or Technorati.
- Compare the blog sources identified by SYNC3 with the sources identified by the other blog search engines in terms of quantity.
- Assess the quality of blog sources listed by other search engines which are not listed by the search interface of the first SYNC3 prototype against SYNC3 blog source credibility criteria [3] and list the relevant blog sources.

A number of 28 rather general queries were run in January 2011 in the Exploration and Browsing Interface of the first SYNC3 prototype and in Technorati and/or Google Blog Search, covering some of the most common news categories. In terms of quantity, as expected, on average the difference in the number of results returned by the SYNC3 prototype and Technorati and/or Google Blog Search was enormous: 266 results to 10426017.

Approximately 30% of the queries in the SYNC3 search interface listed no blog results, in spite of the queries representing events of global importance such as "Ground Zero," thus extremely likely to be covered by the 17 news agencies used to define news events, and in the blogosphere. To examine whether the problem lies in one of the components of the SYNC3 system such as the search or indexing component, 50 of the 500 manually collected blog sources were scraped for the presence of the eight phrases that returned no results in the SYNC3 prototype with the Google Scraper [4]. It was found that about half of the fifty sources mention four of the phrases several times (See Annex 2). The highest number of blog results listed by the SYNC3 prototype was registered for the query "Obama."
The exercise resulted in the identification of 73 blog sources relevant to the queries that have not been listed by SYNC3, although not all pages of results returned by the other two blog search engines were explored but just the first three or four. The results of the exercise can be found in Annex 1.
3. Mapping of blog sources with original content

The second subtask, the identification and monitoring of blogs that act as independent news sources, particularly from countries with a less developed media landscape where blogs do not only contribute to the formation of public opinion but are breaking news as well, was performed with the assistance of the EJC network of correspondents and international collaborators. This additional exercise was necessary since, due to the fact that the SYNC3 system identifies and defines news events only once they have been recognised and followed up by traditional news media organisations, such blogs might remain underappreciated by the SYNC3 system.

Blogs with original content, that is blogs that report factual information related to issues of public interest which were not picked up by mainstream news media, were identified and examined. One such example of blog with original newsworthy content is bloggingportal.eu, a European blogging platform, which collects several hundred individual and collective blogs about EU affairs in a variety of languages. The content of this platform is news oriented rather than opinion oriented and appears to be complementary to traditional news media. The published blog posts are commentary or analysis of EU affairs facts and typically elaborate on details of background and bring original insights into them. The originality of the content of these blogs however does not lie in their potential to break news in the sense of announcing a significant news event before the traditional news media does, and thus does not conform to the traditional notion of what is news worthy, which has been picked up by SYNC3. A potential explanation for this fact lies in the apparent professionalisation of blogs. With ever more blogs written by professional journalists and published under the auspices of well-respected publication brands, this trend may be the key factor. Typically, a journalist will use the news for the main outlet, and exploit the blog to deliver the context that is not covered by news reports. Other blogs are drafted by all kinds of stakeholders, which by definition rarely break original news, but comment on what the official actors do, or on events pertaining to their specific interest.

Whereas such blogs may not contribute original content that conforms to the SYNC3 definition of a news event (which is according to the DoW “something that happens at some specific time and place,” [5]), and were thus disregarded for the purposes of this subtask, the value of their contributions to public debate is incontestable as they offer a detailed understanding of the background, context and impact of news events.

The difficulties encountered in our attempt to identify blogs with original content are confirmed by the findings of a series of academic studies on the relationship between traditional news media and the blogosphere.

A study [6] of the University of Texas at Austin (USA) on the relationship between blogs and the traditional professional news media in the US and internationally, which takes as a corpus the 50 most popular blogs devoted to news and politics, found that the blogosphere heavily relied on news stories released by professional media organizations. Foundational narratives of the blogosphere which circulated particularly in the United States, positioned bloggers as competitors to traditional news media and as taking over the traditional journalistic pursuit of fact-checking by unmasking mainstream media through fact-checking and grounding journalistic claims. Contrary to these narratives, the blogosphere was found to incorporate journalistic accounts either by linking to news content or by cut-and-paste, instead of being “the source of relentless criticism of press bias that one may have sensed from some higher profile anti-media moments.”
Another study [7] conducted by Stanford and Cornell University researchers confirms our inability to identify blog sources with original content in the sense of the traditional notion of news. The study "Meme-tracking and the Dynamics of the News Cycle," demonstrates by means of analysis of the transition of phrases from news media to the blogosphere and vice versa for a sample of 1000 threads that, in terms of the migration of news from the news sphere to the blogosphere and vice versa, the majority of phrases first appear in news media and then diffuse to blogs with a much smaller yet very important number of phrases that appear in the blogosphere and are then picked up by news media. Thus, in the majority of cases, news first appears in mainstream media and in much fewer cases it appears in the blogosphere. Another empirical study [8] conducted by the University of Leicester on online news confirms that information flows in much larger volumes from traditional news media to the blogosphere than vice versa. Journalist professor Jim Hall remarks [9]: “Blogs do not routinely generate new or original content, rather they depend on other sources (both primary and secondary) for their sources and as the basis for their commentary and annotation. Thus, rather than the ‘grand narratives’ of current affairs they will serve up small fragments of stories.”

The observations resulting from our blog content collection and annotation exercise under task 7.2: Content collection and characterisation, detailed in the SYNC3 Deliverable “D7.2.2: Content Package with Simulations”, [10] namely the difficulty of the human annotators to identify blog posts that comment on news events in an opinionated way and the relatively high frequency of blog posts that copy-and-paste news content, confirm these findings as well. The human annotators systematically had to navigate well beyond the first page of Google Blog Search results to identify opinionated blog posts. In our annotated collection to date a number of 203 blog posts, that is about half of the total of about 389 annotated blog posts express either positive or negative sentiment, as result of efforts of human annotators to specifically identify opinionated blog posts.

The difficulty in indentifying opinionated blog posts was linked with an apparent professionalization of blogs the number of which diminishes, which yet function as serious yet informal sources of background information and comment, as Twitter and Facebook now absorb most of the real-time amateur-level communication. Many of the ones that survive and sustain are written by either journalists or subject matter experts, rather than by laypeople, which makes the expression of sentiment by (semi-)professionals more subtle. Our observation is confirmed by the finding of the same University of Texas study [11], whereby bloggers engaged with facts present in news accounts without challenging them, instead using them to produce their own arguments or reinforce the presented views, thus acting as disseminators of news content rather than as providers of alternative scrutinizing views on news events: “We may thus regard them ironically as in some ways preserving and reinforcing professional norms of journalism as they disseminate content generated by traditional reporting.”

However, even if opinion is not expressed implicitly or explicitly in the blog post content, it can be implied from the broader context, such as profile of the blog or of its author(s). If a fragment of a news story about, for instance, BP having increased their revenues in the past year, is copied-and-pasted on an environmentalist blog, the sentiment of the author towards the news event can be implied to be negative, whereas if the same fragment were copied-and-pasted on a business blog, the sentiment of the author could be implied to be positive. The mere volume of circulation of a news event in the blogosphere can be an indication of sentiment, recognition of the quality of that news story and thus an indication of positive sentiment towards it, or the opposite.
The proliferation of the variety of social media platforms, micro-blogging services such as Twitter, whereas still used primarily for dissemination of news, could constitute an alternative direction for the mapping of original content in the sense of breaking news. Whereas traditional news media outlets are still ahead in reporting breaking events, Twitter scoops professional media on breaking news in a manner similar to the omnipresent CCTV, according to academic and journalistic accounts [12].
4. Conclusions

Given the early stages of development of the SYNC3 system, in which the SYNC3 blog crawler indexes a manually compiled list of sources, expert validation of sources automatically indexed by SYNC3 by means of the EJC network of correspondents and international collaborators will become valuable at later stages of development. In this context the first subtask documented in this first blog validation report shifted focus to validating the proportion and relevance of the blog sources listed by the first SYNC3 prototype with emphasis on the identification of relevant sources that have not been listed by the SYNC3 prototype, and resulted in a list of 73 blog sources.

The findings of the second subtask, namely the difficulty to map blog sources that contain original content that fits the SYNC3 definition of a news event, indicate the necessity to shift the core purpose of task 7.3 away from mapping sources with original content to expanding the list of blogs. In the light of the rapidly changing nature of the online environment and the proliferation of the variety of social media platforms such as micro-blogging platforms and social network sites and their increased prominence in the networked news environment, this report highlighted or confirmed a series of issues which showed up in other WP7 tasks as well, such as the user evaluations, on which future research and development efforts should focus.

As the results of the first user evaluations detailed in the User Evaluation Report [13] indicate an area on which further research and development should potentially focus is complementing the organization of content in SYNC3 around the extremely granular news event definition which might not be useful at a practical level for all the targeted user groups, and misses blogs with original content that do not conform to the traditional definition of news, with organization of news items around stories or topics. Our research and the material collected by SYNC3 so far indicate that blogs breaking original news are next to non-existent. Rather, they comment on details and elaborate on facts that form the background and framework of news events. Based on the findings of this exercise, we recommend that future research and development efforts give a prominent place to the possibility to extend the news event definition by a topic or story definition.
5. References

6. **Annexes**

6.1. **Annex 1: Validation of proportion and relevance of blog sources in the first SYNC3 prototype**

The Excel file will be uploaded on the SYNC3 wiki along with the deliverable.

6.2. **Annex 2: Occurrence of queries in a sample of blog sources**

```
Ground Zero (166)
```

```
Missour Health insurance (2) Roma France (2) Pope visit Britain (2) Clinton Abbas meeting (2) tanan shifted defects (2) Justice department BP (2) Oracle Hurs (2)
```